

MINISTRY OF HEALTH
National Ethics Committee for Research — INAEP

REFERENCE NOTEBOOK

ETHICAL, SCIENTIFIC, NORMATIVE AND METHODOLOGICAL FOUNDATIONS

Artificial Intelligence in Research with Humans

Foundational document for the Guide to the Ethical Use of Artificial Intelligence in Research with Human Beings and MARIAH

Preservation of the Working Group's technical work — Version 1.0

LIST OF FIGURES

Figure 2.1— AI model as a mathematical function: input, transformation and output.....	18
Figure 2.2 — Architecture of a deep neural network: layers of artificial neurons...	29
Figure 2.3 — Internal workings of an artificial neuron in a deep neural network	29
Figure 3.1 — Indicative relationship between performance and explainability of the main algorithms in health research	39

LIST OF ACRONYMS

AI — Artificial intelligence

EIA — Algorithmic Impact Assessment (Canada)

ANPD — National Data Protection Authority

Anvisa — National Health Surveillance Agency

APPI — Japan's Act on the Protection of Personal Information
Information)

CDSCO — Central Organisation for Drug Standards Control of India (Central
Drugs Standard Control Organization)

CEP — Research Ethics Committee

CFM — Federal Council of Medicine

CGI.br — Internet Steering Committee in Brazil

CHS — Humanities and Social Sciences

CNPq — National Council for Scientific and Technological Development

CNS — National Health Council

CNN — Rede neural convolucional (Convolutional Neural Network)

Conep — National Ethics Committee for Research

DNN — Deep Neural Network

DPDP — Digital Personal Data Protection Act of India
Act)

EDPB — European Data Protection Board

EHDS — European Health Data Space

ELM — Extreme Learning Machine

EMA — European Medicines Agency

FDA — United States Food and Drug Administration
Administration)

FHIR — Fast Healthcare Interoperability Resources
Interoperability Resources (HL7 standard)

GAN — Generative Adversarial Network

GDPR — General Data Protection Regulation of the European Union
Protection Regulation)

Grad-CAM — Gradient-weighted Class Activation Mapping

HL7 — Health Level Seven (international organization for health interoperability standards)

ICMJE — International Committee of Medical Journal Editors

ICMR — Indian Council of Medical Research

ICO — Office of the Information Commissioner of the United Kingdom
Office)

IDATEN — PMDA's Timely Innovation Assessment System (Improvement Design within
Approval for Timely Evaluation and Notification)

INAEP — National Ethics Committee for Research

LGPD — General Law for the Protection of Personal Data (Law No. 13.709, of August 14, 2018)

LIME — Local Interpretable Explanations Independent of Models (Local Interpretable
Model-agnostic Explanations)

LLM — Large Language Model

LNCC — National Laboratory for Scientific Computing

MARIAH — Risk Assessment Matrix in Artificial Intelligence for Research with Living Beings
Humans

MCTI — Ministry of Science, Technology and Innovation

MFDS — Ministry of Food and Drug Safety of South Korea

MHLW — Ministry of Health, Labour and Welfare of Japan

MEXT — Ministry of Education, Culture, Sports, Science and Technology of Japan

MHRA — Medicines and Healthcare products Regulatory Agency of the United Kingdom
(Medicines and Healthcare products Regulatory Agency)

MRCT — Multi-Regional Clinical Trials Center

NHC — National Health Commission of China

NHMRC — National Health and Medical Research Council of Australia
Medical Research Council)

NIC.br — Information and Coordination Center of the .BR Domain

NMPA — National Medical Products Administration of China
Administration)

WHO — World Health Organization

PAHO — Pan American Health Organization

OWASP — Open Worldwide Application Security Project Foundation

PIPL — China's Personal Information Protection Act
Law)

NLP — Natural Language Processing

PMDA — Japan Medical Devices and Pharmaceuticals Agency
(Pharmaceuticals and Medical Devices Agency)

RDC — Resolution of the Collegiate Board (Anvisa)

RIPD — Data Protection Impact Assessment Report

RNDS — National Health Data Network

RNN — Recurrent Neural Network

SaMD — Software as a Medical Device

SEIDIGI — Secretariat for Information and Digital Health of the Ministry of Health

SHAP — Shapley Additive Explanations

SINEP — National System of Ethics in Research with Human Beings

SISA — Fragmented, Isolated, Sliced, and Aggregated Machine Unlearning Architecture
(Sharded, Isolated, Sliced, Aggregated)

SUS — Unified Health System

Informed Consent Form

TCPS — Canada's Tri-Agency Policy Statement

UNESCO — United Nations Educational, Scientific and Cultural Organization

WCG — U.S. organization for regulatory services and ethics in clinical research.

XAI — Explainable Artificial Intelligence

SUMMARY

1	Ethical Questions in the Age of Artificial Intelligence.....	7
1.1	A GUIDE ON ASKING QUESTIONS	7
1.2	THE ETHICAL TIMEFRAME OF THIS GUIDE.....	9
1.3	FOR WHOM THIS GUIDE WAS WRITTEN.....	11
1.4	HOW TO USE THIS GUIDE.....	11
1.5	The SUS as Territory and Horizon	13
1.6	METHODOLOGICAL NOTE: THE USE OF ARTIFICIAL INTELLIGENCE IN THE CONSTRUCTION OF THIS GUIDE.....	14
2	CONCEPTS FOR UNDERSTANDING ARTIFICIAL INTELLIGENCE	17
2.1	ARTIFICIAL INTELLIGENCE: ROOTS, HISTORY AND FUNDAMENTALS.....	17
2.2	HOW AI SYSTEMS LEARN: DATA, TRAINING AND MODEL	20
2.3	TYPES OF AI SYSTEMS IN HEALTH RESEARCH	24
3	Critical Properties of AI Systems: Bias, Explainability, Supervision, and Lifecycle	32
3.1	THE FIVE DIMENSIONS OF PROPORTIONAL RISK	32
3.2	Algorithmic Bias: What it is, how it arises, and why it persists	33
3.3	EXPLANABILITY AND TRANSPARENCY.....	37
3.4	HUMAN SUPERVISION AND DECISION AUTOMATION	43
3.5	The AI Lifecycle in Research: From Collection to Disposal.....	47
4	THE BRAZILIAN CONTEXT	52
4.1	A HEALTH SYSTEM THAT DEFIES ANY IMPORTED MODEL	52
4.2	THE BRAZILIAN REGULATORY FRAMEWORK: A LAYERED CONSTRUCTION.....	53
4.3	The National Health Data Network and the Infrastructure for AI.....	55
4.4	SEIDIGI AND THE CONSTRUCTION OF BRAZILIAN DIGITAL HEALTH	57
4.5	Algorithmic Bias, Regional Inequality, and the Risk of the Imported Model	59
5	INTERNATIONAL OVERVIEW: EXPERIENCES IN ETHICAL EVALUATION AND REGULATION OF AI IN HEALTH RESEARCH.....	62
5.1	INTRODUCTION: WHY LOOK OUTSIDE.....	62
5.2	The Multilateral Ethical Framework: The World Health Organization as FOUNDATIONAL REFERENCE	62
5.3	CANADA	64
5.4	Australia	65

5.5 UNITED STATES	66
5.6 EUROPE.....	67
5.7 ASIA.....	68
5.7.1 China.....	68
5.7.2 Japan	69
5.7.3 South Korea	70
5.7.4 India.....	71
5.8 RUSSIA.....	72
5.9 Comparative Summary and Lessons for Brazil	72
6 SCIENTIFIC INTEGRITY AND AUTHORSHIP IN AI RESEARCH	75
6.1 WHAT CHANGES WHEN AI ENTERS RESEARCH.....	75
6.2 HUMAN AUTHORSHIP: PRINCIPLE, FOUNDATION AND LIMIT	77
6.3 Algorithmic Plagiarism, Negligence in Misconduct, and Fabrication of EVIDENCE.....	78
6.4 Civil Liability and the Extended Chain of Liability	80
6.5 WHAT THE PROTOCOL SHOULD CONTAIN — AND WHAT THE CEP VERIFIES	82
7 DATA GOVERNANCE AND PARTICIPANT PROTECTION.....	86
7.1 HEALTH DATA IN AI SYSTEMS: A DIFFERENT NATURE	86
7.2 AGENTS, ROLES AND RIGHTS IN DATA RESEARCH.....	88
7.3 RESEARCH DATABASES: CONSTITUTION, FUTURE USE AND WAIVER OF CONFIDENTIAL INFORMATION	92
7.4 DATA SOVEREIGNTY AND INTERNATIONAL TRANSFER	94
7.5 Federated Learning and Privacy Architectures.....	98
7.6 WHAT THE PROTOCOL SHOULD CONTAIN AND WHAT THE CEP VERIFIES.....	101
8 MARIAH: ARTIFICIAL INTELLIGENCE RISK ASSESSMENT MATRIX (AEB VERSIONS)	106
8.1 The Construction Process and the Paths Open to Ethics Committees in RESEARCH.....	106
8.2 FUNDAMENTALS, SOURCES AND DESCRIPTION OF MATRICES	107
8.2.1 Qualitative Version A	108
8.2.2 Quantitative Version B	113
9 SPECIFICITIES BY NATURE OF THE RESEARCH	121
9.1 HEALTH RESEARCH: ASSISTANCE IMPACT, ANVISA AND SAMD INTERFACE	121
9.1.1 When the AI system is also a medical device.....	121
9.1.2 The CEP-Anvisa sequence: complementary, non-concurrent mandates.....	122
9.1.3 What the SaMD protocol should contain — and what the CEP verifies	123

9.2 RESEARCH IN HUMAN AND SOCIAL SCIENCES (HSS)	125
9.2.1 A territory with its own risks.....	125
9.2.2 Consent in CHS with AI: what changes	129
9.2.3 Collective risk and vulnerable groups in CHS	132
9.2.4 What the AI-based CHS protocol should contain — and what the CEP verifies	134
10 RELATIONSHIP WITH THE INAEP GUIDE AND MARIAH	136
11 INTERNATIONAL REGULATORY UPDATE — 2024–2026	137
11.1 EUROPEAN ARTIFICIAL INTELLIGENCE REGULATION (AI ACT).....	137
11.2 EMA AND THE MEDICINE LIFE CYCLE.....	137
11.3 Joint EMA–FDA Principles of Good AI Practices in Development OF MEDICINES	138
11.4 European Health Data Area (EHDS)	138
11.5 EDPB AND SCIENTIFIC RESEARCH — GUIDELINES 1/2026	139
11.6 EDPB AND ANONYMIZATION — GUIDELINES 02/2026	139
11.7 BRITISH RECOMMENDATIONS FOR THE REGULATION OF AI IN HEALTH	139
11.8 CONSEQUENCES FOR THE BRAZILIAN FRAMEWORK AND FOR MARIAH.....	140

BACKGROUND NOTE

This Handbook does not establish additional requirements for the ethical evaluation of protocols. Its purpose is to present the ethical, scientific, normative, and methodological basis for the guidelines contained in the INAEP Guide and the criteria operationalized by MARIAH.

This Reference Notebook preserves and organizes the rationale produced by the Working Group that supports the INAEP Guide for the Ethical Use of Artificial Intelligence in Research with Human Beings and MARIAH — the Risk Assessment Matrix for Artificial Intelligence in Research with Human Beings.

Its function is explanatory and referential. The document brings together the conceptual, ethical, normative, methodological, and comparative framework that underpins the guidelines and operational instruments, allowing the reader to understand why certain questions, risk dimensions, and safeguards were incorporated into the Brazilian model.

The Reference Notebook does not replace the INAEP Guide, does not constitute a risk classification instrument, and should not be used in isolation as an evaluation checklist. For practical application, the INAEP Guide and MARIAH should be consulted.

The decision to maintain this document as a standalone document acknowledges the accumulated technical work of the Working Group and preserves the methodological memory of the construction of the Brazilian framework, without transferring to the main Guide the scope inherent to a foundational document.

This work is intended for Brazilian society, the scientific community that conducts research with artificial intelligence in health in Brazil, the researchers who propose the protocols, the Research Ethics Committees that evaluate them, the institutions, and the data managers who make them possible.

1. Ethical Questions in the Age of Artificial Intelligence

1.1 A GUIDE ON ASKING QUESTIONS

Ethics in research involving human subjects has always depended more on the ability to... to question rather than the ability to dominate. Researchers know in detail the protocols that they develop. In the Brazilian context, their peers are members of Ethics Committees. Research Ethics Committees (CEPs) are familiar with the principles that the protocol must meet. Among these Two pieces of knowledge, the right question is the instrument that allows dialogue.

This guide was built on that premise. Its purpose is not to transform your Readers are experts in artificial intelligence (AI). It is about offering the questions that Research ethics professionals need to educate those who develop and use technology. These questions don't belong only to those who evaluate. Whoever proposes the study must answer them. They are not part of the protocol itself. This is the dialogue upon which this guide was built.

This distinction has immediate practical consequences. Specialized training in neural network architectures, model performance metrics or protocols Cryptography requires years of study, which is neither required nor assumed by composition. multidisciplinary approach of the CEPs.

The capacity for ethical questioning demands conceptual precision regarding what is at stake. For the research participant, transparency regarding what guarantees the protocol must provide is essential. to offer and the firmness to demand them even in the face of scientific enthusiasm and asymmetry of Technical knowledge. This is the ability this guide can build, and it is to build it that it was created. escrito (MRCT Center; WCG, 2025; OMS, 2021).

The right question, asked at the right time, is the most powerful tool for protection. The ethics that a committee possesses. Not because it replaces the researcher's technical judgment, but because it obliges him to make it explicit, document it, and submit it to the scrutiny of those he represents. The interests of people who are not in the room when the protocol is approved. Ask Regarding the demographic composition of the training database, the withdrawal mechanism... consent after model training and the system's destination after termination. The research demonstrates an understanding of the risks and prevents the approval of what cannot be approved. to be properly evaluated. The speed of development of artificial intelligence.

This places any technical knowledge in a peculiar position, since it can be...
to become partially (or totally) obsolete before being fully learned.

The neural network architecture that dominates clinical performance today may be replaced by a different approach in two or three years. The language model that
impressive to experts this semester, it will present documented limitations next semester.

For a guide intended to provide guidance on the writing and ethical evaluation of protocols of
Research spanning a decade reveals that this speed represents a challenge that cannot be solved simply by...
through technical means, but it also demands an ethical response.

The questions this guide offers don't become outdated at the same rate as the...
architectures. The question of the demographic representativeness of the training data remains.
valid for any system, of any generation, trained on any mode of
clinical data. Verification of actual human supervision, distinct from merely nominal supervision,
It permeates all architectures and all contexts of use. The reflection on what will happen
Having the model after the research is completed should precede any discussion about its
Technical performance and survives any protocol update.

These are ethical questions presented in a technical manner. This guide aims to make them accessible.
but without requiring the reader to master this terminology, because terminology changes and ethics change.
permanence (Youssef *et al.*, 2024; Barucci *et al.*, 2026).

There is an additional dimension to this choice. Those who learn to ask the right questions,
Working on a committee or research panel develops a transferable skill. It can be applied...
there are protocols with systems that this guide did not anticipate, in contexts that have not yet been implemented.
There are technologies that don't yet have names. The person with technical expertise who
Those who learn a specific architecture need to relearn it when the architecture changes.
Learn to ask; you don't need to start over. Ask again.

The Inaep Guide to the Ethical Use of Artificial Intelligence in Research with Human Subjects
It offers a repertoire of questions. Not a technical vocabulary to impress.
researchers, nor a compliance manual to meet formal requirements. The Guide
proposes a set of precise, well-founded, and operational questions to protect the
People participating in artificial intelligence research in healthcare in Brazil.

The difference between a compliance guide and an ethics guide lies precisely in the type
of the question that each person proposes. The first one verifies if the protocol meets the requirements.

established principles, the second question is whether the choices made in the research are ethical. justifiable.

The form can be fulfilled with a Free and Informed Consent Form (TCLE). which mentions AI without explaining what that means for the participant. The substance requires that The informed consent form (TCLE) should describe, in accessible language, what the system will do with the person's data, who... will have access to the resulting model and how the right of withdrawal will be exercised in a system that He has already learned from that data.

The form can be satisfied with a statement that a doctor will supervise the The algorithm's decisions. The substance requires demonstrating that this professional understands the The basis of the algorithmic decision has time to contest it, in addition to proving that the The system was designed to record when you disagreed.

Asking the right questions is also a political act. A question formulated precisely about... topics such as data sovereignty, digital colonialism in health research, and... The use of SUS (Brazilian public healthcare system) data on foreign AI platforms serves more than just a function. regulatory, but civilizing function.

It is in this sense that the opinion can contribute to intelligent research. Artificial intelligence in Brazil responds to the interests of the people who generated the data, not just to... The interest of those who developed the system. The National System of Ethics in Research with Beings Human Resources (SINEP), established by Law No. 14,874/2024, is the evolution of the CEP/Conep System, to which happens in the structure designed by the Law and its regulations. The competencies assigned by resolutions of the National Health Council prior to the Law — among them the Resolution CNS No. 738/2024, widely cited in this guide — reads here as SINEP competencies (Brazil, 2024a). The system is based on the understanding that the Protecting participants requires an independent system with a mandate to do so. These are questions that scientific enthusiasm and market pressure tend to silence. This guide is... The instrument that makes this mandate exercisable in the field of AI in healthcare.

1.2 The Ethical Timeframe of this Guide

Artificial intelligence is already integrated into health research protocols submitted to... Brazilian postal codes. Among the examples are systems that support diagnostic imaging. trained on university hospital databases, predictive risk models

epidemiological tools applied to populations within the Brazilian Unified Health System (SUS), using natural language processing tools. natural sciences regarding electronic health records and generative algorithms that produce data synthetic methods are used to circumvent privacy restrictions. These protocols are evaluated with... Instruments built for another world, because specific regulations haven't arrived yet. with the necessary depth to cover what the technology already does. This guide stems from that. interval.

It would be comforting to promise that the interval is temporary, that regulation will come, that the The gap will be filled and this guide will become unnecessary. That promise would be imprecise. The gap between the speed of technological innovation and the speed of production. Regulation is a structural condition of accelerating fields, not a correctable administrative flaw. Bioethics did not wait for the conclusion of the philosophical debate before needing to respond to... The reality of clinical trials with new drugs.

Ethics in data research did not wait for data science to mature. Before needing to protect participants. Ethics in AI research will not wait for regulation. complete to exercise its mandate. The CEP decides now, based on the existing principles, applied to technologies that the principles did not nominally anticipate, but that the Principles are achieved with precision when the right questions are asked (Soares; Chiavegatto) Son, 2026; Wilton Park, 2026).

This guide is the institutional response from SINEP, through the National Ethics Committee. in Research (INAEP), linked to the Brazilian Ministry of Health, at this time. It is not This is neither a waiting document nor an academic exercise. It is a stance taken on what the system Brazilian research ethics standards require that AI research be carried out today, based on principles established and based on the experience accumulated by the CEP/Conep System throughout its history, initiated in 1988, with the publication of Resolution No. 01 of 1988, and considering the milestone current legislation, Law No. 14,874/2024 and Decree No. 12,651/2025 which regulates it (Brazil, 2024a; Brazil, 2025a). Brazil has a legitimate research ethics system. Historical significance, territorial reach, and multidisciplinary composition are qualities few countries in the world possess. They built it with the same solidity. This system has the conditions to advance in leadership. from the field of ethics in AI research. This guide is a step in that direction.

1.3 WHO IS THIS GUIDE WRITTEN FOR?

This guide is intended for Brazilian society and the scientific community that conducts research. AI in healthcare in Brazil: to the researchers who propose it, to The institutions and data managers that enable it, and the Ethics Committees that evaluate it. Brazilians bring together specialists such as doctors, nurses, dentists, social workers, philosophers, jurists, theologians and user representatives, and this multi-professional composition It is a deliberate force of the system, not a limitation. None of its readers need it. Understanding neural networks to fulfill your ethical role: every technical concept that appears here It exists because it is necessary to formulate a precise ethical question. No concept The technician appears by himself.

The researchers who develop or propose research with AI. Healthcare professionals in Brazil are the primary target audience for this guide. For them, the guide offers clarity on... What the CEP will ask and why — not to prepare convenient answers to questions. Predictable, but necessary to build protocols that deserve affirmative responses. A Transparency regarding ethical requirements serves the honest researcher as much as it protects them. the participant. The protocol anticipates the questions in this guide and answers them with Honesty is the protocol that AI research in healthcare in Brazil needs to produce.

The structure of the Guide allows for both full reading and searching by topic or need.

Those who propose or evaluate a protocol with federated learning can go Directly to Chapter 7. Those who need to understand what data derivation is before evaluating it. A longitudinal protocol finds that answer in Chapter 2. Who wants the roadmap? The complete ethical evaluation process covers Chapters 6 and 7 (scientific integrity and governance). (of data). Those who need the proportional risk assessment instrument should go to Chapter 8. (Risk Assessment Matrix in Artificial Intelligence). Who wants the specification by Nature of the research: Computer Program as a Medical Device (Software as a Medical Device (SaMD) and Human and Social Sciences, go to Chapter 9. For complete reading... It shapes both the proposer and the evaluator. The focused consultation provides the tools for the evaluation process.

1.4 HOW TO USE THIS GUIDE

The guide that the reader holds in their hands was constructed with deliberate architecture. The Chapter Form 2: It provides the minimum vocabulary necessary for ethical questions about AI to make sense. without requiring technical expertise.

Chapters 4 and 5 provide context for the Brazilian context and the international panorama that... determine the territory where these questions will be asked and the power asymmetries that... make them necessary.

Chapter 6 addresses accountability: it deals with scientific integrity and authorship in research. AI, which defines the chain of responsibility for the protocol that reaches the CEP.

Chapter 7 governs: it translates the General Data Protection Law (LGPD) and the CNS Resolution No. 738/2024 on operational requirements regarding the data that the protocol (Brazil, 2024a). Chapter 8 provides the tools: it presents the Risk Assessment Matrix. in Artificial Intelligence in Research with Humans (MARIAH) as a tool for proportional risk assessment.

Chapter 9 specifies: it deals with the specificities of protocols with the Program of Computer as a Medical Device (SaMD) and with research in Human and Social Sciences. Parts V and VI of MARIAH provide support for the training that makes the practice of An ethical mandate that becomes progressively more qualified over time.

Parts III (Version A) and IV (Version B) of MARIAH are operational instruments of Support for assessment. The Risk Assessment Matrix in Artificial Intelligence, available in The qualitative and quantitative versions convert the principles of the guide into concrete criteria for Application in protocols, with individual analysis of each dimension and questioning. The guide The matrix shapes the person who asks the question. The matrix instrumentalizes the act of evaluation. Both are necessary and The two were designed to work together.

This booklet is not a legal manual, a clinical guideline, or a document of... Technical certification of AI systems. Its function is to provide a basis for evaluation. Ethics. It does not replace the committee's judgment, it does not dispense with collegial deliberation, and it does not... It offers ready-made answers to dilemmas that require careful consideration. It offers the questions that... They make this assessment possible with the precision that the field demands.

Ethics in research has always been, ultimately, a collective exercise and based on human judgment about what can be done to other human beings in name of scientific knowledge.

Artificial intelligence does not alter this fundamental principle. It demands that it be exercised with... more rigor, more information, and more awareness of the consequences than a bad opinion. a well-founded analysis can produce results on populations that were not in the room when the protocol was implemented. It was approved.

There are also things this guide doesn't change. A large portion of the protocols involve some use of AI. The data that reaches a CEP today involves studies with tabular data and predictive models. consolidated. These are practices that are already part of the daily work of the committees, and for them this The guide does not create a new barrier. The depth of the assessment remains proportional to what the The system performs the protocol. MARIAH's input filter classifies usage as routine. It does not influence the decisions of the study nor does it produce effects on the participant, and the The protocol follows the standard CEP (Brazilian Postal Code) flow. Uses that pass through the filter are found at the following levels... The initial stages of the matrix offer a path compatible with the risk they present. Level I, the lightest level, already... It includes protocols such as those supporting exploratory analysis of de-identified data. For these studies, this guide only asks for what the committee already does well.

1.5 The SUS as Territory and Horizon

Brazil's Unified Health System (SUS) is responsible for the right to health for more than... 214 million people (Brazil, 2026e) in continental territory, with diversity epidemiological, ethnic, linguistic, and regional data that no global private database can provide. This diversity is simultaneously the greatest scientific asset of the public system. Brazilian healthcare and its greatest ethical responsibility. The data that the SUS (Brazilian Public Health System) produces daily, Over decades of universal service, they represent a wealth of information about Human health in contexts of inequality that the world doesn't have anywhere else. Researching with this data is a privilege that comes with proportional obligations.

This guide was written for this territory. Not for an abstract health system or... for the clinical context of high-income countries where most AI systems in healthcare It was developed and validated. Writing it from the perspective of the SUS (Brazilian Public Health System) does not restrict its scope: the same These guidelines apply to research conducted in supplemental healthcare and private institutions. subject to the same ethical and normative framework as research with human beings in Brazil. The questions it offers were constructed with the specificities of the SUS (Brazilian Public Health System) in mind: a regional diversity that makes algorithmic bias a public health risk, the fragility of

infrastructure that limits what can be demanded and what can be monitored, the presence of indigenous peoples, quilombola communities, and riverine communities in contexts that no model Imported knowledge. Ignoring this specificity would be like producing a guide on ethics in AI research that is not useful for Brazil. It would be just another imported document without translation to the South.

The horizon of this guide is the SUS (Brazilian Public Health System) as a civilizing project, not merely as... healthcare infrastructure. A system built on the principle that health is universal right, that the research that feeds it must respect the people who support it and that the artificial intelligence operating within it must be governed by the public interest. This The horizon determines what the questions in this guide protect: not only the participants. Individual aspects of each protocol, but the collective collection of data, care, and trust that The Brazilian Unified Health System (SUS) is represented. Research using AI in the SUS that does not answer the questions in this guide is not... This is simply ethically inadequate research. It is research that compromises public assets. which makes it possible.

The following chapters develop these questions. They begin with vocabulary. necessary to formulate them, in Chapter 2.

1.6 METHODOLOGICAL NOTE: THE USE OF ARTIFICIAL INTELLIGENCE IN THE CONSTRUCTION OF THIS GUIDE

This guide was produced in accordance with the ethical principles you recommend. CNPq Ordinance No. Law 2.664/2026, which established the Policy of Integrity in Scientific Activity, requires the declaration. explicit use of generative AI tools at any stage of intellectual production of research or funding (CNPq, 2026). Chapter 6 of this guide translates this requirement into Operational parameters for protocols submitted to Research Ethics Committees. A Editorial consistency demands that the guide itself be transparent about its use of AI. construction, and applies to itself the same standard that it proposes to others. The construction of this guide It involved the coordinated use of five generative AI systems between March and April 2026. under explicit human editorial direction at all stages. Claude Opus 4.7 (Anthropic) acted in the complete editorial review of the chapters in four layers (spelling, style, content and structure), in the rewriting of passages under the organizer's direction, in the harmonization of acronyms and of bibliographic references, in the preparation of navigation charts and maps of

changes in the layout of the book, in the construction of the index, and in Connection between chapters.

ChatGPT 5.4 (OpenAI) was used in exploratory discussions of the structure of chapters, initial bibliographic survey and suggestion of conceptual alternatives on the topics specifics. Gemini 3.1 Pro (Google) was used in comparative research on regulation. International AI in healthcare, with emphasis on Brazil, Asia (China, Japan, South Korea and India) and Russia, which served as initial input for Chapter 5. Microsoft Copilot was employed in Assistance in preparing tables and synoptic charts, and in verifying consistency. terminological throughout the writing process. DeepSeek was used for cross-validation of technical content and in identifying gaps in specific learning topics of Machine and data governance.

No chapter was produced independently by any system. The structure of The book encompasses the selection of arguments, editorial decisions, and the validation of each claim. The responsibility for these tasks fell to the human authors, whose names are listed on the identification form.

All content generated or suggested by AI systems has been reviewed. entirely by the authors before composing the final text. Bibliographic references The information was confirmed in the original sources; quotations were compared with the cited documents. Regulatory provisions were included in the official text of laws, decrees, and resolutions. The verification followed the standard that Chapter 6 establishes as a minimum: identification of The tool, a description of the function performed, and certification from the human responsible for validation.

The human personnel responsible (supervisor and editor) named on the identification sheet. They declare that no AI-generated content has been submitted as human-authored without substantive review, which third-party projects were not incorporated into AI tools for preparation of reports, and that full responsibility for the content of this guide rests with them. It belongs. AI is not listed as a co-author, a position consistent with the World Health Organization. Health, from the International Committee of Medical Journal Editors and CNPq Ordinance no. 2.664/2026, resumed in Chapter 6, §6.2.

The use of AI accelerates text production, but introduces specific risks that this guide will address. He acknowledges in his own chapters: hallucination of references and facts (Chapter 2, section on algorithmic hallucination; Chapter 6, §6.3); inherited bias from training data (Chapter 3, §3.2; Chapter 4, §4.5); tendency toward rhetorical confirmation in grand models of

language; and terminological imprecision in Brazilian normative devices in the systems trained predominantly in English.

Three mitigation protocols were adopted: (a) verification of all citations (a) bibliographic in primary source; (b) comparison of each normative statement with the official text of law, decree or resolution referred to; (c) full reading, by a responsible human researcher, review each chapter before inclusion in the book.

The statement provided here complies with CNPq Ordinance No. 2,664/2026 (CNPq, 2026). This is linked to Law No. 14.874/2024, which establishes SINEP and sets the full responsibility of researcher throughout the protocol (Brazil, 2024a), and is consistent with CFM Resolution No. 2.454/2026 regarding the use of AI as a tool to support human decision-making, never a substitute (CFM, 2026).

The transparency provided here serves the dual purpose of upholding ethical standards. proposed by the guide and to offer authors and researchers a practical example of what This means fulfilling the requirement to declare the use of generative AI in scientific production.

2 Concepts to Understand Artificial Intelligence

2.1 ARTIFICIAL INTELLIGENCE: ROOTS, HISTORY AND FUNDAMENTALS

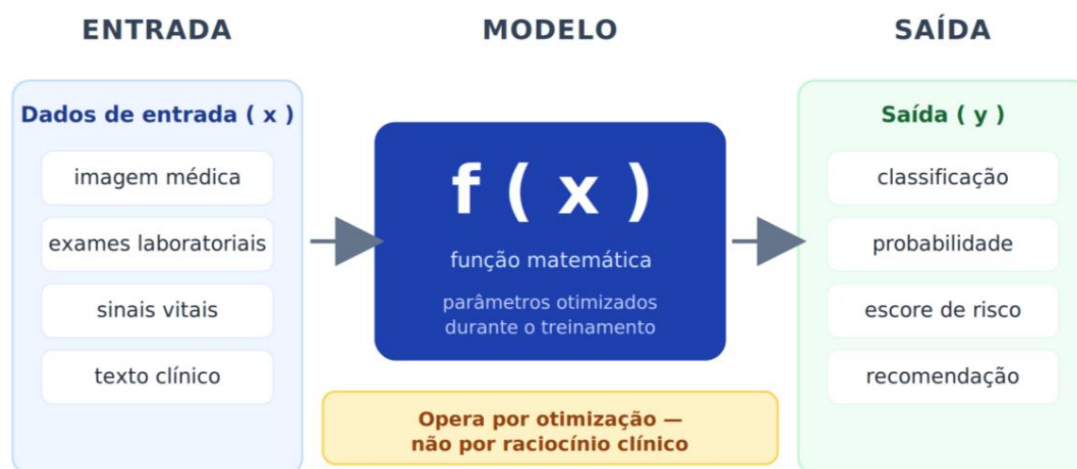
This chapter presents the necessary foundations for understanding what a artificial intelligence model, how it learns from data, and which families of Algorithms appear more frequently in research protocols submitted to National System of Ethics in Research (SINEP). The properties that transform these Technical choices in ethical matters — such as bias, explainability, human oversight, and the cycle. of life — will be examined in Chapter 3.

Artificial intelligence is already present in research evaluated by Ethics Committees. in Research (CEPs), including in diagnostic support systems, predictive models, natural language processing tools and generative architectures. These Systems often arrive at postal codes accompanied by technical descriptions that They presuppose specialized knowledge on the part of the evaluators.

The purpose of this chapter is not to train machine learning experts. but to offer CEP members the minimum vocabulary necessary to understand the intended use, identify missing information, and formulate ethically relevant questions. prior to protocol approval. Each section functions as an autonomous reading unit. For example, someone evaluating a protocol that uses a generative system can consult... directly to Section 2.3.

A model is a mathematical function (Figure 2.1). It receives input data and produces... one output: a classification, a probability, a risk score, a recommendation. The model operates by optimizing a previously defined mathematical function. defined during training, without performing clinical reasoning in the human sense of the term.

Figure 2.1— AI model as a mathematical function: input, transformation, and output.



Source: adapted from Salih *et al.* (2023) and Santana, Moreno and Santos (2025), with examples documented in health research in Brazil (Barbosa *et al.*, 2021a, 2021b; Santos *et al.*, 2026).

The distinction between optimization and reasoning defines the scope in which they focus. The main ethical risks associated. An optimized model to maximize accuracy in a set of historical data can produce technically correct results in that context. context and clinically inadequate in another. Mathematical optimization does not have access to meaning of the data it processes (Santana; Moreno; Santos, 2025).

In a brief historical context, the founding concepts of AI emerge from contributions accumulated throughout the 19th century and the first half of the 20th century. Figueiredo and Aguiar (2025) document this trajectory starting from Charles Babbage, who conceived the Analytical Engine, and Ada Lovelace, a direct collaborator of Babbage and Recognized as the first programmer in history, Lovelace was the first to... to understand that a calculating machine could go beyond numbers and operate on symbols of any kind. This intuition, formulated in 1843, anticipated by a century the which Alan Turing would formalize with the Turing machine and the concept of computability.

It is important to highlight that women's contribution to the history of AI is not... It limits Ada Lovelace. Many women have played an essential role in the history of AI, but many They were not recognized. The six programmers of the Computer and Numerical Integrator The Electronic Numerical Integrator and Computer (ENIAC) developed the The first electronic computer in the US in 1946, and only then were publicly credited... decades later (Light, 1999).

Grace Hopper developed the first compiler in 1952 and coined the concept of human-readable programming language (Beyer, 2009). Hedy Lamarr patented, with George Antheil, spread spectrum, the foundation of modern wireless communications. (Figueiredo; Aguiar, 2025). These contributions are not included in the canonical books of the area. This absence reflects the same exclusion pattern as algorithms trained on data.

Historical narratives tend to reproduce and amplify these patterns.

The erasure is not limited to women. Shun'ichi Amari published, in 1967, the first neural network with gradient-based weight adjustment, mathematical foundation of Modern machine learning (Amari, 1967). Kunihiro Fukushima developed, in 1980, The Neocognitron, the first deep convolutional network, a direct precursor to neural networks. Convolutional neural networks (CNNs) dominate diagnosis by Image today (Fukushima, 1980).

The 2024 Nobel Prize in Physics was awarded to John Hopfield and Geoffrey Hinton for contributions that Amari and Fukushima had anticipated and, in key aspects, substantiated (Figueiredo; Aguiar, 2025). Durham University documented this. Systematic erasure: Japanese AI pioneers have been written out of history. canonical framework of the field (Durham University, 2024). The field that is built by forgetting its founders produce instruments that tend to forget the populations that do not dominate the narrative.

In 1956, researchers gathered at Dartmouth coined the term intelligence. artificial intelligence (AI) and pioneered the field as a discipline. Frank Rosenblatt created the Perceptron in 1958, the first neural network capable of learning patterns through learning. Supervised. The promise was transformative.

However, disappointment came early: the mathematical limitations of the Perceptron, The findings documented by Minsky and Papert generated what are known as AI winters . Periods of funding cuts and academic discrediting. The thaw came in 1986. when the backpropagation algorithm *enabled* the training of networks capable of representing non-linearly separable relationships and reopened the field for a new generation of researchers (Figueiredo; Aguiar, 2025).

The next rupture was structural. In 2017, Google researchers published the article "*Attention Is All You Need*" (Vaswani *et al.*, 2017), introducing the architecture of transformers . This work made possible the development of

Large Language Models (LLMs), systems trained on vast volumes of text capable of generating language, summarizing documents, and supporting decisions. clinics with increasing sophistication (Figueiredo; Aguiar, 2025).

In parallel, deep learning *began* processing images. medical, physiological signals, and genomic sequences with performance comparable to that of Human experts in specific tasks.

For postal codes, this trajectory offers a perspective that goes beyond their dimension. Historically, the evolution of artificial intelligence also reveals processes of inclusion and exclusion. present in the development of the systems and in the selection of the data used for their Training. Algorithms learn patterns recorded in historical data, including absences, underrepresentations, and forms of erasure present within them. History itself. this field, marked by the late or insufficient recognition of the contribution of Some of its precursors demonstrate that invisibility is not just a problem for... data, but also the institutions and practices that guide the production of knowledge. In research involving human subjects, this legacy demands attention to the groups that are absent. or underrepresented in databases and the discriminatory effects that the systems They can reproduce or enlarge them.

This historical and ethical dimension is materialized, in a concrete way, in the technical stages. that structure the very process of building the models.

2.2 HOW AI SYSTEMS LEARN: DATA, TRAINING, AND MODELS

An artificial intelligence system in healthcare begins with clinical data produced... In real-world contexts, by real people, and under conditions unknown to the model. directly. Before the training, this data has already been subjected to successive decisions. Humans: what information to collect, which populations to include, which instruments to use and Under what infrastructure conditions should the collection be carried out? These decisions influence the quality. The representativeness and limitations of the resulting system.

Youssef *et al.* (2024), in an article published in the *European Cardiology Review*, highlight that the AI research protocol should describe the complete data flow, including collection, preprocessing, splitting of datasets, training, validation, testing and eventual post-processing of the model outputs. The data splitting strategy.

It follows this sequence and organizes the model's learning process into steps.

distinct.

The training set contains the labeled examples with which the model adjusts its...

Internal parameters for associating input patterns with the correct outputs. The set of

Validation tracks performance during training and guides adjustments.

hyperparameters, without directly altering the model weights. The test set

It remains isolated until the end of the process and allows for an impartial assessment of its capacity.

generalization to new data (Youssef *et al.*, 2024). This separation allows estimation,

With greater rigor, the model's ability to generalize, without ensuring that the performance

What is observed will be reproduced under real-world conditions of use.

Therefore, postal codes must have adequate information about the data, including traceability, linkability, and auditability, to ensure that training data

be compatible with the intended use of the AI system. For CEP, this means that the

The protocol should specify not only what data was used, but also from which population it was collected.

extracted, with what selection criteria, did they undergo pre-processing as

standardization or harmonization, if any part was generated synthetically and, if so, with what means.

justification.

O Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard (MRCT Center) and the WIRB-Copernicus Group (WCG) further recommend that the protocols Clearly specify the levels and privileges of access to the data and the AI system.

during training, including whether the activities take place in dedicated local infrastructure.

or on cloud platforms and third-party resources (MRCT Center; WCG, 2025).

Transparency regarding the origin, quality, and processing of data is

indispensable for evaluating whether the performance observed during development may

to be maintained under real-world conditions of use. Satisfactory results in training data or

Operating in controlled environments does not guarantee that the system will be equally useful, safe, or

I need it in other contexts.

A diabetic retinopathy screening system, for example, may present

Good performance when validated in a reference center equipped with the necessary equipment.

High-resolution and specialized professionals. The same system can produce results.

different when used in basic health units with less expensive equipment.

Quality, different collection conditions, and professionals without specific training. This

This occurs because the model learns relationships present in the data and the context in which they are used. were produced. When this context changes, the model does not adapt automatically. the learned parameters: continues to produce results based on the standards of original atmosphere.

For this reason, the protocol must justify the correspondence between the context of development and intended use context, as well as providing for local validation when differences in population, infrastructure, equipment, workflows, or qualifications of Users could affect system performance. In the protocol evaluation, this The properties of the models have a direct implication.

The MRCT Center and the WCG (2025) recommend that Committees examine the context intended use from the initial protocol assessment. Relevant issues include:
For which population was the model trained and validated? Is this population sufficiently large? representative of the people and groups that may be affected by the system under certain conditions Real-world usage? The model does not automatically identify differences between the population. represented in the development data and the one in which it will actually be applied. It is up to the researcher to recognize these differences, evaluate their possible consequences, and propose appropriate validation and monitoring measures.

This correspondence can also change over time. Models of Artificial intelligence does not automatically keep up with changes occurring in population, in healthcare practice or in the environment in which they are used. When a system It is trained with clinical data from a specific period, and it learns relationships associated with... existing conditions, such as available technologies, current regulations, profile epidemiological data on the population, care flows, and diagnostic criteria.

If these conditions change and the system is not reassessed or updated, it continues. producing results based on previously learned patterns. The divergence progressive correlation between the data used in development and the data found during the process. This usage is called **data drift** . Youssef *et al.* (2024) highlight this dynamic. as a dimension that has been insufficiently examined in AI research protocols, although it may compromise system performance, participant safety, and the scientific validity of the results.

Data drift takes different forms. Covariate drift . This occurs when the distribution of the input data changes without a change in the relationship between the input and

The output may change. A screening model trained on a middle-income urban population may to degrade when applied to a rural population with distinct comorbidity patterns, even if the diagnostic criteria remain the same.

Concept drift, on the other hand, is more profound: it occurs when the very... The relationship between the input variables and the outcome changes. The introduction of a new therapeutic protocol, a change in the criteria of a diagnostic classification or the The emergence of a variant of an infectious agent may render obsolete the logic that... The model learned, but the system didn't signal this obsolescence.

For health research in Brazil, this risk has an additional dimension. The diversity Regional integration of the Unified Health System (SUS) means that a model validated in a specific context. It can degrade rapidly when transported to another location. Soares and Chiavegatto Filho (2026) demonstrated that algorithms developed without local validation produce results Inadequate when applied to populations with epidemiological profiles distinct from the data. training.

Data drift is not a future phenomenon to be managed after approval of protocol. It begins the moment the model is developed in a context and intended operating in another. The CEP that evaluates a multicenter protocol with AI in multiple regions. Brazil simultaneously assesses the risk of drift between the participating contexts.

The AI research protocol should include mechanisms for continuous monitoring. of the model's performance throughout the entire duration of the research. Youssef *et al.* (2024) They describe three relevant phases in the lifecycle of an AI system: training, testing, and... Actual conditions before commissioning and monitoring after commissioning.

The absence of any of these phases indicates that the researcher did not map the Expected system behavior beyond the development phase. In the CEP assessment, A protocol without a continuous monitoring plan describes a system that will be... abandoned to its own devices soon after approval.

In view of this, the MRCT Center and the WCG (2025) advise ethics committees to to verify the definition of clear criteria for minimum acceptable performance. The protocol It also needs to establish degradation limits that trigger system review and predict communication mechanisms for participants regarding relevant changes in The model's behavior. This set of requirements supports ethical integrity and... Public trust in health research.

These criteria constitute essential conditions for the continued validity of Informed consent. Consent is based on the declared performance of the system at the time of the participant's decision. Changes resulting from data drift. They modify this framework and demand a proportionate ethical response. The ethics committees in Research should require prior planning for these scenarios, with safeguards that... preserve the legitimacy of consent over time.

2.3 Types of AI Systems in Health Research

Choosing the algorithm is the first technical decision in a research protocol with AI. It's also an ethical decision. Each family of algorithms occupies a distinct position in This spectrum creates tension between predictive performance and explainability, and this position determines... What risks does the system carry and what questions should the CEP (Ethics Committee) ask? This section examines the... main families present in AI research protocols in the Brazilian public health system (SUS), based on studies. Performed in Brazil on the diagnosis of infectious diseases, classification of breast lesions. and prediction of arboviruses (Barbosa *et al.*, 2021a, 2021b; Santos *et al.*, 2026). For each family, The chart indicates the level of explainability and the central ethical question that the Ethics Committee must know. formulate. This question does not require the evaluator to master the mathematics of the algorithm. It requires so that you know what to ask the researcher about the system being evaluated.

The level of explainability indicates how easily the model's reasoning can be understood. understood by healthcare professionals and participants. High explainability does not imply greater predictive performance; the balance between the two should be evaluated according to the purpose and the Context of use. Examples of use in healthcare in Brazil refer to research conducted in... Brazilian Unified Health System (SUS) and in Brazilian university hospitals, as documented in the references at the end of this document. chapter.

AI systems used in clinical research are distributed, in their most widespread form. elementary, in two large families, according to the temporal axis of the question they answer.

Diagnostic support systems answer a question about the present: what What does this data reveal about this patient's current clinical state? Risk prediction systems They answer a question about the future: what is the probability that this patient will improve? for a specific outcome?

This distinction defines the type of ethical risk the protocol carries and the type of question. which the Ethics Committee must formulate before any analysis of scientific merit.

Diagnostic support systems process input data such as medical images, laboratory profiles and clinical records, producing classifications or recommendations. probabilistic. The Federal Council of Medicine (CFM) established, through Resolution No. 2,454, of February 11, 2026, which algorithmic tools support diagnosis. They must act exclusively as support for medical decision-making (CFM, 2026). The resolution prohibits Direct communication of diagnoses to patients and maintaining full responsibility for physician regarding the final clinical act.

Risk prediction systems operate on multimodal variables such as historical data. clinical, social determinants, epidemiological data and environmental factors, producing probability scores for future outcomes.

Predictive risk models significantly enhance surveillance capabilities. epidemiological data from the Brazilian Unified Health System (SUS). However, such systems present their own ethical risks. most notably the phenomenon of *self-fulfilling prophecy*.

A self-fulfilling prophecy is characterized by a dynamic in which a prediction, when... Influencing decisions and behaviors contributes to the achievement of the desired outcome: Expectation becomes reality.

For example, a patient classified by an algorithm as being at high risk of Those who drop out of therapy receive less follow-up care and, therefore, abandon treatment. confirming his own prediction. The predicted outcome is now occurring as a result of the decision. guided by the model, and not by inevitable clinical evolution. Therefore, the system contributes to produce the result that it announced.

Diagnostic support and predictive systems share a structural risk that the CEP You should check for differential performance between subgroups in any protocol. population-based systems. Soares and Chiavegatto Filho (2026) demonstrated that systems trained with data predominantly from specific ethnic, gender, or socioeconomic groups They produce worse results for populations that are underrepresented in the training base. One A system validated in a leading university hospital may degrade when applied in... basic health unit in the interior of the Northeast or the Amazon, with equipment distinct, different epidemiological profiles and populations missing from the training data (Fiske *et al.*, 2025).

This challenge of performance and representation also extends across the systems that They operate on unstructured clinical language. Natural language processing. (Natural Language Processing — NLP) stems from a given paradox in medicine: up to 80% of Clinically relevant information resides in free text, not in structured fields. Database. Nursing progress notes, narrative histories, reports. Anatomopathological data and family histories carry knowledge that no spreadsheet can contain. capture. NLP converts these narratives into data that can be analyzed by algorithms, through Techniques for tokenization, lemmatization, and vector representations of words.

The main ethical challenge of natural language processing is not related to technology. The free text of medical records contains many indirect identifiers, such as unusual professions, unique accident dates, and unusual family structures common. Through reverse inference and open-base triangulation, these fragments can Re-identifying the participant without their consent, even after suppressing their name and CPF (Brazilian taxpayer ID).

LLMs expand the power of NLP and expand its risks proportionally. Unlike conventional models, which classify or extract targeted information, LLMs generate answers, produce reports, formulate hypotheses, and synthesize literature with fluency that simulates technical authority. Two specific risks emerge strongly in literature.

The first risk is latent memorization: if researchers introduce texts Identifiable physicians in the interaction windows with cloud-based LLMs, the architecture can absorbing this confidential data and unintentionally sharing it with other users. (Figueiredo; Aguiar, 2025).

The second is hallucination : the model fabricates scientific citations. nonexistent, recommends clinically inappropriate dosages, or describes findings. fictitious laboratory devices disguised as evidence.

CNPq Ordinance No. 2,664, of March 6, 2026, established that responsibility for Clinical inaccuracies, plagiarism, or fabricated data generated by LLMs in research falls under... entirely about human researchers, preventing the concealment of authorship by systems. computational (CNPq, 2026).

For a CEP (Brazilian Postal and Research Center), this regulation has a direct implication: protocols that utilize LLMs (Laboratory Management Languages) in any phase of the research must declare this use and describe the mechanisms of Human verification of the generated results. Generative systems constitute a

A distinct category within AI in healthcare. Generative adversarial networks (Generative Adversarial Networks (GANs), diffusion models, and LLMs can create medical images. synthetic, evolving clinical text, pharmacological hypotheses and tabular data entirely artificial.

The generation of synthetic data holds the promise of escaping severe constraints. Privacy concerns, offering voluminous material without any connection to real individuals. The Agency The National Data Protection Authority (ANPD) warned, through a technical note, that the uniqueness The individual cannot be underestimated: generative models can replicate fragments. idiosyncratic characteristics of the real subjects that served as the basis, circumventing theoretical anonymity and violating the General Data Protection Law (LGPD) if reverse engineering reveals the original patients (Brazil, 2024d).

The additional risk is epistemic substitution: synthetic data used as a substitute. Real data without rigorous validation of statistical reliability function as a chamber of echo, amplifying the inequalities of the original base and creating the risk of manufacturing evidence with a mathematical appearance of legitimacy.

A common risk factor among PLN, LLMs, and generative systems is bias. Linguistic issues arising from global asymmetry. Most available libraries and models She was trained predominantly in English, capturing Anglo-Saxon cultural nuances. When applied to data in Brazilian Portuguese, these models distort meanings. They fail to process regional medical jargon and make systematic errors in... reading standardized abbreviations of the SUS (Soares; Chiavegatto Filho, 2026).

The Soberania Project, launched by the Ministry of Science, Technology and Innovation (MCTI) In partnership with the Artificial Intelligence Secretariat of Piauí in 2025, it represents the answer. structural solution to this problem: a foundational model trained on 150 billion tokens of public scope in the Portuguese language, with sensitivity to regional accents and nuances. sociocultural aspects of the Brazilian territory (Santos *et al.*, 2025).

For the CEP, the question about the model's training language constitutes a question of scientific validity and fairness for the population studied.

The Brazilian response to this scenario does not, in itself, eliminate the challenges associated with it. Reuse of previously trained models in other contexts. The transfer of *Transfer* learning solves a practical problem that affects most people. Brazilian public health institutions: the prohibitive cost of training an AI model of

Zero. A pre-trained model accumulates general knowledge by processing terabytes of data. in large supercomputers.

The researcher uses the existing architecture and quickly trains it with their own... data, adjusting the model to the research context in just a few days and at a reduced cost. For laboratories at peripheral public universities with underfunded budgets, This approach equalizes access to cutting-edge architecture worldwide.

The risk lies precisely in this apparent shortcut: the imported model carries biases. of the populations with which it was originally trained, predominantly European and Americans. A researcher from the Brazilian Unified Health System (SUS) who performs learning transfer to The diagnosis of skin lesions inherits the insensitivity of the original model to manifestations in Black and brown patients, crystallizing silent racial inequities without the protocol acknowledging them. declare (Soares; Chiavegatto Filho, 2026). In the protocol analysis, the essential question about Pre-trained models are about the demographic background of the original training base and about The statistical power of the fine-tuning phase to correct these deficiencies in the Brazilian context.

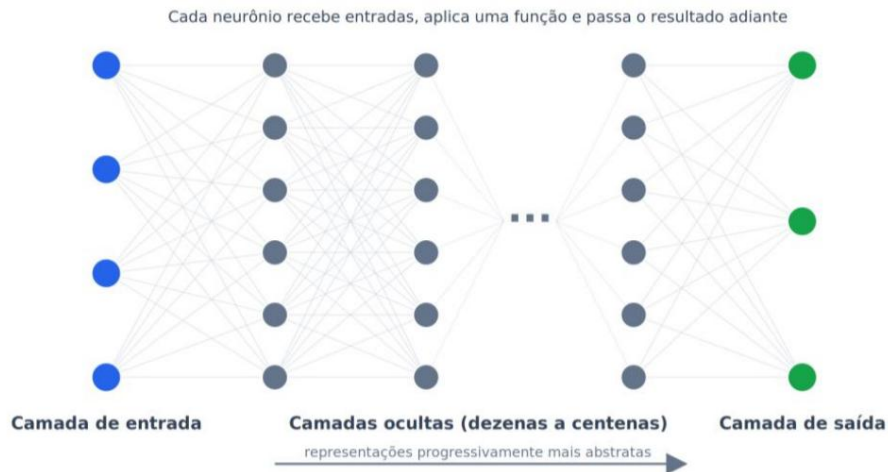
Nevertheless, the transfer of learning only partially solves the problem. The environmental impact that training large models imposes. The fine-tuning of a pre-model The trained training consumes a minimal fraction of the energy required by the original training. This reinforces, From a sustainability perspective, the choice for reusable architectures instead of Models built from scratch.

The persistent environmental cost is that of continuous inference: the use of the model after the Training accumulates consumption proportional to the volume of queries, and systems are operating. in international cloud infrastructure with a carbon-intensive energy matrix They transfer this impact to territories far from the beneficiaries of the research. The Guide to Brazilian user of Artificial Intelligence acknowledges the environmental impact of the systems. AI constitutes a relevant ethical dimension (Brazil, 2026c). For a CEP, the question is not only It's not about what the system does, but where it operates, with what infrastructure, and what its matrix is. Energy. This dimension rarely appears in the protocols. It is part of the risk analysis of a protocol committed to equity between the generations that inhabit the same territory that the SUS serves.

Deep Neural Networks (DNNs) are structured in Dozens or hundreds of consecutive layers of artificial neurons. Each neuron receives Each input is taken, a mathematical function is applied, and the result is passed on (Figure 2.2).

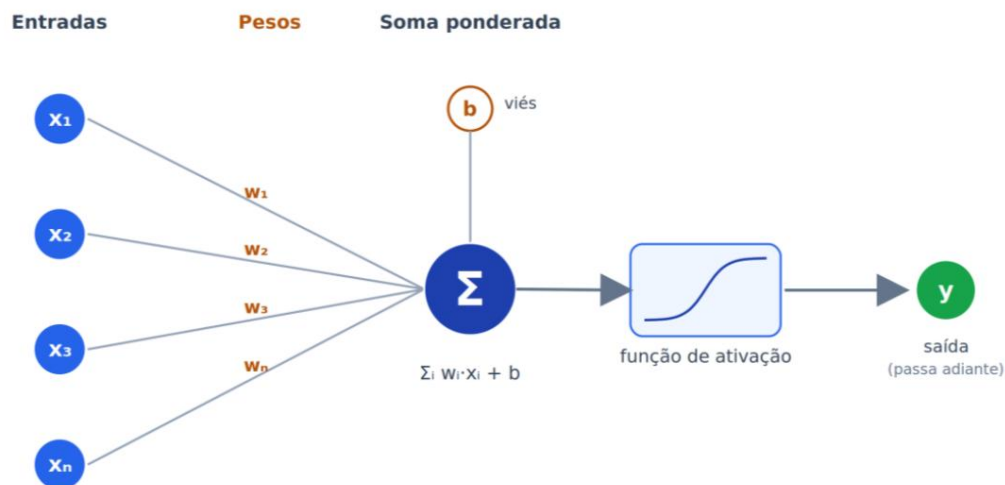
The layer learns progressively more abstract representations of the input data.
(Figure 2.3).

Figure 2.2 — Architecture of a deep neural network: layers of artificial neurons



Source: author's own elaboration.

Figure 2.3 — Internal workings of an artificial neuron in a deep neural network



Source: author's own elaboration.

Three architectures dominate health research: convolutional networks (Convolutional Neural Networks — CNNs) for medical image classification, the networks Recurrent Neural Networks (RNNs) for time series such as Continuous electrocardiograms, and transformers for deep mining in extensive medical records (Santana; Moreno; Santos, 2025).

The depth of DNNs ensures high precision, but makes them difficult to explain. functioning like black boxes. Another risk is overfitting: networks trained with too few Data tends to retain noise, exhibiting high accuracy in development but failing Given the real diversity (Youssef *et al.*, 2024). The CEP that receives a protocol with DNN you should check if there are documented cross-validation metrics in separate sets and if the protocol provides for explainable artificial intelligence (XAI) techniques for to make the model's decisions contestable by the healthcare professionals involved.

Lightweight models address a real need of the Brazilian public health system (SUS) : Running AI systems on devices with limited computing power and no internet connection. Stable with cloud servers. Model compression techniques, distillation of Knowledge *distillation* and quantization reduce size and cost. The computational efficiency of DNNs is achieved without critical performance sacrifice.

The specific ethical risk of lightweight models is selective degradation: the processes of Compression can reduce performance unevenly among population subgroups. harming precisely the atypical populations that the SUS (Brazilian Public Health System) most needs to reach. The protocol It should be documented that the compression was audited to preserve performance fairness. between groups.

Extreme Learning Machines (ELMs)
They represent a radically different approach from DNNs. Instead of the iterative process In backpropagation adjustment, the ELM randomly selects the weights of the hidden layer, locking These weights are applied, and the result is solved by direct matrix inversion. The gain is significant: training speed incomparably superior to that of DNNs, at a lower cost. minimum computational.

The central ethical risk of ELM is stochastic instability: because it depends on weights. Due to randomness, the model may produce fluctuating results when trained repeatedly. using the same data, thus violating the principle of scientific reproducibility. In sets of With small or unbalanced data points, this instability worsens. The CEP (Centers for Ethics and Professional Development) should verify if the... The protocol documents repeat tests with attestation of the consistency of the classifications and whether The database volume is sufficient to support the decision-making power of the architecture. proposal.

The type of AI system determines the type of ethical risk and the type of question the Ethics Committee (CEP) asks. must formulate. This statement, which runs through all the movements in this section, requires a

Practical implementation: how does the committee identify the type of system when reading a protocol? MRCT Center and WCG (2025) propose that committees classify systems not by disease studied, but due to the stage of development of the tool at the time of Research. In the discovery phase, the system is still under construction and the data is used to train the architecture. In the translation phase, the system is tested under real-world conditions by first time. In the initial implementation phase, the system operates in an interventional manner on Clinical routines. Each stage imposes distinct ethical obligations and requires oversight from the Ethics Committee. different dimensions.

Barucci *et al.* (2026), from a European perspective on European Union Regulation Regarding the Artificial Intelligence Act (AI Act) for ethics committees, they add a dimension. Relevant regulatory issue for Brazil. Regulation (EU) 2024/1689 classifies it as high risk. all AI systems used in a clinical context as medical devices or integrated systems to existing medical devices.

In Brazil, RDC No. 657, of March 24, 2022, classifies substances in this category. computer programs for diagnostic, preventive, monitoring or treatment of disease (Anvisa, 2022). For the Brazilian CEP, this articulation has Practical consequence: before evaluating the scientific merit of the protocol, the committee must determine if the proposed system qualifies as a computer program as Medical device (Software as a Medical Device — SaMD) and what is its risk class? This classification precedes any performance analysis.

This chapter concludes here. The reader who has completed it will be able to identify the... The type of system in front of you, you know the database that feeds it and the family of algorithms. that drives it. What do these technical choices produce when the system encounters research participants — bias, opacity, failed oversight, the life cycle that It doesn't end here — that's the subject of the next chapter.

3 Critical Properties of AI Systems: Bias, Explainability, Supervision and LIFE CYCLE

The concepts from the previous chapter describe what artificial intelligence systems are. What they are and how they work. This chapter discusses what they produce when they enter a Research involving human subjects. Four properties focus on ethical issues. Bias, which is not a programming flaw and occurs at three distinct moments. The explainability, which needs to work for different audiences. Human supervision, which fails for reasons predictable. And the life cycle, which doesn't end when the research ends. Before that, the section 3.1 presents the five dimensions that grade the requirement regarding the protocol and permeate the four.

3.1 THE FIVE DIMENSIONS OF PROPORTIONAL RISK

Law No. 14,874, of May 28, 2024, and Decree No. 12,651, of October 7, In 2025, they consolidated in Brazil the gradation of requirements and procedures in proportion to risk. from the research (Brazil, 2024a; 2025a). The ethical principles of evaluation do not vary with risk. What varies is the stringency of the protocol's requirements and the designated body for review. lo. Applied to AI systems, this gradation is organized into five dimensions that the CEP should Go through it when reading any protocol.

The first is the nature and sensitivity of the data processed: systems that operate Texts containing free medical records or biometric images require more rigorous scrutiny. which systems process unidentifiable tabular variables.

The second is the degree of automation and the clinical consequence of the decision: a system that Producing recommendations subject to mandatory human review carries a distinct risk from a A system whose outputs are directly incorporated into conduct protocols.

The third is the generative or classificatory profile of the architecture, given that systems Those capable of creating new content, such as synthetic data or generated reports, present risks. hallucinations and fabricated evidence, absent in conventional classifiers.

The fourth dimension is the plasticity of the parameters: the system operates with weights. Frozen after training, or does it continue learning during research? Systems with Continuous or adaptive learning introduces instability that can alter the The model's behavior throughout the study compromised informed consent.

data based on system characteristics that may not be the same at the end of the process.
search.

The MRCT Center and the WCG (2025) recommend that protocols with systems
Adaptive measures define predetermined thresholds for change that trigger ethical review.
mandatory. The fifth dimension is the level of explainability of the architecture: systems with
Explainable AI capabilities allow healthcare professionals to identify the basis of
algorithmic decision-making, challenging results, and documenting clinical reasoning in a way that
auditable (Santana; Moreno; Santos, 2025; Youssef *et al.*, 2024).

These five dimensions require precision in reading the protocol and a willingness to...
to formulate the questions that the researcher, immersed in scientific enthusiasm, may not have considered.
Done. The system classification precedes any performance evaluation. The CEP that
evaluates how well the system works before asking what it is and how it decides to approve.
A promise, not research.

3.2 Algorithmic Bias: What it is, how it arises, and why it persists

Algorithmic bias results from the structure of the data used to train the model.
Not due to technical errors. A machine learning algorithm does not have access to the world;
has access to the records that the world has produced.

These records carry decades of human decisions about who received care.
who was documented, who received an accurate diagnosis, and who was underreported.
Learning patterns, the model also learns absences and distortions. According to Barucci *et al.*
(2026), the main factors of opacity in AI are the reliance on historical data with
implicit, subtle, systemic biases that are difficult to detect.

Opacity sets in at three distinct points in the development cycle.
These are: 1) At the time of collection, when the data available for training is not
They represent the diversity of the population that will be affected by the system. 2) At the time of
design, when choices are made about which variables to include, which outcomes to predict, and which
The performance metrics adopted incorporate unstated assumptions about who
What matters and what counts as a relevant outcome. 3) At the time of evaluation, when the
The model's performance is reported as an overall average, hiding serious discrepancies in...
specific subgroups.

The absence of differential performance analysis by population subgroup is a One of the most frequent ethical gaps in clinical trial protocols involving AI is precisely... because good average performance masks flaws concentrated in the smaller populations. vulnerable (Youssef *et al.*, 2024).

The geographic origin of training data is the primary vector of bias in AI systems. applied to global health. The World Health Organization (WHO) has documented that most One of the large models available for clinical use was trained predominantly with data. from high-income countries, with predominantly white and European populations (WHO, 2024).

Algorithmic bias therefore has three overlapping layers that the CEP needs to... distinguish. The first is technical: the model fails measurably in a subgroup because This subgroup was underrepresented in the training data. The second is structural: the model It reproduces hierarchies of race, class, and territory as if they were natural regularities. because historical data encodes them in that way. The third is political: the absence of The requirement for equity analysis in research protocols is not a neutral omission. Choice determines who receives ethical protection and who suffers from system failures. recognized.

Technical bias and structural injustice are distinct concepts. The Ethics Committee must differentiate between them. To act precisely. Technical bias is measurable: the model produces error. statistically superior in one subgroup compared to another, and this error can be identified. quantified and reduced through methodological intervention.

Structural injustice operates on a different scale. When the model codifies it as the norm. It's normal for a reality built upon historical exclusion; he not only errs more in... certain groups. It reproduces as an objective fact the hierarchy that generated the exclusion.

Rajkomar *et al.* (2018) empirically demonstrated the central mechanism: models Predictive technologies with high overall performance may mask serious disparities in subgroups. specifics. A model with 90% overall accuracy may have 70% accuracy for the group. Underrepresented in training. The overall performance report does not reveal this data. Only Stratified analysis by subgroup makes it visible.

Fiske *et al.* (2025) advance in this direction by demonstrating that the problem is not The solution is to simply include race and ethnicity variables in the models. When these variables They are incorporated without adequate historical and social contextualization; the model may reinforce

spurious associations between identity markers and clinical outcomes, producing
Algorithmic discrimination disguised as scientific evidence.

The MRCT Center and the WCG (2025) posit that ethics committees should verify whether the
protocols document differential performance analyses by subgroup before any
Approval, because unverified bias is not nonexistent bias.

In Brazil, this distinction between technical bias and structural injustice has consequences.
Direct practice. The protocol that presents only global performance metrics, without
stratified analysis by race, gender, age group, geographic region, and condition.
socioeconomic, it offers no evidence of equity. It offers an absence of evidence of
iniquity.

In the Brazilian Unified Health System (SUS), which serves populations with epidemiological, ethnic, and regional diversity that
No private database replicates this; approving a protocol without this analysis is tantamount to approving it.
a system whose actual behavior for part of the studied population remains
unknown. Bill No. 2,338/2023 (Brazil, 2023), whose purpose is the regulation
Regarding AI in Brazil, it explicitly acknowledges the risk of algorithmic discrimination in services.
public and demanded independent audits for high-risk systems.

The requirement for stratified performance analysis is the ethical operationalization of this.
principle in the evaluation of research protocols. Soares and Chiavegatto Filho (2026) state
that the uncritical incorporation of AI into the SUS (Brazilian Public Health System), without robust local validation, risks reinforcing
A technocratic and exclusionary model that deepens pre-existing inequalities.

The Pan American Health Organization (PAHO/WHO, 2024) highlights that low-income countries
The media face a specific challenge: the models available for adoption were
developed in contexts of superior infrastructure, and their actual performance in
older equipment and staffed by professionals with different training remain
often unknown.

The data from the SUS (Brazilian Unified Health System) allows for the development of genuinely representative models of...
Brazilian population. The use of this data without adequate safeguards, however, transforms the
The diversity of the Brazilian Unified Health System (SUS) is being used as raw material for systems that will never restore equity to...
populations that generated them. Santos *et al.* (2025) identify this risk as colonialism.
Internal digitalization: extracting data from the most vulnerable populations to feed systems.
which primarily serve those who already have access.

This set of risks translates into concrete demands regarding how the data are described and justified in the protocol.

The first requirement is documentation of the demographic composition of the training base. The document should describe the origin of the data used to train the model, including the distribution by race, gender, age group, geographic region, and socioeconomic status, and compare this distribution with the profile of the target population of the research. When the base of training comes from pre-trained foreign models; the documentation must identify the geographic and ethnic origin of the original data and describe how the fine-tuning phase addressed the differences in relation to the Brazilian population. Protocols that describe only the size of the training base, without disclosing its demographic composition, provides the CEP (Brazilian Postal Code). Insufficient information to assess the risk of bias (Youssef *et al.*, 2024).

The second requirement is the analysis of differential performance by subgroup. It is expected... that the protocol presents, or commits to producing during the research, metrics of performance stratified by population groups relevant to the clinical context. Performance stratified by population groups relevant to the clinical context. Global metrics of accuracy, sensitivity, and specificity are insufficient when the model will be applied to heterogeneous populations.

The South Korean Ministry of Food and Drug Safety published, in 2022, specific guidelines for designing clinical trials with AI were released, instructing... researchers and ethics committees to verify three dimensions missing in the protocols of conventional medicines: data leakage prevention, auditing for detecting hidden population biases *and* methodologies to mitigate them. Performance variations between different hospital units (MFDS, 2022). The CEP Brazilians can and should adopt an equivalent requirement, proportional to the clinical risk of the application. proposal.

The third requirement is a plan for monitoring equity during and after the research. Bias is not a static property of the model. It can emerge or expand over time as the population served diverges from the training profile. The MRCT Center and the WCG (2025) recommend that AI protocols define predetermined criteria for performance degradation that triggers mandatory ethical review, including criteria specific to disparities between subgroups.

3.3 EXPLANABILITY AND TRANSPARENCY

Explainability corresponds to the ability to produce, for each relevant decision, an understandable narrative about the factors that determined it, without requiring full access to the source code of the AI system.

Santana, Moreno and Santos (2025) define explainable AI as the ability of these systems to provide intelligible explanations about their decisions and internal processes, making the workings of algorithms transparent to users and creators. This definition carries an operational distinction that the CEP needs to internalize: explaining to an engineer is different from explaining it to a doctor, and explaining it to a doctor is different from explaining it to a patient. Each level has distinct requirements and serves ethical purposes. distinct.

Three levels of explainability structure this spectrum. The first is explainability. Technical description: the model's internal weights, gradients, and activations for engineers. Machine learning. This level allows you to audit the architecture and identify layers. problematic and propose corrections to the model.

The second is functional explainability: identifying the most influential variables in decision-making for data scientists and clinical researchers. Tools like SHapley Additive exPlanations (SHAP) e Local Interpretable Model-agnostic Explanations (LIME) They operate at this level, producing maps that indicate which characteristics of the data... The inputs contributed most to each specific output.

The third aspect is clinical explainability: translating the logic of the decision into language. Understandable for healthcare professionals, but contestable by research participants. Santana, Moreno and Santos (2025) document applications of explainable AI in the diagnosis of COVID-19, leprosy, and retinoblastoma precisely because the clinical adoption of these The tools depended on the doctor's ability to understand, verify, and question the algorithmic reasoning.

CEP focuses its evaluation on the third level, aimed at ensuring explainability. The analysis does not focus on the code or the internal workings of the model. The central question revolves around the ability to justify decisions in a comprehensible way. produced by the system.

In assessing explainability, CEP starts with three concrete checks. The first: Does the protocol describe which explainability technique will be adopted and for which audience? A technique suitable for engineers may be useless for doctors. Secondly: explainability. is it planned as a mandatory component of the system or as an optional feature available? Who should we ask? Systems where explainability is optional replicate in practice the The black box problem, because productivity pressure in the clinical environment tends to suppress the use of optional features.

Thirdly, the explanations produced by the system were validated with professionals. What is the health context where the system will be used? A valid explanation for a specialist. The role of oncology in a tertiary hospital may be incomprehensible to a general practitioner. in a basic health unit in the interior of the Legal Amazon. The MRCT Center and the WCG (2025) They establish that ethics committees must verify whether the limitations of the system and its Potential variability was adequately communicated to physicians, participants, and regulators.

The conflict between performance and explainability is a result of the structures of Deep learning, not engineering failures. Santana, Moreno and Santos (2025) Empirically, they document that deep neural networks with high clinical performance... They tend to operate with less explainability than simpler models, such as trees. Decision making and logistic regressions.

More precise models tend to operate like boxes that are difficult to interpret (Salih *et al.*) (*al.*, 2023). DNNs achieve high performance by identifying complex patterns that escape to direct reading by the clinician. Conversely, models such as decision trees allow To trace the logical path of the prediction with greater transparency, even if with less... accuracy. This tension doesn't arise by chance. It results from the ability of neural networks to to construct abstract representations in multiple dimensions, far removed from clinical concepts. traditional.

Figure 3.1 represents this spectrum with the algorithms most used in search of health in Brazil, positioned according to their predictive performance and their level of explainability. On the left pole, architectures such as deep neural networks with wavelets, Convolutional networks and computational reservoir networks focus on high performance and low explainability.

Figure 3.1 — Indicative relationship between performance and explainability of the main algorithms in health research.



Source: own elaboration, based on Barbosa *et al.* (2021a, 2021b), Salih *et al.* (2023), Santana, Moreno and Santos (2025) and Lima *et al.* (2026).

On the right side, models such as decision trees, Bayesian networks, and logistic regression. They offer high explainability with comparatively more modest performance. The model The ideal, which combines high performance with explainability, remains an objective outside of this context. within the scope of current architectures. The positioning is indicative and may vary according to the data volume, clinical context and configuration choices (Barbosa *et al.*, 2021a, 2021b; Salih *et al.*, 2023; Santana; Moreno; Santos, 2025; Santos *et al.*, 2026).

For a CEP (Ethics Committee), this tension has concrete implications for protocol evaluation. researcher proposing a DNN for clinical diagnosis in research with living beings. Humans are implicitly accepting a degree of opacity. This acceptance needs to be declared and justified.

The question the committee should ask is precise: the performance gain in relation to More explainable models ethically justify reducing transparency for the participant.

And what about the attending physician?

This question requires the protocol to present a performance comparison between the proposed architecture and more interpretable alternatives, with explicit analysis of the exchange between Accuracy and explainability (Barucci *et al.*, 2026). Protocols that adopt neural networks. Deep interpretations without such comparison assume that superior performance justifies opacity.

The literature documents techniques that mitigate, without eliminating, the opacity of DNNs.

Heatmaps and *saliency maps* identify regions within an image.

medical factors that most influenced the model's classification, allowing the physician to verify whether the system focused on anatomically relevant structures or irrelevant artifacts.

Gradient-weighted class activation mapping (Grad-CAM) and SHAP operate in this register.

Translating internal gradients into interpretable visual representations.

These techniques also have limitations that need to be made explicit. Maps of

The salience indicates the regions of greatest attention in the model. The interpretation of the reasons for this

The focus and its causal relevance require further analysis.

A map highlighting the lung region on a chest x-ray does not demonstrate

that the model learned lung anatomy. It demonstrates that this region was

statistically correlated with the outcome in the training data. The difference between

Statistical correlation and causal relevance are the areas where serious clinical errors occur.

without the system signaling any anomaly. In practice, the presence of techniques of

Explainability in a protocol does not guarantee that the system is clinically sound.

Interpretable. It is equivalent to an instrument that makes interpretation possible. Supervision.

Significant human input regarding this instrument remains the responsibility of the researcher and

Committee requirement.

Three additional guarantees structure the explainability requirement in the approval of any protocol with AI.

The first is that the system produces, for each decision with clinical consequences on

The participant needs a justification that is understandable to the responsible healthcare professional. This

The justification does not need to describe the model's internal weights. It needs to answer, in

In clinical language, the question a doctor might ask a colleague is: why this result, for

This patient, with this data?

The second is that this justification can be challenged and overturned by a court ruling.

documented human. In the context of research, it is the protocol, under the responsibility of

researcher, who describes the flow by which the system's output can be challenged and

superimposed. When the system is intended for clinical use, the resulting clinical act observes the

full responsibility of the physician as stipulated in CFM Resolution No. 2,454, of February 11

of 2026 (CFM, 2026). For this oversight to be real, there needs to be a basis for contestation. One

A system that produces outputs without justification offers a result that can be accepted or rejected.

arbitrary. The protocol should describe how the responsible professional documents their
Disagreement is recorded, and the alternative judgment is registered. This flow is a condition for supervision.
human being is auditable.

The third is that participants should be informed, in the informed consent form, about the level of
The explainability of the system and what that means for your right to review the decision.
Article 20 of the LGPD guarantees the data subject the right to request a review of automated decisions.

Fornasier (2022) demonstrated that this right remains abstract when the
The participant is not informed about the degree of opacity of the system affecting them. The Informed Consent Form of
A protocol with DNN should inform, in accessible language, that the system produces
recommendations whose internal logic is not completely transparent, describe how the
The doctor will use this recommendation and indicate the channel through which the participant can request it.
Clarifications or revisions. Explanations that do not reach the participant do not protect their rights.
rights only create the appearance of protection.

The South Korean Ministry of Food and Drug Safety requires, in
Guidelines for AI clinical trials from 2022, which vendors provide data
Transparent information about the origin of the training data and prediction logic. Reports should be transparent.
document methodologies for mitigating performance variations between hospitals (MFDS,
(2022). This South Korean regulatory requirement is more precise than any guidance.
Current Brazilian guidelines on explainability in AI research protocols. The CEP that adopts
An equivalent requirement does not exceed his mandate. He rigorously performs the function of guardian of
dignity of participants in a field where regulation has not yet arrived with the
Required specificity.

Explainability and fairness are architectural decisions in AI, defined beforehand.
even during data collection, and not added after the system is ready. This distinction
defines the principle of ethics *by design*: ethical choices must be
incorporated into the system structure as engineering criteria, not superimposed as
a subsequent compliance layer. The principle extends to the ethical field what privacy...
From the very conception (*privacy by design*) it already establishes data protection.

Article 46 of the LGPD (Brazilian General Data Protection Law) stipulates that data processing agents must adopt measures.
safety, technical and administrative aspects, from the product or service design phase onwards.
(Brazil, 2018). Applied to AI in healthcare, this legal mandate means that the protocol of

Research should demonstrate that fundamental ethical decisions were made at the time.

from the system design, not identified as pending issues after the model was built.

The WHO has documented that the ethical governance of AI systems in healthcare is articulated in The three phases of the value chain are: design and development, and the provision of the service or product. and entry into clinical use (WHO, 2024). In each phase, different actors carry out Distinct responsibilities exist, and certain decisions can only be made in the earlier phase.

The choice of explainability, for example, needs to happen before training, because Some more opaque architectural styles are difficult or impossible to interpret. retroactively.

The decision regarding the demographic composition of the training base needs to be made before... data collection, because structural representativeness biases cannot be corrected with adjustments. subsequent to the model.

The protocol that provides for equity analysis only as a final evaluation step. It implicitly states that the data has already been collected, the model has already been trained, and the decisions have already been made. Irreversible architectural changes have already been made. The postal code that receives this protocol receives a A system that has been formed, not a project that can be evaluated.

The position of equity in the system's development cycle reveals whether the decisions Ethics were incorporated from the outset or treated as an afterthought, outside of the process. evaluative scope of the CEP.

For the Ethics Committee, the principle of ethics from conception has direct operational implications. in the reading of any protocol involving AI. The committee must verify whether the ethical decisions Structural elements were identified and documented prior to development. Barucci *et al.* (2026) propose that protocols explicitly describe architectural choices with ethical justification, including the rationale for the architectural selection, the methodologies of Mitigation of overfitting and documentation of the quality and composition of the database. A protocol that justifies these choices solely based on performance criteria, without Explicit consideration of fairness and explainability does not satisfy the ethical standard since conception. The distinction between a protocol that treats ethics as a project and a protocol that The treatment of ethics as a final checklist is evident in the sequence of sections of the document. The CEP that Learn to read this sequence; develop an evaluation tool that no metric... Performance replacement.

3.4 Human Supervision and Decision Automation

Human oversight in AI varies continuously in terms of its influence on decisions. Humans have an influence on algorithmic results. At one extreme, the system produces information that The researcher considers freely and may ignore without justification. At the opposite extreme, the The system makes decisions that the researcher ratifies out of habit or institutional pressure, or that They are executed autonomously without real human intervention. Between these extremes there exists a gradation that research protocols rarely describe precisely, but that In practice, it determines who decides about the participant (WHO, 2024).

The ability to disrupt the system is human oversight in its purest form. Specifically, the "kill switch" is not a technical safety feature, but a... Ethical requirement. A system that cannot be turned off by the principal investigator during The study is not under human supervision. It is under nominal supervision. The AI Act determines that professionals responsible for supervising high-risk systems must be able to to intervene and reverse algorithmic decisions when necessary (Barucci *et al.*, 2026). The protocol The Ethics Committee (CEP) should describe how a researcher can disagree with leaving the system and how to suspend the process. The system will be affected when performance compromises participant safety, in addition to... specify who has the authority to trigger this suspension. This ability to However, interruption only exists in practice when the researcher recognizes that it is necessary. intervene.

The WHO has identified automation bias as one of the key risks. The use of large language models and AI systems in diagnosis and clinical care. The phenomenon occurs when the healthcare professional passively accepts the patient's exit from the system. without critical scrutiny, due to work overload, overconfidence in Technology or institutional pressure for productivity (WHO, 2024). WHO documentation describes this risk as predating LLMs: it is present in any support system for clinical decision whose architecture favors acceptance over contestation. Soares and Chiavegatto Filho (2026) points out that this phenomenon is consolidating itself as a specific ethical risk. In the integration of AI into the Brazilian public health system (SUS), the workload of professionals and the scarcity of resources create... Favorable conditions for algorithmic support to transform into effective delegation.

The type of system and the clinical context determine where each protocol falls on the spectrum. It is situated in practice. A triage system that produces a risk score on a difficult-to-read screen.

interpretation, in a service environment with long queues and reduced time due to
The patient's approach is functionally closer to automation than to recommendation.
regardless of how the protocol describes it. When the system is intended to support
Regarding clinical decision-making, the requirement of CFM Resolution No. 2,454/2026 applies, stating that the tool
Support only the professional's decision, without communicating diagnoses directly to the patients.
patients, and that responsibility for the clinical act remains complete (CFM, 2026).
In research, supervision is the responsibility of both the protocol and the researcher. The protocol must prove...
that the system achieves this goal in real-world use, not just in theory.

In the context of CEP, the main issue regarding human supervision does not refer to
The position stated by the researcher is not the one defined by the system's design and elaboration.
of the outputs, the institutional conditions of use, and the productivity demands of the environment.
Clinical factors tend to favor this. Barucci *et al.* (2026) identified that the physician's ability to
Blocking or overriding algorithmic decision-making is a fundamental criterion in regulatory evaluation.
South Korean clinical AI systems. The MRCT Center and WCG (2025) recommend that the
The protocols describe the processes that will be available for researchers.
Identify and respond to adverse events related to system output. Both
The demands stem from the same premise: human supervision that was not designed to be...
The option exercised will not be exercised.

Meaningful human supervision requires three simultaneous conditions. The first: the
The researcher understands the basis of algorithmic decision-making with sufficient depth to
to contest it. The second: has the time and institutional conditions to exercise this.
Scrutiny before acting. The third: the system is designed to facilitate, not inhibit, the
Disagreement. When any of these conditions fail, supervision becomes nominal.
Fornasier (2022) demonstrated that the absence of mandatory real human intervention
The review of automated decisions under the LGPD (Brazilian General Data Protection Law) opens up space for a reduction in oversight.
a formal gesture that does not alter the result produced by the algorithm.

The first condition depends on the explainability discussed in the previous section. A system
that which does not produce a justification comprehensible to the researcher does not offer a basis for
Dispute. The researcher who disagrees with the algorithmic output without understanding why the system...
having reached that result, it does not exercise supervision, it issues a competing opinion without basis.
Dialogical with the model. The WHO establishes that explainability and intelligibility are

conditions for human supervision to be real, not just nominal, precisely because

Without them, the professional does not have access to the reasoning that they should be supervising (WHO, 2024).

The second condition depends on the institutional context, not the protocol. A doctor
eight-minute consultations at a private clinic that accepts popular health insurance plans.
does not have the same conditions to exercise scrutiny over an algorithmic output as a
A specialist in a thirty-minute elective consultation. The research protocol does not control the
Time available per patient in the usage environment. However, it can describe for what purpose.
In a clinical context, the system was designed and validated, and it should be made explicit that its use in such contexts...
Reduced supervisory capacity requires additional safeguards.

The third condition is the one most directly controllable by the system design.
Systems that present their outputs as definitive results rather than recommendations.
Graduated systems inhibit disagreement through linguistic formulation. Systems that do not record the
Disagreements between the researcher and the algorithmic output eliminate the audit trail.
which would make the supervision verifiable. Systems that require additional steps to record.
Disagreement regarding simplified steps for acceptance creates an asymmetry of effort that
It favors passive acceptance. Barucci *et al.* (2026) propose that protocols describe
explicitly the mechanism by which the doctor can block or override the decision.
algorithmic, and that this mechanism is auditable.

Agentic AI represents the most advanced point on this spectrum.
Unlike conventional systems, which produce a single output for a single
At the input stage, AI agents plan sequences of actions and utilize external tools.
They execute sequential steps and make intermediate decisions without human review between them.
each step. An agent can, in a single autonomous cycle, consult records, cross-reference
Using information from medical literature, formulating diagnostic hypotheses, and proposing courses of action.
Therapeutics before any researcher has seen the case. The point of human supervision.
It shifts. Instead of reviewing each result, the professional approves or rejects a plan.
complete product produced by a process not monitored step by step. When the system
Even if the agent is intended for clinical use, full responsibility for the clinical act remains with the agent.
from the physician, in accordance with CFM Resolution No. 2,454/2026 (CFM, 2026). At any stage of
In research, exercising supervision requires that the protocol, under the researcher's responsibility,
Describe not only what the system produces, but also how each autonomous step was defined.
monitored and made contestable. A CEP that receives a protocol with agentic components.

It assesses risks that increase with autonomy and requires careful human supervision.
designed.

In practice, this set of requirements translates into objective evaluation criteria.
that guide the actions of the Ethics Committee.

Therefore, the Ethics Committee should formulate four questions about any protocol involving
algorithmic decisions with clinical consequences for participants. Each question addresses
a distinct dimension of human supervision and cannot be replaced by others.

The first question concerns the formulation of the system output. The protocol describes
How is the algorithmic output presented to the researcher? A system that displays "diagnosis:
"Pneumonia" operates differently from a system that displays "probability of pneumonia: 78%,
based on infiltration in the right lower lobe." The linguistic formulation of the output influences
directly influences the researcher's behavior. The WHO documents that the way in which...
The way health information is presented significantly affects clinical decisions.
as well as the content of this information (WHO, 2024). The protocol that does not describe the
The system output presentation interface does not provide the CEP with sufficient information to
Assess the risk of automation bias.

The second question concerns the time available for review. The protocol describes the
The clinical context of system use is sufficiently precise for the Ethics Committee to assess whether the...
Do institutional conditions allow for real-world supervision? A system validated in a service of
Radiology, with an average reporting time of twenty minutes, can produce a dynamic of supervision.
completely different when applied in an emergency room triage with average time.
Three minutes per patient.

The third question is about recording disagreements. The system documents...
automatically generates algorithmic outputs and final clinical decisions, allowing for the identification of...
Discrepancies and building an audit trail? Without this record, the declared human supervision
The protocol becomes unverifiable.

The MRCT Center and WCG (2025) advise that protocols should describe the
processes that will be available for researchers to identify and respond to
Adverse events related to system performance. The record of disagreements is the...
A minimum requirement for these processes to function: without it, no adverse event.
The algorithmic output can be tracked, attributed, or analyzed.

The fourth question concerns informed consent regarding the system's role in decision chain. Participants were informed, in the informed consent form, that clinical decisions regarding Your care during the research will be supported by an AI system, to what degree? What is the influence of this system and how is human supervision ensured? Law No. Law 14.874/2024 establishes that informed consent must cover the procedures that will be adopted in the research (Brazil, 2024a). An AI system that influences decisions diagnostic or therapeutic procedures on the participant are relevant to the consent. A protocol that does not mention it in the informed consent form results in incomplete consent. The Ethics Committee that accepts this Informed Consent Form approves a study in which the participants do not know how the decisions are made about them.

To the four points above, a fifth question is added, stemming from a threat. Contemporary to human supervision: prompt injection . This refers to... unauthorized instruction that infiltrates the inputs of a large-scale system. Language models and subvert expected behavior (OWASP Foundation, 2025). A The direct form comes from the user themselves; the indirect form is embedded in data that the The system processes data such as medical records, pages consulted in automated searches, or articles. members of AI-assisted literature reviews (Greshake *et al.*, 2023). In research in In healthcare, this vector can compromise the confidentiality of participant data, falsify information. summarizations and qualitative coding, and circumventing the automated supervision itself. The CEP you should ask, in protocols involving large language models with access to External sources or sensitive data, if the researcher describes the safeguards in place. against direct and indirect prompt injection, if there is a clear separation between system instructions. and processed content, and whether there is monitoring of anomalous outputs with human review. periodic.

3.5 The AI Lifecycle in Research: From Collection to Disposal

The lifecycle of an AI system in healthcare research does not coincide with the cycle conventional clinical research. The research has a defined beginning, middle, and end in the protocol. The AI system goes through phases with distinct ethical obligations, some prior to the protocol. to the recruitment of the first participants and others subsequent to the formal closing of the study.

Six phases structure this cycle. The first is the collection and preparation of data: the data Clinicians are gathered, organized, given pseudonyms, and divided into training groups. Validation and testing. Decisions regarding demographic representativeness, inclusion criteria, and Exclusion and preprocessing methods are implemented at this stage and are irreversible in later stages. subsequent phases.

The second step is training the model: the algorithm processes the training data and adjusts its... internal parameters. The participant whose data feeds into this phase is not present in The process is mandatory, but you have the right to be informed about it during the consent process.

The third step is validation and testing: the model's performance is evaluated on the datasets. reserved, with analysis of general metrics and stratified by subgroup. Barucci *et al.* (2026) They establish that this phase corresponds to testing under real-world conditions before the device is put into use. with regulatory and ethical obligations distinct from the training phase.

The fourth phase is the entry into use in research: the system begins to produce outputs that influence decisions about real participants. Human supervision, explainability, and Documenting disagreements becomes an operational requirement, not just a statement. intention.

The fifth phase is monitoring during the study: the system's performance is... monitored over time to detect data drift, accuracy degradation, or The emergence of disparities between subgroups that were not visible in the validation phase.

The MRCT Center and the WCG (2025) recommend that protocols define thresholds predetermined changes in performance that trigger mandatory ethical review, including the possibility of reassessment by the Ethics Committee.

The sixth phase is the discontinuation or transfer of the system: the model is withdrawn from Use at the end of the research, transferred to another institution, licensed, or published in a repository. open. Each of these trajectories has ethical implications for the participants whose data They formed the model, and it is for the populations that the system will continue to affect.

Understanding these six phases is a necessary condition for the Ethics Committee to assess whether the Does the submitted protocol cover the complete system lifecycle or only the usage phase? research. A protocol that accurately describes the training and validation, but is silent on The model's fate after termination describes half of the cycle.

The formal closure of a research project does not end the life cycle of the model that... she produced it. When the study ends, the AI system remains. The data that the

The elements they fed remain incorporated into their mathematical parameters, distributed as diffuse and mathematically inseparable form due to the model's architecture, as discussed in Section 7.1 of this guide. The participant who signed the Informed Consent Form consented to a specific research study. in a specific context, with a stated purpose. He did not consent to what happens to model after the research ends.

The possible fates of the model after the research concludes are numerous and Each carries distinct ethical implications. The model can be transferred to an institution. national or foreign partner. Can be licensed to a technology company. Can be Published in an open repository of AI models, making it accessible to anyone. researcher or developer in the world. It can be incorporated into a commercial product that It will be offered to the market. It can be used as a base model for transfer of Learning opportunities for future research on populations that have never participated in the research. original. None of these destinations are ethically neutral. They all prolong the lives of Participant data in systems that operate beyond the scope of original consent. (MRCT Center; WCG, 2025).

Brazilian and international literature documents cases in which health data Data collected for research purposes were subsequently used to develop Commercial products without new consent from participants. The DeepMind-Royal Free case. NHS, discussed in section 7.4 of this guide, is the best-documented example of this pattern. Law No. Law 14.874/2024 establishes that any secondary use of data collected in research requires Specific ethical assessment (Brazil, 2024a). The LGPD determines that the processing of data Personal information for a purpose other than the original one requires a new legal basis or new consent. (Brazil, 2018). Both standards apply to the model as well as to the raw data. The The model is not a neutral product of the research. It is a repository of the data that formed it.

The risk is not limited to the complete model. Knowledge extraction techniques. (*Knowledge distillation*) allows a derived model to learn the patterns of the model. Original without direct access to the raw data. Models published in open repositories. They can be subjected to gradient inversion attacks that reconstruct fragments of Original data, as documented by Geiping *et al.* (2020). The consenting participant There is no control mechanism in place when collecting your data for a specific research project. Regarding none of these derived uses. The protocol does not address the model's destination.

After the research is concluded, it not only omits information, it creates a zone of... deliberate opacity regarding the longer-term trajectory of participant data.

The AI research protocol should cover the entire system lifecycle from the first... Data collected regarding the discontinuation or transfer of the model. Three specific requirements. They structure the CEP's assessment regarding this dimension.

The first requirement is a description of the model's destination after the termination of the process. research. The protocol should explicitly state what will happen to the resulting model: if It will be discontinued and deleted, transferred to a partner institution with precise identification and contractual conditions described, published in a repository with specification of the conditions of access, or incorporated into a product or service with recipient identification. When the If the intended destination is transfer to third parties or publication in an open repository, the The protocol should describe the safeguards that will prevent secondary uses incompatible with The participants' consent. The informed consent form must inform the participants about this destination. in accessible language, before signing. Law No. 14.874/2024 establishes that the Consent must cover the research procedures (Brazil, 2024a). The destination of The model is a procedure with consequences that extend beyond the termination of the... This should be studied and included in the consent form with the same clarity as any other intervention. direct contact with the participant.

The second requirement is the definition of criteria for discontinuing the system during the research. The protocol should establish which conditions trigger the suspension of the use of System before the planned study closure: a performance degradation threshold. Below which the system no longer offers justified clinical benefit, the identification of Performance disparity between subgroups that constitutes differential harm to populations specific, or the occurrence of adverse events linked to the system output that exceed the anticipated risks.

The MRCT Center and WCG (2025) recommend that protocols describe the processes available for researchers to identify and respond to adverse events. related to system performance, including predetermined shutdown criteria. Protocols that do not define these criteria operate without safety checks: the system It remains in use until its scheduled termination, regardless of what happens with... your actual performance.

The third requirement is coverage of the entire life cycle in the informed consent form. Consent for a protocol involving AI should inform participants about six points that Conventional models do not account for: the specific purpose of the model to be trained with your data; the fate of the model after the research ends; the conditions of transfer to third parties and eventual commercial use; the technical mechanism by which the The participant may exercise the right of withdrawal after the model training; the period during which the model will remain active; and the channel through which the participant can obtain information on the use of the model after closure. Fornasier (2022) demonstrated that The absence of this information in informed consent forms for AI research creates an area of legal opacity. which compromises both the rights of the participants and the ethical validity of the research.

The Ethics Committee's opinion when approving AI research approves the complete life cycle of the project. The system that the research will build, not just the research that the protocol describes. The system will continue to exist and operate after the study is completed, after the researchers move on to other projects and then the participants cease to be accompanied. Ethical evaluations that fail to reach this horizon approve only the fraction. visible from a chain of consequences that extends far beyond it.

4. THE BRAZILIAN CONTEXT

This chapter situates this guide within the contemporary Brazilian context and highlights the...
The integration of Artificial Intelligence (AI) into health research. None of the models
The foreign regulatory frameworks analyzed in Chapter 5 were designed for scale and diversity.
Population-wise, and the federative architecture of the SUS (Brazilian Unified Health System). The path to follow starts from this uniqueness.
(4.1) goes through the layered normative architecture that Brazil has built in less than a
Decade (4.2), examines the data infrastructure upon which a sovereign AI can be
developed (4.3), describes the state policy that the agency (4.4), identifies the risks
structural issues that it reproduces when poorly conceived (4.5) and concludes with the normative gaps
that the postal code needs to fill in today, in practice (4.6). Each section functions as a unit
autonomous reading.

4.1 A HEALTH SYSTEM THAT DEFINES ANY IMPORTED MODEL

The Unified Health System (SUS) serves the entire Brazilian population. It functions
in 5,571 municipalities (Brazil, 2026e). It ranges from large university centers to posts
primary care in communities in isolated locations in the Brazilian hinterland. No other
The world's most comprehensive universal healthcare system brings together, on a comparable scale, this combination of
Ethnic diversity, regional inequality, and a constitutional commitment to universality.
Any technology that intends to operate within this system needs to understand it before entering.
in it.

AI follows the same principle. The international models analyzed in this guide were...
built for contexts with more homogeneous technological infrastructure, larger populations
Demographically uniform and health systems organized differently. The model
The Canadian one was designed for government public services. The Australian one, for a
high-income system with reasonably distributed digital coverage. The Indian system, the most
While similar in scale and complexity to the Brazilian Unified Health System (SUS), it still lacks the protective regulatory framework that...
Brazil established the General Data Protection Law (LGPD) with Law No. 13.709, of August 14, 2018.
Personal LGPD (Brazil, 2018). None of them were designed for indigenous and riverside communities.
from the Amazon, nor for the quilombola territories.

This is an affirmation of specificity, not a lament. With rapid progress
Regarding AI in healthcare, Brazil already has an initial foundation and is not starting from scratch. It starts from a
A distinct place, with its own challenges and assets that many countries do not have. The SUS is, in
At the same time, the greatest challenge and the greatest potential for the development of a
Brazilian approach to ethical evaluation of AI in health research.

4.2 The Brazilian Regulatory Framework: A Layered Construction

AI regulation in healthcare in Brazil is built in layers. Each instrument
It addressed a specific social issue: the protection of personal data and ethics in research.
with human beings, the safety of digital medical devices, scientific integrity, the
Protection of children and adolescents in digital environments. These instruments, when combined,
They form one of the most robust protective frameworks among countries that are currently facing the advance of
The accelerated use of AI in healthcare.

The LGPD (Brazilian General Data Protection Law) is the cornerstone of this architecture. The law classifies health data as personal data.
sensitive. Requires a specific legal basis, a stated purpose, and identification of the Data Protection Officer.
Data for your processing. Dourado and Aith (2022) demonstrated that the LGPD introduces into
Brazilian legal system: the right to explanation and review of automated decisions.
(art. 20). This right is the normative expression of the principle of algorithmic transparency, point
central to the regulation of AI systems. The regulation of AI in healthcare in Brazil begins with...
LGPD. Matsushita and Almeida (2025) observe that the articulation between the LGPD and the resolutions
Specific health regulations constitute the most advanced framework for protecting health data.
Latin America.

Law No. 14,874, of May 28, 2024, and Decree No. 12,651, of October 7,
The 2025 law, which regulates it, has elevated ethics in research with human beings to the level of
federal law and established the National System of Ethics in Research with Human Beings (SINEP)
(Brazil, 2024a; 2025a). The text of the law does not use the terms "algorithm" or "intelligence".
"artificial". Its provisions, however, apply directly to the processing of large
databases and the development of computer programs for the purpose of
diagnostic or monitoring.

Resolution CNS No. 738, of February 1, 2024, complements this milestone.
The resolution guided the Ethics Committees on ethical obligations in the creation and management of repositories.

data for scientific research and conditioned the transfer of identifying data to third parties — including AI system developers — to the prediction in the protocol, to justification and approval by the committee, with its execution being the responsibility of the Controller (art. 12 and paragraphs; Brazil, 2024b).

The National Health Surveillance Agency (Anvisa) integrates into this framework with the Resolution of the Collegiate Board No. 657, of March 24, 2022. The rule established the Regulatory framework for Computer Programs as Medical Devices (Software as a Medical Device (SaMD)). An AI system falls into this category when its purpose encompasses diagnosis, prevention, monitoring, treatment, or compensation of diseases (Anvisa, 2022). The system's life cycle now has two mandatory phases and sequential. The Ethics Committee safeguards the rights of participants during the research. Anvisa certifies the derivative product is safe for use on a large scale. One does not replace the other.

The CNPq's Policy on Integrity in Scientific Activity, established in March 2026, mandates that all use of generative AI be declared, specifying the tool and... of the purpose, at any stage of the research (CNPq, 2026).

Law No. 15.211, of September 17, 2025, known as the Digital Statute of the Child and the Child and Adolescent Statute (ECA Digital), adds a specific layer of protection to the framework for research involving children and adolescents in digital environments (Brazil, 2025b). A The law prohibits behavioral profiling of minors for any purpose other than strictly necessary for the operation of the service. Requires privacy settings at the level Highest level of protection since the product's conception. Requires regular review of... AI tools, with the participation of experts and competent bodies, to ensure their safety and suitability for use by this audience. For AI research protocols that If children and adolescents are involved, the protocol must demonstrate compliance with this. The statute, and the CEP verifies it before any other analysis. Full protection is not a principle. Abstract. It is a legal obligation that has been in effect since March 2026.

The Brazilian Artificial Intelligence Plan 2024–2028 projects investments of around of R\$ 23 billion by 2028 and positions health as a priority sector (Brazil, 2025c). A The speed of this progress makes the consolidation of the ethical framework even more urgent. The Project Law No. 2,338 of 2023, authored by Senator Rodrigo Pacheco, proposes the Regulatory Framework. Artificial Intelligence in Brazil and establishes specific obligations for high-risk systems. category that encompasses clinical and diagnostic applications (Brazil, 2023a). The text of the bill also

It provides for regulatory *sandbox* mechanisms for the controlled testing of high-risk systems.

While it is being processed in the National Congress, the current framework already provides a regulatory basis. sufficient for the ethical evaluation of protocols using AI. The Ethics Committee is not waiting for new legislation to act. Act within the framework that Brazilian law has already established.

4.3 The National Health Data Network and the Infrastructure for AI

The National Health Data Network (RNDS) is a national platform for Interoperability created to connect health information systems from different countries. origins, public and private, throughout the Brazilian territory (Brazil, 2020a). Its logic lies in Standardization of the language with which different systems communicate. Information A document produced at a primary care center in Maranhão becomes readable by a system. Hospital in São Paulo. The technical standard adopted is HL7 FHIR (Health Level Seven — Fast Healthcare Interoperability Resources, or Quick Resources for Interoperability in Level Seven Health), version 4.0.1, international reference for interoperability in health. digital.

This standard structures clinical data into discrete, machine-readable features. Each record becomes a computational object with a predictable, queryable structure and Integrable with other systems. A consultation, an examination, a dispensing of medication: They all follow the same organizational logic. For AI, this structure is a prerequisite for entry. An AI system that operates with SUS (Brazilian public healthcare system) data depends on organized, consistent data. and standardized. The RNDS offers this infrastructure on a national scale. Without it, sovereignty Brazilian AI development in healthcare would remain an unfulfilled promise. technique.

The data transmitted over the network includes vaccination records, test results, laboratory tests, dispensing of medications, hospital discharge summaries and records of clinical care (Brazil, 2020a). Access occurs through different channels according to the profile. From the user's perspective. Healthcare professionals access it through the SUS Digital Professional portal. Citizens access it through Meu SUS Digital. Researchers access it by extracting aggregated data and anonymized, with safeguards anchored in the LGPD (Brazilian General Data Protection Law).

Researchers' access to RNDS data, whether through aggregate extraction or by This specific project is governed by CNS Resolution No. 738/2024, which establishes the rules.

in the creation and use of databases for scientific research purposes involving human beings. The Resolution defines the roles of Controller and Operator, and requires Presentation of institutional terms according to the bank's origin — especially the Term Institutional approval is required when the database is established outside the scope of the research—, and Articles 20 and 25 regulate consent for the future use of data and the circumstances under which it may be waived. of the Free and Informed Consent Form (TCLE) (Brazil, 2024b). Chapter 7 develops These figures and the institutional Terms are detailed in Chapter 8; they are integrated as criteria. Operational aspects of Axis 3 of the Risk Assessment Matrix in Artificial Intelligence.

The distinction between clinical use and research use has direct ethical consequences. In healthcare settings, the data circulates under the legal framework of healthcare, for the purpose of... Immediate clinical use and under the supervision of the responsible professional. In research use, the legal basis As things change, consent takes on a central role and the purpose needs to be stated precisely. Anonymization, in this context, is a necessary but not sufficient safeguard. Rocher, Hendrickx and Montjoye (2019) demonstrated that seemingly anonymized databases can allow the re-identification of individuals from the intersection of a few almost identical attributes. identifiers. In a collection the size of the RNDS, this risk multiplies. A protocol Using data from the RNDS, it is necessary to demonstrate not only that the data has been anonymized, but also... Under what conditions does this anonymization withstand cross-referencing with other available databases?

For AI research, this collection has strategic value. It brings together data produced by the Brazilian population as a whole, in clinical, epidemiological and social contexts of enormous diversity. No private hospital consortium, national or foreign, replicates it. This breadth. An AI model trained on RNDS data learns from the real Brazil, not with a cut-out of it. This distinction defines the difference between a useful tool for the SUS (Brazilian Public Health System). and a tool transferred to the SUS (Brazilian Public Health System).

Soares and Chiavegatto Filho (2026) warn that models developed with data Foreigners or private hospitals in the Southeast may be ineffective or harmful to to be applied in populations with different epidemiological profiles.

The problem is structural: the data used to train the AI reflects historical inequalities. When the model learns from data from British biobanks or private hospitals. People from São Paulo also learn about the patterns of exclusion embedded in these collections.

This logic has already produced verifiable consequences in Brazil: prediction algorithms for dengue fever, COVID-19 surveillance, and analysis of death certificates through processing of

Natural language processing functions with acceptable performance when validated across populations. that generated their training data, and the validity of the models collapses when they are applied to distinct epidemiological and social contexts, without local revalidation (Soares; Chiavegatto Son, 2026).

RNDS responds to this risk with the only structurally viable countermeasure: data. produced by the real Brazil, on a national scale, about the totality of people served by the SUS (Brazilian Public Health System). It meets the requirements. A model trained on this collection has a chance to learn what no other data can. Imported would teach. The dispute over data sovereignty is, therefore, also a dispute. for the usefulness and fairness of AI in Brazilian public health.

4.4 SEIDIGI AND THE CONSTRUCTION OF BRAZILIAN DIGITAL HEALTH

The Secretariat of Information and Digital Health, known by the acronym SEIDIGI, was created by Decree No. 11,358, of January 1, 2023, as the implementing arm of the digital health policy. of the Ministry of Health. The institutional mission of the Secretariat encompasses planning, incorporation and supervision of information technology products and services in the SUS (Brazilian Public Health System), including the management of the RNDS, the development of interoperability systems and the Health data protection (Brazil, 2023b). For the field of AI, SEIDIGI occupies the position of a key player in defining the technical standards that any AI system will need to adhere to. to operate with SUS data. Decree No. 11,358/2023 was revoked by Decree No. 11.798, of November 28, 2023, which approved the current Regulatory Structure of Ministry of Health, in which SEIDIGI was maintained (Brazil, 2023c).

Haddad *et al.* (2025) documented the trajectory that led to the creation of the secretariat. Brazil has accumulated pioneering experience in telehealth over nearly two decades. public. The Brazilian National Telehealth Program, initiated in 2007, preceded most of the middle-income countries in the organization of teleconsultation and telediagnosis networks in public health system. The COVID-19 pandemic accelerated the regulatory framework, with the Law Law No. 14,510, of December 27, 2022 (Brazil, 2022), which authorized and regulated telehealth. throughout the entire national territory. This trajectory was not linear. It faced resistance. corporate challenges, funding gaps, and systemic fragmentation. SEIDIGI was created precisely because of these challenges. to overcome this fragmentation.

What distinguishes SEIDIGI from previous secretariats with similar functions is its scale and... The integration of responsibilities. SEIDIGI incorporates DATASUS, historically responsible through the SUS information systems, and coordinates actions through the Health Management Committee. Digital, which brings together the National Health Surveillance Agency and the National Health Agency. In addition, the National Council of Health Secretaries and the National Council of Municipal Health Departments (Haddad *et al.*, 2025). This governance architecture It transforms digital health from an isolated initiative into a state policy with formal bodies. of federative coordination.

The SUS Digital Program, established by Ordinance GM/MS No. 3,232, of March 1, 2024 is the operational expression of the SEIDIGI agenda. The program organizes digital health. in strategic actions that include telehealth, interoperability, electronic health records, Digital identity of citizens and artificial intelligence applied to management and assistance (Brazil, 2024c). Each action has defined goals, responsible parties, and evaluation criteria. Digital health It ceases to be a project and becomes a structured program, with a life cycle logic and accountability.

Public Call for Proposals No. 01/2026 of Inova SUS (Brazil, 2026a), coordinated by SEIDIGI is moving in this direction with a methodological innovation. The Brazilian State is ceasing to be It is merely a regulator of technologies developed by the private sector and begins contracting them with Ethical and technical criteria are defined beforehand. The calls for proposals specify the requirements for... algorithmic transparency, representativeness of training data, and compatibility. with the RNDs. A system that does not respect these criteria does not pass the selection process. Ethics leaves It goes from being a requirement that comes after development and becomes a condition of entry.

According to Santos *et al.* (2025), initiatives such as the federal program "Agora Tem Experts", the gradual federalization of the RNDs and Inova SUS constitute an agenda called digital health sovereignty. This expression refers to the capability of a country in the Global South. Global to create and control health technologies based on the populations themselves, under Public management with a focus on equity. This direction has a concrete technical translation. An AI. developed specifically for the reality of the SUS (Brazilian Public Health System), taking into account the conditions Brazilian epidemiological and regional inequalities differ technically and ethically from an AI. Imported and adapted.

This movement has a direct consequence for AI research. Systems developed in a research context and intended to operate within the SUS (Brazilian Public Health System) will need, at some point...

Currently, it needs to pass through the SEIDIGI review process. The research protocol submitted to the Ethics Committee is not the... final product. This is the stage that precedes regulatory and technological evaluation. The more rigorous the ethical evaluation of the protocol, the greater the likelihood that the resulting system will be successful. be approved in subsequent stages. CEP and SEIDIGI do not compete. They are part of the stages. distinct from the same chain of responsibility.

4.5 Algorithmic Bias, Regional Inequality, and the Risk of the Imported Model

Algorithmic bias is inherited, not a programming flaw. The data that trains... AI systems reflect the patterns of those who produced them, those who use the healthcare system, and of that which was deemed sufficiently qualified to be registered. When these Patterns reflect historical exclusion; the algorithm learns from exclusion. Soares and Chiavegatto Filho (2026) documents that even state-of-the-art models reproduce biases from their training bases, including the so-called healthy volunteer *bias* and systematic patterns of missing data associated with data collection processes with scope limited.

Brazil has conditions that make this risk more acute than in countries with Homogeneous sanitation infrastructure. Regional inequality is not a demographic variation. Distinct epidemiological profiles coexist in the same country: the burden of arboviruses in the North and Northeast Brazil, maternal mortality among Black women, tuberculosis in populations in prisons and those experiencing homelessness, neglected diseases that don't appear in large populations Private hospital consortia in the Southeast region of Brazil. A model trained on hospital data. University students in São Paulo learn about a Brazil that doesn't exist for most people. attended by SUS.

Algorithmic racism is the precise name for what happens when this process... It focuses on the Afro-Brazilian population. Poorly collected data, systemic absence of specific populations in training sets and technical adjustments blind to race/color. converge to produce systems that, under the guise of mathematical neutrality, They reproduce patterns of exclusion built up over centuries. This is not a failure. Accidental. It is the predictable result of teaching an algorithm about records that have never been recorded. designed to capture the experience of excluded populations.

International literature confirms this asymmetry with increasing precision. Rajkomar *et al.* (2018) demonstrated that predictive models with acceptable performance in

Majority populations produce substantially worse results in groups.

historically underrepresented in training data. The problem won't be solved with more

The problem is solved with the right data, from the right populations, collected using the right protocols.

That's right. That's why the representativeness of training data is no longer an issue.

methodological and became an ethical issue with normative support expressed in the main international guidelines analyzed in this guide.

The racial issue adds a layer of complexity that no technical adjustment can overcome.

It solves it on its own. Race and ethnicity data are essential to reveal health inequities.

Maternal mortality among Black women, underreporting of sickle cell disease in regions with low laboratory coverage and differential access to diagnostic tests

These patterns only appear when the racial variable is present and well-collected.

Fiske *et al.* (2025) demonstrated that the use of these variables without adequate contextualization

It can reproduce stigmas and reinforce structural discrimination within the algorithm itself.

They should correct them. The misuse of racial data transforms the equity tool into...

instrument of exclusion.

Explainability is the other side of this problem. Models with low transparency, often called black boxes , they prevent professionals and

Users need to understand how a decision was made. This opacity weakens control.

Socially, it hinders institutional accountability in case of error and makes it impossible to identify

At what stage of the process was the bias introduced (Soares; Chiavegatto Filho, 2026)?

In the context of the Brazilian Unified Health System (SUS), where algorithmic decision-making can affect access to a procedure, the

Whether it's regulating a queue or triaging a patient, opacity has consequences that go beyond simply managing a queue or triaging a patient.

Beyond the technical aspects, they also affect rights.

The inequality in algorithmic literacy among healthcare professionals themselves.

amplifies this risk. Soares and Chiavegatto Filho (2026) identified that most of the

Professional categories still lack the training to interpret predictive algorithms.

to understand the probability of outcomes and to recognize when a system is operating outside of them.

parameters for which it was validated. A professional without algorithmic literacy does not

It questions the model. It accepts the result as given. The automation of error, in this scenario, is...

faster and more comprehensive than any individual human error could be.

Assessing the representativeness of training data is an ethical obligation of research.

There is no methodological detail to be verified only by the reviewers of a scientific journal.

The protocol submitted to the committee needs to answer specific questions: for which population?
Was the model trained? Does this population share the epidemiological, social, and... profile?
What is the demographic profile of the people who will be affected by the research? The training data includes
Are there groups that have been historically underrepresented in Brazilian health information systems?
The MRCT Center and the WCG (2025) position these questions as entry criteria in
Ethical evaluation of any protocol involving AI. The guide that the reader has in their hands adopts the
same principle.

The tool developed by the Pan American Health Organization (PAHO) for
The assessment of countries' institutional readiness for the adoption of AI in public health incorporates
technical, legal, ethical and organizational dimensions (PAHO/WHO, 2024). The ethical dimension
It explicitly includes the evaluation of algorithmic equity and justice as a condition for the
Responsible adoption of any system. It's not necessary to master the fundamentals.
mathematicians of a machine learning model. The essential thing is to ask the questions.
certain. The Risk Assessment Matrix in Artificial Intelligence presented in Chapter 8
This guide organizes these questions into five axes, of which Axis 4 is dedicated to
The representativeness of the data and the transferability of the model directly address the risk.
of inherited bias and algorithmic racism.

Algorithmic fairness is not a future aspiration. It is a present-day regulatory requirement.
Rajkomar *et al.* (2018) demonstrated that ensuring equity in predictive models in health
This is a condition for advancing health equity as a whole. In Brazil, where the SUS (Unified Health System) carries...
Given the constitutional commitment to universality and equity, this statement carries weight.
normative. An AI system that reproduces inequalities in access to diagnosis or in
The outcome of the treatment contradicts the fundamentals of the public health system. Evaluate this.
The risk before the system is put into use is the most concrete contribution that CEP can make to...
The health project that Brazil chose to build.

The corresponding operational material is in Parts V and VI of MARIAH.

5. INTERNATIONAL OVERVIEW: EXPERIENCES IN ETHICAL EVALUATION AND AI REGULATION IN HEALTH RESEARCH

5.1 INTRODUCTION: WHY LOOK OUTSIDE

International experience in regulating artificial intelligence in healthcare has advanced with rapid growth in recent years. The WHO, along with European countries, Canada, Australia, The United States and East Asian nations have established regulatory frameworks. ethical review instruments and guidelines for research committees that provide Concrete and verifiable contributions. Brazil learns from each of these experiences. It doesn't matter what they are, but learn from them to build your own.

The path that led to the construction of the Intelligence Risk Assessment Matrix. Artificial Intelligence in Research with Human Beings (MARIAH) of the National Ethics System in The research (SINEP), described in Chapter 8, went through the WHO's ethical framework and nine models. Regulatory frameworks across three continents. None of them were designed for this context. Brazilian epidemiological and social context, for the ethnic and regional diversity of the population served. by the SUS (Brazilian Public Health System) or for the protective framework of the LGPD (Brazilian General Data Protection Law). These specificities are not gaps. It's regrettable. These are the areas where the Brazilian model differentiates itself and justifies its position.

The following overview starts from the WHO's multilateral ethical framework and examines nine National regulatory experiences: Canada, Australia, United States, Europe, China, Japan, South Korea, India, and Russia. For each experience, the text identifies the regulatory framework. Relevant to research ethics committees, the role assigned to the ethical oversight of systems algorithmic and the specific contribution that each model offered to MARIAH. Section 5.9 It synthesizes the lessons and explains what the Brazilian context adds to what none of these [lessons] offer. He predicted experiences.

5.2 The Multilateral Ethical Framework: The World Health Organization as FOUNDATIONAL REFERENCE

The first globally applicable regulatory instrument on the subject is the Recommendation. on the Ethics of Artificial Intelligence, adopted by UNESCO member states on 23rd November 2021, which establishes respect for human rights and human dignity. based on principles such as transparency, fairness, sustainability, and oversight.

Human performance in AI systems (UNESCO, 2022). The ethical governance of AI in healthcare is addressed by the WHO. its most established multilateral reference. In 2021, the WHO published the Ethics document and Governance of Artificial Intelligence for Health, resulting from a consultation with twenty international experts. The document established six consensus-based ethical principles for to guide governments, developers, researchers, and healthcare providers in development and adoption of AI systems (WHO, 2021).

The six principles have guided international ethical debate ever since. 1) The first is... Protection of autonomy: human beings must remain in control of their decisions. clinics, with access to the information needed to use AI systems safely and 1) Effectiveness. 2) The second is the promotion of human well-being, safety, and interest. public: the systems must meet regulatory requirements for safety and effectiveness for intended uses. well defined. 3) The third is the guarantee of transparency, explainability and intelligibility: the AI systems must be understandable to developers, healthcare professionals, Patients and regulators. 4) The fourth is the promotion of responsibility and accountability: The use of AI should occur under appropriate conditions, with points of human oversight. established both before and after the algorithm. 5) The fifth is the inclusion guarantee and Equity: Systems should be designed for broad and equitable use, without encoding biases. to the detriment of identifiable groups. 6) The sixth is the promotion of responsive AI and Sustainable: the systems must be consistent with the sustainability of the systems. health, environment and workplaces (WHO, 2021).

These six principles did not arise from speculation. The WHO formulated them based on... concrete evidence regarding the risks of diagnostic support systems, of the models predictive and screening systems that were already operating in real-world clinical settings in 2021. Low- and middle-income countries face specific challenges in adopting AI in healthcare, such as... the reliance on models developed in contexts with superior infrastructure and the lack representative data of their populations. This dimension of global equity resonates directly within the context of the SUS (Brazilian Public Health System) and grounds the local validation requirements that this guide establishes.

In 2024, the WHO published a specific update for large models. multimodais (Large Multi-Modal Models — LMMs): Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. The document acknowledges that The principles of 2021 remain valid, but the emergence of LMMs has introduced...

New challenges that require specific attention (WHO, 2024). Generative models were adopted faster than any other consumer application in history, without that individuals, healthcare professionals, or governments were prepared for the risks involved.

The 2024 update identifies three dimensions of risk that were absent from the 2021 discussion. The first is content creation: LMMs produce responses that appear authoritative, but may be factually incorrect, including scientific citations. non-existent and inadequate clinical recommendations. The second is opacity of the bases of Training: the data used to train the main available LMMs were not disclosed, making it impossible to verify whether they are biased, whether they were obtained legally, and whether The model's performance reflects genuine capability or memorization of the data. training. The third is exclusive control by large corporations: the concentration of The development of LMMs in a few private companies creates structural dependence and an asymmetry of power that public health systems need to recognize and address.

The WHO documents constitute, in this guide, the minimum ethical standard for... approval of protocols with AI, regardless of technical sophistication or prestige of institution. The six principles of 2021 and the specific recommendations for LMMs of 2024 are operational parameters. This guide operationalizes them within the context of SINEP through... MARIAH and the guidelines developed in the following chapters.

5.3 CANADA

The Canadian model is the most directly influential in the construction of MARIAH. The Secretariat of the Treasury Board of Canada developed and maintains the Assessment of Algorithmic Impact Assessment (AIA), a mandatory tool. for AI systems deployed by Canadian federal government agencies (Canada, 2024). The AIA operates with binary-answer questions distributed across four risk dimensions: Impact, project context, data, and technology. It classifies systems into four levels, each One associated with mandatory minimum requirements. The rule that mitigations do not alter the level. Regarding risk, only the required criteria originate directly from this model.

Two elements of the AIA were incorporated into MARIAH with adaptations. The first is the The weighted scoring logic of Version B: thematic blocks with differentiated weights, whose The sum determines the risk level on a numerical scale. The second is the separation between

risk classification and required criteria, which preserves the integrity of the instrument and It prevents statements of intent from replacing evidence of effective risk control. The AIA It was designed for government public services, not for protocols submitted to ethics committees. This transition required a reformulation of the axes, the questions, and the basis. normative framework of the instrument.

Canada is also a benchmark for the Three-Agency Policy Statement: Ethics of Research with Human Subjects (TCPS 2), a document that guides ethics committees in Canadian research. In its most recent edition, TCPS 2 addresses the use of AI in research and It establishes that ethical oversight must accompany the entire research lifecycle, from From conception to dissemination of results (Canada, 2022). This principle guides the structure Mariah's steps.

5.4 Australia

The Australian model is the main reference for the qualitative logic of Version A. (qualitative) of MARIAH. The National Council for Health and Medical Research of Australia The National Health and Medical Research Council (NHMRC) published the Guide in 2023. Evaluation of Research Involving AI (*Guide for Assessing Research Involving AI*), document specifically designed to guide research ethics committees in evaluating protocols with algorithmic systems (Australia, 2023). The Australian guide starts from a The principle that guides all of Chapter 8 of this guide: the ethics committee does not need to understand The algorithm's code. You need to understand the impact it can have on the person. participant.

The NHMRC organizes the evaluation around the AI system lifecycle and distinguishes three stages with distinct risk profiles: exploratory discovery (*Discovery*), translation Clinical (*Translation*) and Execution (*Deployment*). A system in the exploratory discovery phase. It operates with retrospective data and open hypotheses. A system in the implementation phase. It influences clinical decisions in real time. The ethical risks of each stage are... fundamentally different and require different questions from the committee. This distinction was incorporated into Block 1 of Version B of MARIAH, which requires the protocol to state in which The system is at the stage it is in, and at Step 0 of both versions, which guides the filtering of Entry based on the system's actual role in the protocol.

The most structuring contribution of the Australian model to Version A is the logic of Sequential flowchart. The NHMRC demonstrated that ethical assessment instruments are more effective when they go through dimensions in progressive order, from the simplest to the most complex. They are complex and guide the evaluator's judgment without replacing it.

This logic was fully adopted in Version A: the five axes are traversed in The sequence continues, and the risk level only increases, never decreases. The Australian guide is also the source. The main concept of continuous human supervision, which underlies Axis 5 of both... the versions. The presence of a qualified researcher with the real ability to review and Challenging the system's results is not only advisable, it is a prerequisite for the The protocol does not automatically reach the highest risk levels.

The Australian model also contributed to Axis 4, dedicated to representativeness and Transferability of the model. The NHMRC reported that systems developed for certain Populations tend to underperform when grouped together with different characteristics. ethnic, socioeconomic, or geographic. The requirement that the protocol demonstrate validation. The model, when applied in a context comparable to that of the research, is directly grounded in this principle.

A significant limitation of the Australian model for Brazilian adaptation is... absence of specific criteria for vulnerable populations and for high-risk contexts inequality. The guide was developed for a high-income healthcare system, with relatively homogeneous technological infrastructure. MARIAH incorporates questions specifics regarding population diversity and validation in a context comparable to the SUS for to fill this gap.

5.5 UNITED STATES

The main American contribution to MARIAH comes from the directive of United States Food and Drug Administration Administration (FDA), published in January 2025. The document addresses the use of AI in life cycle of pharmaceutical and biological products and introduces two concepts that informed The structure of MARIAH: the context of use and the consequence of the decision (FDA, 2025).

The main contribution from the US was in the context of use, which defines the specific role. of the system and the scope of the question it was developed to answer. The consequence The decision assesses the severity of the potential adverse event when the model operates with high...

influence and low human supervision. These two concepts underpin, respectively, Step 1 and Block 4 of Version B.

The FDA also contributed to the concept of lifecycle maintenance: the requirement that adaptive systems be continuously monitored after they begin to be used, with Defined protocols for data drift *and* model retraining. This The concept is incorporated into Axis 4 of Version A and into the risk reduction measures of Block A. 7 of Version B. The FDA guideline demonstrated that the performance evaluation of a system AI used in contexts with human supervision should consider the team's outcome. Human-AI interaction is not just the isolated performance of the model. This principle underlies the... Questions regarding effective oversight mechanisms in both versions of MARIAH.

The US regulatory context also includes the benchmark of the Center for Multiregional Clinical Trials from Brigham and Women's Hospital and the University of Harvard MRCT Center e WCG (Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard) which published a set of tools for use in 2025. Responsible for AI in clinical research (MRCT Center; WCG, 2025). This reference was the source. the distinction between routine and non-routine use that underlies the input filter of both versions of MARIAH, and the concern about the capabilities of large models of A language for inferring identities and sensitive attributes from seemingly given data. anonymized, incorporated into Axis 3 and Block 6.

5.6 EUROPE

The European contribution to MARIAH comes from two complementary sources. The first is the General Data Protection Regulation (GDPR). — GDPR), in force since 2018 (European Union, 2016), which establishes the most stringent standard. of the personal data protection law in force worldwide and served as a reference for the construction Brazilian General Data Protection Law (LGPD). The GDPR classifies health data. as a special category of personal data, with reinforced requirements regarding legal basis and purpose. declared and minimization of treatment. This normative architecture underpins Axis 3 of Version A and Block 6 of Version B, which have the greatest regulatory weight in both instruments.

The second source is the European Artificial Intelligence Regulation (EU AI Act). Approved in 2024 and in the process of being progressively implemented. The regulation classifies systems.

AI risk is categorized into four risk categories: unacceptable, high, limited, and minimal. AI systems used in healthcare and clinical research contexts, they fall into the high category. risk, with mandatory requirements for transparency, human oversight, technical robustness and Data management (European Union, 2024).

This logic of proportional classification influenced the four-level structure of MARIAH, although the Brazilian adaptation replaced the European criteria with its own criteria. oriented towards the context of research with human subjects and the national regulatory framework.

The WHO's contribution to MARIAH, including the six principles of 2021 and the Specific recommendations for large multimodal models by 2024 are described in Section 5.2. The principle of fairness is especially relevant to the European context and Brazilian: the WHO documented that models trained predominantly with data from High-income populations systematically perform worse in other populations. with different ethnic and socioeconomic characteristics (WHO, 2024). This finding This underpins Axis 4 of both versions of MARIAH and the emphasis on questions about representativeness and validation in a context comparable to the Brazilian Unified Health System (SUS).

5.7 Asia

The Asian experience in regulating AI in healthcare is marked by distinct approaches. among themselves, but converging on a central point: in all three countries examined, the The research ethics committee is recognized as the first line of technical defense against Algorithmic biases, system opacity, and data protection violations. This Convergence legitimizes and reinforces the purpose of this guide.

5.7.1 China

China has built Asia's most stringent regulatory framework for AI systems in health. The National Medical Products Administration (NMPA) The National Medical Research Institute (NMRI) adopts a three-class hazard system for medical devices. with AI. Autonomous diagnostic systems and clinical decision support in critical conditions. They are classified as Class III, with mandatory clinical validation in Chinese territory and double-level ethical scrutiny (China, 2021). This principle of maximum precaution influenced

The logic of MARIAH: protocols involving autonomous systems without human supervision. Effective performance automatically reaches Level IV, regardless of performance in other axes.

The most relevant regulatory instrument for ethics committees is the set of Measures for the Ethical Review of Life Sciences and Medical Research Involving Human Beings Humans, in effect since February 2023. This document formally expanded the The scope of ethical oversight should encompass algorithms, the use of large databases, and... brain-computer interfaces (China, 2023).

He also introduced the mechanism of double ethical review for protocols with Direct algorithmic interventions: submission to an institutional committee and, frequently, to higher instance. Article 32 of these measures establishes exemption from further ethical review. for retrospective studies with previously anonymized data and minimal risk. This The solution engages with the Brazilian debate on the secondary use of data in research and with MARIAH's Axis 3 questions on consent and anonymization.

The main shortcoming of the Chinese model, for comparison with Brazil, is the absence... national program for training members of ethics committees for evaluation of algorithmic systems. Supervision falls on methodologies developed autonomously. by large hospitals, without federal standardization. The Data Protection Law China's Personal Information Protection Law (PIPL), in effect since 2021, requires that data be kept locally and that explicit consent be given. main legal basis. This creates a system of data sovereignty that, in practice, prevents the Conducting international multicenter studies with the sharing of raw data. This restriction is the most direct contrast to Brazil's LGPD, which allows data transfer. International data transfer based on contractual guarantees and specific consent.

5.7.2 Japan

Japan has adopted one of the most pragmatic regulatory approaches in Asia for Adaptive AI systems. The Pharmaceuticals and Medical Devices Agency. The Pharmaceuticals and Medical Devices Agency (PMDA) developed the Assessment system. Timely Innovation (IDATEN, from the English *Improvement Design within Approval for Timely Innovation*) *Evaluation Notice*), a mechanism that allows iterative updates to existing AI systems.

approved without requiring a complete new regulatory process, provided that the modifications occur within previously agreed parameters (PMDA, 2023). This concept was incorporated into Axis 5 of Version A and Block 7 of Version B of MARIAH, in the questions about Model retraining protocol and data drift monitoring.

Ethical Guidelines for Medical and Biological Research Involving Human Subjects Published in 2021 by the Japanese Ministries of Health and Education, they brought three Relevant innovations for ethics committees: centralized protocol review. multicenter evaluations by a single committee; digital validation of consent; and *Opt-out* mechanism for secondary use of data in artificial intelligence research. artificial (MHLW; MEXT, 2021).

The exclusion mechanism allows researchers to utilize vast databases. clinicians without obtaining explicit consent, provided they publish the study details and ensure that patients have an accessible way to retrieve their data. This principle has been incorporated. to MARIAH in a more protective way: question 2.8 of Axis 2 requires that the protocol provide for mechanism for the participant to withdraw consent at any time, with exclusion from your data from the study databases and a guarantee that it will not be used in training or future retrainings.

5.7.3 South Korea

South Korea offers the two most innovative contributions from the Asian experience. for MARIAH. The first one comes from the Ministry of Food and Drug Security. (Ministry of Food and Drug Safety — MFDS), which developed two criteria for classification. Risk factors of AI systems in healthcare: the potential for failure (*Malfunction Potential*) and the *Opportunity to Override* (MFDS, 2022).

The failure potential assesses the likelihood that algorithmic errors will cause harm. clinical care for the patient. The opportunity for rebuttal assesses whether the researcher is qualified. There are real reasons to question the system's outcome before its execution. These two criteria were incorporated into Axis 1 of both versions of MARIAH and Block 2 of Version B, in Questions regarding system autonomy and effective human oversight capacity.

The second contribution comes from the Framework Law for the Development of Intelligence. Artificial Intelligence and *the Establishment of Trust (Framework Act on the Development of Artificial Intelligence)*

Intelligence and Establishment of Trust), approved in January 2025 and in force since January 2026 (South Korea, 2025).

The law imposes ethical requirements for high-impact sectoral systems, including health, and This makes South Korea the second jurisdiction in the world, after the European Union, to enact such law. Comprehensive AI. The MFDS guidelines for clinical trials with AI, published in 2022, They instruct ethics committees to audit three phenomena absent from the evaluation of Conventional drugs: data leakage between training and test sets. hidden population biases and performance variation among hospital units with different equipment (MFDS, 2022). These three dimensions are incorporated into the 3 Axes and 4 from both versions of MARIAH.

5.7.4 India

India is the country that most closely aligns with the purpose of this guide. The Indian Council of Medical Research (Indian Council of Medical Research — ICMR) published the Guidelines in 2023. Ethics for the Application of Artificial Intelligence in Biomedical and Health Research (Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare), o The only document in the world specifically designed to empower ethics committees in evaluation of protocols with algorithmic systems (ICMR, 2023).

The guidelines structure the evaluation with ten bioethical principles and highlight that AI It is intended for diagnostic enhancement. It does not replace clinical judgment. The professional The person in charge holds irrevocable sovereignty over the system's outcomes. This principle is... The central foundation of Axis 5 in both versions of MARIAH.

The most innovative Indian contribution to MARIAH is the exclusion mechanism. Late withdrawal: the research participant has the right to request the deletion of their data from the database. repository used by AI at any time, even after the study has begun. None of the Western models analyzed foresee this right. MARIAH incorporates it as an ethical requirement. in question 2.8 of Axis 2. The Central Organization for Drug Standards Control (CDSCO, from the English Central Drugs Standard Control Organization) contributed to the Protocol Algorithm Change Protocol, which requires manufacturers to submit In advance, a plan describing how the model will be retrained within

technical parameters previously agreed upon (CDSCO, 2025). This principle underlies the Questions about model retraining in Axis 5 and Block 7.

The main drawback of the Indian model is legislative. The Data Protection Act. Digital Personal Data Protection Act (DPDP Act), approved in 2023 (India, (2023), completely removed the sensitive data category from its text and equated medical records, genetic records, and psychiatric data, as well as any other data. personal (Burman, 2023). This choice transfers responsibility to the Ethics Committees. to protect the participant that the law has left unprotected. The Brazilian LGPD makes a choice. Conversely, health data constitute a sensitive category with enhanced protection. This distinction It is one of the pillars of Block 6 of Version B and Axis 3 of Version A.

5.8 Russia

The Russian experience offers the guide two contributions of distinct natures: one a regulatory model of maximum precaution and a cautionary case regarding the risks of Subordinating ethical oversight to the logic of accelerated innovation.

The Federal Agency for Health Surveillance (Roszdravnadzor) adopts a classification automated: any healthcare system based on artificial intelligence falls under the category of Class 3, the highest risk level, regardless of its application (Russia, 2020). This The decision forces all AI-based protocols to go through the most rigorous validation process. Clinical and ethical approval. It consists of a simple and comprehensive administrative solution. A MARIAH grades the requirements in proportion to the risk, but acknowledges the need for maximum Caution is advised when autonomous decisions could have an irreversible impact on the participant.

Federal Law No. 152-FZ on Personal Data of Russia (Russia, 2006), as amended Recent studies classify health data as a special category and require explicit consent. for its treatment. It also imposes mandatory data localization on Russian servers, which, in practice, makes international multicenter studies with data sharing unfeasible. raw data.

5.9 Comparative Synthesis and Lessons for Brazil

The WHO ethical framework and nine national regulatory experiences show that evaluating Ethically sound protocols with artificial intelligence go beyond simple verification of Documentary compliance. The assessment requires an instrument that reflects complexity. A technique that impacts the participant. That is exactly what MARIAH aims to achieve. to do.

Brazil inherits a specific contribution from each model. From the WHO, the six principles. Foundational ethics and warnings about the specific risks of large multimodal models. From Canada, the logic of weighted scoring, and the critical rule that mitigations do not alter the risk level. From Australia, the sequential flowchart and the distinction between stages of the cycle of The life of the system. From the United States, the concept of context of use as a preliminary step. mandatory and the principle of maintaining the life cycle with drift monitoring data. From Europe, the proportional risk classification and the enhanced protection standard for Sensitive data. From South Korea, the criteria for potential failure and opportunity of contestation. From Japan, the model retraining protocol within parameters previously agreed upon. From India, the participant's right to request the exclusion of their Training data can be accessed at any time, and the principle of irrevocable medical autonomy is upheld. Russia, the principle that maximum precaution is justified when autonomous systems They can cause irreversible damage to the participant.

What none of these models predicted is the Brazilian context. None of them consider the The ethnic and regional diversity of the population served by the SUS (Brazilian Public Health System). None of them are anchored in... LGPD, which classifies health data as a sensitive category with enhanced protection and It offers research participants a set of rights that surpasses those of all others. Asian legislation analyzed. None was designed for a universal healthcare system. which serves the entire Brazilian population under radically unequal conditions of access. These Specificities are not adaptations on the margins of the instrument; they are the core of what makes it... MARIAH is necessary and distinct.

An additional distinguishing feature of the Brazilian framework escapes any of the models. international comparisons. CNS Resolution No. 738/2024 operationalizes the LGPD in this context. specific to research with databases involving human subjects, with detailed information. which none of the nine national experiences examined presents: definition of the figures Legal distinctions between the Controller and the Operator, operational distinction between banks incorporated in within the scope of the research and databases already established outside of that scope, the institution of the four Terms.

Institutional agreements, consent, commitments, and transfers that structure the chain.

data custody, and an exhaustive enumeration of the circumstances under which informed consent is not required for future use, different depending on the origin of the bank. This level of operationalization of data protection.

Research involving human subjects is a unique feature of the Brazilian system and forms the basis. which MARIAH bases her argument on in Chapters 7 and 8.

6. Scientific Integrity and Authorship in AI Research

Artificial intelligence has not entered scientific research merely as a tool for analysis. Entered as an agent capable of drafting protocols, generating bibliographic references, To produce synthetic data and formulate hypotheses fluently in a way that simulates technical authority. This entry creates a new set of responsibilities within the Brazilian legal system. He began to treat it with precision.

Five instruments currently structure scientific integrity in AI research in Brazil. CNPq Ordinance No. 2,664/2026, which establishes the Integrity Policy in the Activity. Scientifically, it requires mandatory declaration of the use of generative AI. Law No. 14.874/2024 and the Decree No. 12,651/2025 establishes the researcher's full responsibility for all aspects of the project. protocol. CNS Resolution No. 738/2024 regulates the establishment and management of banks of data for research, and consolidates integrity and governance duties in the supply chain. treatment. CFM Resolution No. 2,454/2026 preserves medical responsibility regarding AI-supported decisions. The LGPD (Brazilian General Data Protection Law) establishes responsibility for the data that... The system processes.

This chapter translates this regulatory framework into the requirements that a protocol must meet. satisfy requirements before merit evaluation, when using AI at any stage of research.

6.1 What changes when AI enters research?

AI enters research in radically different ways, and this distinction determines the a set of ethical responsibilities that applies to each case. There are two main ways of The use of AI in the context of research.

In its primary use, AI operates as a processing tool: it analyzes images. medical, classifies laboratory tests, identifies patterns in databases epidemiological, processes time series of physiological signals. The researcher defines the The question is asked, the data is selected, the results are interpreted, and all those choices are taken into account. AI is a tool, just like the microscope and statistical software were.

In the second use, AI operates as a content-generating agent: it writes sections of the The protocol is developed, hypotheses are formulated, literature reviews are produced, and the Informed Consent Form is drafted. Free and Informed Consent (TCLE) generates synthetic data and suggests analyses. In this use, the boundary between

The central ethical question becomes what the researcher created and what the machine produced. This chapter addresses the second use, acknowledging that both can occur in the same context. protocol, which is currently common.

The Brazilian legal system responded to this challenge with normative instruments. accurate, although gaps remain in the extended algorithmic chain that are addressed. in 6.4 and in Chapter 7. CNPq Ordinance No. 2,664, of March 6, 2026, established the obligation to declare the use of any generative artificial intelligence tool in Promotion and research activities, with a description of the tool used and the function performed. and the method of human verification of the results (CNPq, 2026).

Decree No. 12,651/2025, which regulates Law No. 14,874/2024, consolidated the The principle that no stage of research involving human beings escapes responsibility. identified and verifiable human source (Brazil, 2025a). Responsibility for inaccuracies, plagiarism or data generated incorrectly by AI systems falls entirely on the human researchers. This principle does not admit exceptions due to technical complexity, for limitations of the system or due to the researcher's lack of knowledge about how it works. The algorithm it used.

Scientific integrity in AI research is not an academic issue about authorship and Intellectual property credit. It has a direct and verifiable consequence on the people who participate in it. Research. A protocol that uses a large-scale language system to draft sections. The informed consent form can produce fluent but legally imprecise language, with omissions regarding... specific risks that the system was unable to identify or that it misinterpreted.

A protocol that uses generative AI to formulate hypotheses without human verification. rigorous measures can expose participants to poorly substantiated interventions that appear to be... Scientific evidence. A protocol that presents synthetic data as real data. This undermines the scientific validity that justifies the risk imposed on participants. The protocol A document that does not explicitly state the use of generative AI at each stage is incomplete. The evaluation focuses on what has been described and implicitly accepts what has been omitted.

Seven regulatory instruments structure the researcher's obligations and mandate. The role of the Ethics Committee in evaluating protocols with AI. The triangulation between Law No. 14,874/2024, the Decree No. 12,651/2025 and CNPq Ordinance No. 2,664/2026 define the core of The researcher is responsible for the integrity of each stage of the research.

Resolution CNS No. 738/2024 complements this core by offering, in its article 3, XII, The Brazilian normative definition of research integrity is "commitment to values." ethical principles and good practices in conducting research," and by imposing, in Article 5, VII, the It is the duty of researchers, sponsors, and institutions to respect this principle. It is the first since the Brazilian legal system establishes this definition in a rule of national scope.

CFM Resolution No. 2,454/2026 adds the dimension of medical responsibility. Regarding AI-supported decisions. The LGPD (Brazilian General Data Protection Law) establishes responsibility for data. Personal data produced and processed throughout the system's lifecycle. CNS Resolutions Decrees No. 466/2012 (Brazil, 2013) and No. 510/2016 (Brazil, 2016) preserve the bioethical principles that They permeate the entire SINEP system, and the arrival of AI makes them more demanding, not less.

6.2 HUMAN AUTHORSHIP: PRINCIPLE, FOUNDATION AND LIMIT

The discussion about scientific authorship almost always revolves around credit: who Who signs the article, who appears in the acknowledgments, who receives recognition for work. In AI research, the central question is who is responsible for the damage.

AI systems do not appear before CEPs, they do not respond to legal proceedings. In the context of investigations into scientific misconduct, they cannot be held liable for damages caused to research participants do not have professional registration, do not have institutional affiliation, and do not They have assets subject to civil redress. Human authorship is the mechanism that ensures that Someone responds. Removing human authorship from any stage of research is not a matter of Intellectual property credit. Removes the protection that guarantees accountability to the participants (Brazil, 2024a; 2025a).

The international scientific community has converged in recent years on a position Clear on AI and authorship. The *International Committee of Medical Journal Editors* (ICMJE, 2026) defines scientific authorship in health by four cumulative criteria: (1) contribution substantial for the conception or design of the study, or for the acquisition, analysis or (1) interpretation of the data; (2) substantive drafting or critical revision of the manuscript; (3) final approval of the version to be published; (4) agreement to be responsible for all of aspects of the work, ensuring that issues related to accuracy or integrity of Any part can be investigated and resolved. An AI system does not fulfill any Of the four: disapproves, does not respond, and does not take responsibility. The main publications

International medical bodies, including the WHO, have already established that AI should not be listed as a co-author, and is only permitted as a recognized tool (WHO, 2021).

Brazilian regulations, through CNPq Ordinance No. 2,664/2026, aligned themselves with this position.

Law No. 14,874/2024 and Decree No. 12,651/2025 consolidate this principle without Ambiguity: the principal investigator is responsible for the entire protocol, for all its stages and for all its results, including the stages in which it delegated content production or data processing to algorithmic systems.

CNPq Ordinance No. 2,664/2026 operationalizes this responsibility with three Specific requirements: identification of the AI tool used, description of its function. The role played by this tool at each stage of the research, and a description of the mechanism of Human verification of the generated results (CNPq, 2026). The third element is the most demanding and the least fulfilled. To declare that a highly demanding language system was used. The scale for drafting the protocol does not meet the requirements of the ordinance. The researcher must describe how... verified the accuracy, completeness, and ethical suitability of the content produced by the system. before submission to the Ethics Committee, and identify who performed this verification.

Every protocol should clearly state two specific points, the presence of which will be verified by the CEP. First: the protocol explicitly states whether or not AI was used. at each stage of the research? The absence of a statement in a protocol where the use of AI is... The likely or expected outcome should trigger a request for clarification before any assessment of... merit. The protocol's silence on the use of AI is not evidence that AI was not used. It is Missing information that prevents a complete ethical evaluation.

The second: the protocol describes the concrete mechanism for human verification of Results generated by AI? A statement of use without a description of the verification is a statement. Incomplete. The researcher who uses AI without verifying what it produced transfers it to the Ethics Committee, and eventually to the participant, the risk of a step in the protocol that no human being can do reviewed with identified responsibility. Human authorship in AI research is not It's not a nostalgic principle, nor resistance to technological progress. It's the only guarantee available. that someone is held accountable for what happens to the people who participate in the research.

6.3 Algorithmic Plagiarism, Negligence in Misconduct, and Fabrication of Evidence

Algorithmic plagiarism occurs when a researcher presents content generated by AI. as a result of my own work, without declaring the use of the tool and without critical control. Strict. Unlike traditional plagiarism, in this case there is no identifiable copying of a... a specific author, but the dilution of authorship through algorithmic mediation that It masks the origin of the content. The central risk lies not only in misappropriation, but in the silent transfer of scientific authority to a system that does not understand the meaning, truthfulness, or ethical implications of what it produces.

Large Language Models (LLMs) produce text. fluency that simulates technical authority. This fluency is precisely the risk factor. discussed in this chapter. A language system can generate bibliographic citations with Impeccable formatting referencing studies that never existed. Can you describe it with... Apparent accuracy of results from clinical trials that were not conducted. It can produce internally consistent literature reviews that systematically omit evidence. Contrary statements without indicating this omission. Methodological sections may contain inaccuracies. which no reviewer without specific technical training will be able to identify.

The problem is not the intention of the researcher using these tools. It's the Structural properties of systems: large-scale language models are optimized. to produce plausible text, not true text. Plausibility and truthfulness are not the same. synonyms. The distance between the two may be invisible to the reviewer and consequential for the participant (WHO, 2021; Figueiredo; Aguiar, 2025). The National Data Protection Agency (Brazil, 2024d), in its Technological Radar on Generative Artificial Intelligence, documented this structural property and recommended transparency parameters and Human supervision is a condition for responsible use.

Not all inappropriate use of generative AI in research constitutes misconduct. Scientifically, the distinction between three types of failure is important for guiding the institutional response. The first is a methodological error: the researcher uses a language system to write a section of the protocol and does not verify the result with the necessary rigor. There is no intention to Fraud, and the response is to correct and empower. The second is misconduct through negligence: the The researcher knows they should verify this, has the means and time to do so, and yet... does not check — whether out of haste, naive trust in the system, or the assumption that the The problem lies with another authority. This is the most common type currently, and it's what the Ordinance addresses. CNPq Resolution No. 2,664/2026 aims more forcefully, through the requirement of an explicit description of

verification mechanism. The third is intentional misconduct: the researcher presents synthetic data generated by AI as data collected from real participants, or subscribe as The author created a protocol generated entirely by AI without substantive review. Here, there is fabrication of evidence, formal investigation and legal accountability.

The Ethics Committee (CEP) is not an instance for investigating scientific misconduct, but it is the first instance. capable of identifying protocols that present structural risks to evidence production. compromised by the lack of human verification of the results generated by AI.

In practice, the ethical evaluation of protocols involving generative AI requires that the committee... Formulate four concrete questions in response to any proposal that makes use of these [reasons/reasons]. tools at any stage.

First: AI was used in drafting the protocol, the informed consent form, or any other document. Submitted to the committee? The absence of a declaration is not evidence of non-use. It is information that the committee should actively request.

Second: the results generated by the AI were verified by a human expert. with the competence to identify errors, hallucinations, and omissions specific to the field. Scientific research? Generic verification by a researcher without training in the area is not... It satisfies that requirement.

Third: Does the protocol use synthetic data generated by AI? If so, how was it validated? the statistical accuracy of these data in relation to the real population, and this validation is documented no protocol?

Fourth: the Informed Consent Form informs participants, in accessible language, about the role of AI. at any stage of the research that directly affects them?

Algorithmic plagiarism, negligent misconduct, and AI-powered evidence fabrication. Generative technologies are not risks of a distant future. They are part of the protocols that are arriving today. SINEP submission platforms.

6.4 Civil Liability and the Extended Chain of Liability

When conventional research causes harm to a participant, the chain of Responsibility is relatively clear: principal investigator, executing institution, and... When applicable, sponsor. AI research introduces a fourth link in this chain. The traditional approach does not include: the developer or supplier of the algorithmic system —

companies like OpenAI, Google, Anthropic, Microsoft, or national developers of medical software.

When the damage stems from a model flaw, from a structural bias in the architecture. Whether it's a hallucination of the generative system, the question of who is responsible remains unanswered. consolidated in the current Brazilian legal system. What the legal system clearly establishes. It is the researcher's position: the responsibility of the principal investigator remains, even when there is an algorithmic system mediating your decision and the damage. Choosing to use a tool is a human choice. Human choice carries human responsibility. The researcher who delegates research steps to AI systems without proper verification does not... transfers its responsibility to the system. It accumulates responsibility for the choice of Delegating without checking.

This logic outlines four concrete links in the extended chain of responsibility. The principal investigator is responsible for choosing the tool and for the human verification of the... The executing institution is responsible for the results and the final integrity of the protocol. governance structure, through access to secure systems and through the empowerment of researchers. The sponsor, when one exists, is responsible for the contractual terms that They enable or limit the use of AI in the research they fund. The system provider is responsible for product defects. However, there is still no defined criterion for civil liability for damages arising from automated decisions in environments clinical. While this gap persists, the researcher tends to absorb, individually, a risk that should be distributed.

When the protocol involves databases, CNS Resolution No. 738/2024 designates Two of the links in the chain are considered specific legal figures in research. The Controller, According to article 3, IV, it is the natural or legal person responsible for decisions regarding the data processing, which may involve being a sponsor, protocol researcher, or person responsible for... biobank or its designee. The Operator, as defined in Article 3, XIII, carries out the processing in The name of the Controller is responsible for use in accordance with the purpose approved by the CEP. Article 11 It establishes that the Controller's responsibilities are non-waivable, allowing for... Joint comptrollership. This appointment helps to reduce the uncertainty of Responsibilities within the scope of data governance are detailed in Chapter 7, which It addresses agents, roles, and rights in data-driven operations research. The remaining gap,

Therefore, it specifically concerns the link of the external provider of the algorithmic system.
not to the internal links already regulated by CNS Resolution No. 738/2024 (Brazil, 2024b).

Three mechanisms mitigate this concentration in current practice. Law No. 14,874/2024 and Decree No. 12,651/2025 establishes that the principal investigator is responsible for all of the protocol in all its phases (Brazil, 2024a; 2025a), but they acknowledge responsibility.
solidarity from the institution and the sponsor.

When research involves the clinical use of AI, see CFM Resolution No. 2,454/2026.
determines that medical responsibility for AI-supported decisions is comprehensive and
inalienable, it prohibits the direct communication of diagnoses to the patient by algorithmic systems.
and requires registration in the medical record (CFM, 2026). In the research chain, the participant's right
The human review stems from Article 20 of the LGPD (Brazilian General Data Protection Law), which guarantees the data subject the right to review by a human reviewer.
natural, decisions made solely on the basis of automated processing;
The protocol should include this information in the informed consent form, and the ethics committee verifies it.

The LGPD (Brazilian General Data Protection Law) also establishes, in articles 42 to 45, civil liability regarding data.
Personal data of participants processed by AI systems during the research, including when
This processing takes place on third-party platforms (Brazil, 2018). The obligations of
The governance, legal basis, minimization, and security aspects stemming from the LGPD (Brazilian General Data Protection Law) are being developed.
in Chapter 7, which addresses data governance as an operational dimension of
Civil liability is set forth in this section. CNS Resolutions No. 466/2012 (Brazil, 2013) and
No. 510/2016 (Brazil, 2016), which preserve their ethical relevance, preserve the principles that
These principles permeate the entire chain: beneficence, non-maleficence, autonomy, and justice.
Principles haven't become outdated with the adoption of AI; they've become more demanding.

6.5 WHAT THE PROTOCOL SHOULD CONTAIN — AND WHAT THE CEP VERIFIES

Every protocol using AI must contain five integrity requirements, and a sixth if...
Whether using proprietary or third-party databases, the CEP verifies the presence of each.
One. The absence of these elements determines whether the submitted document is evaluable in its
completeness. Of the five initial requirements, three are disqualifying. The lack of
Any one of them makes a merit-based analysis impossible and demands mandatory action.
Two others can be resolved by providing additional documentation. The sixth, when applicable, is...
Eliminatory: without it, the data chain of custody is not formalized.

The first (eliminary) criterion: there is an explicit statement regarding the use or non-use of AI in each Research phase? The absence of a declaration is not evidence of absence of use. It is information. which the committee should actively request before any merit assessment.

The second (eliminary) requirement: the declaration identifies the tool used, the function. What is the role of the individual at each stage, and who is responsible for verifying the results? (Declaration) A generic statement that mentions AI without specifying its use, function, and verification does not meet the requirements. from CNPq Ordinance No. 2,664/2026 (CNPq, 2026).

The third (eliminary) step: the protocol describes a concrete verification mechanism. human factors are involved in evaluating AI-generated results, including who performed the verification and what their role was. Do I have the technical expertise for that?

The fourth (remediable) criterion: the informed consent form was produced or reviewed by a human with the capacity to... to identify inaccuracies, omissions, and hallucinations specific to the clinical context of the research, and Does it inform participants about their right to review automated decisions (Article 20, LGPD)? The fifth (remediable) issue: Does the principal investigator sign the protocol with an express declaration of full responsibility for the content, including content generated or processed by AI systems?

Sixth: if the protocol uses a database, the Controller and the Operator (if any) are identified, and the institutional terms required by CNS Resolution No. 738/2024 are Presented? For banks established within the scope of the research, the minimum requirement is the Term. Data Use Commitment Agreement; in multicenter protocols or with the participation of more than from an institution, the Institutional Agreement Term is added.

For existing databases outside the context of research, the Resolution also requires... Institutional Consent Form signed by the head of the institution of origin of the data. The absence of applicable Terms means that the data chain of custody was not established. Formalized, and the protocol is not evaluable on its scientific merit. Chapter 7 describes the four Institutional terms (Agreement, Consent, Commitment, Transfer) in detail.

The absence of any of the qualifying elements does not constitute a minor issue. Resolvable by formal supplementation. Indicates that the protocol was submitted without the required level of human responsibility as outlined in Law No. 14.874/2024, Decree No. 12.651/2025, and Resolution CNS No. 738/2024 requires all research involving human beings (Brazil, 2024a; 2024b; 2025a).

Four dimensions run through the entire protocol and deserve specific attention. during the evaluation:

- In the bibliographic references: there are indicators of citations generated by AI without verification, such as references with inconsistent data, digital object identifiers nonexistent or descriptions of studies without verifiable correspondence in databases.
Indexed?
- In the methodology: the use of AI at some stage of the analysis is precisely described.
sufficient to allow for the assessment of the specific risks of that system at that time.
Clinical and population context?
- In the data: if the protocol uses synthetic data generated by AI, the statistical accuracy was validated, documented, and compared with the actual population that the data is intended to represent.
represent?
- In the Informed Consent Form: the language is accessible to the target audience of the research, the information about the
The role of AI in decisions affecting participants is present and the text was
Verified by a responsible human identified in the protocol?

These four dimensions do not require technical knowledge about AI systems.

They are used. They require precision in reading the protocol and a willingness to formulate the questions.
which the asymmetry of technical knowledge tends to suppress.

This chapter concludes with an operational distinction that deserves careful attention: the
The difference between an incomplete protocol and an inadequate protocol. An incomplete protocol...
presents remediable gaps: declaration of AI use missing, but requestable; mechanism for
Verification not described, but implementable; Informed Consent Form with inadequate language, but
Correctable. These cases allow for requests for additional information within a defined timeframe and
explicit criteria for supplementation.

The inadequate protocol has structural flaws that cannot be resolved by
supplementary documentation: absence of human responsibility identified in stages.
Criticisms of the research, use of synthetic data without validation presented as real data.
Collected from participants, informed consent form that omits the role of AI in decisions that directly affect them.
the participating individuals. These cases require a reasoned refusal based on the principles of Law No.
14.874/2024, Decree No. 12.651/2025, CNS Resolution No. 738/2024 and the requirements
from CNPq Ordinance No. 2,664/2026 (Brazil, 2024a; 2024b; 2025a; CNPq, 2026).

Distinguishing between the two accurately and with sound reasoning does not create an obstacle to research.
with Brazil. Far from it, it contributes to the only moment in which governance can still

to be complete. Chapters 7 and 8 extend this logic to data governance and to the proportional risk assessment of the system described in the protocol.

7. Data Governance and Participant Protection

Artificial intelligence transforms data into raw material for models that then become... to operate on a scale similar to that of the same population that generated them. This transformation requires Data governance designed for the specific nature of algorithmic data processing: that dissolve into the weights of the model, models that memorize fragments of what they saw, consents whose withdrawal does not exclude already trained data, clinical databases that cross Frontiers under the promise of innovation.

The Brazilian legal system has built, in less than a decade, the most comprehensive framework. detailed information about Latin America is relevant to this context. The LGPD provides the overall framework for data protection. of personal data. CNS Resolution No. 738, of February 1, 2024, operationalizes this. Architecture for research with databases involving human subjects. Law No. Law 14.874/2024 and Decree No. 12.651/2025 establish institutional responsibility for the data processed in research. CFM Resolution No. 2,454/2026 regulates the use of AI in medical decisions. This chapter translates this milestone into concrete questions that the protocol Research using AI in healthcare should provide an answer.

7.1 Health Data in AI Systems: A Different Nature

A relational database stores information in discrete fields and Addressable. A patient's record occupies a defined space and can be located. altered or eliminated with surgical precision. This logic guides most legislation.

Regarding the protection of personal data, including the LGPD (Brazilian General Data Protection Law). When an AI system enters the scene, This logic collapses.

Training a neural network works through a radically different mechanism. The algorithm processes the training data and extracts statistical patterns that are distributed across millions of mathematical parameters, known as model weights (Cota et al., 2024). The The original data is not stored as a file. It dissolves into the architecture of... A model, like how ink dissolves in water. A patient's clinical record ceases to occupy space. a locatable address and begins to inhabit, in a diffuse and mathematically inseparable way, the entire structure of the system.

This distinction has a direct consequence for the protocol design. Protocols of research that treats health data as raw material for training AI models. They operate in a legal and ethical category that conventional norms did not anticipate. A promise to delete the data after the research is completed is sufficient in a database. Relational, it becomes technically empty in a deep learning model. The researcher Anyone submitting such a protocol needs to acknowledge this difference before anything else. Another analysis. Large Language Models (LLMs) and networks. Deep neural networks have a property that distinguishes them from any other information system. previous: the ability to memorize fragments of the data with which they were trained and to reproduce them under specific consultation conditions. This memorization does not occur due to error. Design. It results from the optimization mechanism itself that makes these models effective. The more a model learns, the greater the risk that rare or unique data, such as records, will be lost. Clinicians treating less prevalent conditions should remain precisely focused on their parameters.

Classic pseudonymization does not protect against this risk. Replacing the name and the A patient's identification number, using an alphanumeric code, is a sufficient measure in a... relational database. In a deep learning model trained with data Rich clinical combinations of variables such as diagnosis, date of admission, municipality of Residence and test results allow for re-identification with a high probability, even without any direct identifier. The European Data Protection Board (European Data Protection Board) The Data Protection Board (EDPB), in its Opinion 28/2024 (EDPB, 2024), stated that the True anonymization of high-dimensional clinical data in AI models. Advanced development is virtually impossible with current technologies.

The risk of memorization has a direct regulatory consequence in Brazil. The Agency The National Data Protection Authority (ANPD) published the Technological Radar on generative AI in which recognizes this risk and recommends that any AI system trained with data sensitive personal information should be subject to a specific re-identification risk assessment before... entry into use (Brazil, 2024d). This guidance has a precise target: the protocol of Research must demonstrate that the risk of memorization was assessed, not assumed. Non-existent.

The Free and Informed Consent Form (TCLE) was designed for a world of locatable records. Their traditional models inform the participant about the purpose. of the research, the risks of the procedure, and the right of withdrawal. None of these clauses

It answers the questions that the use of data in AI model training requires.

The World Health Organization (WHO) recognized this in its ethics and governance guidelines.

AI for healthcare, which conventional informed consent fails to capture.

adequately address the specific risks of using personal data in learning systems.

machine (WHO, 2021).

The protocol submitted to the Ethics Committee needs to answer questions that the models
Conventional consent forms have never been formulated. The data will be used to train a
model or just for statistical analysis? The resulting model can be transferred to
third parties? If the participant withdraws consent after the training, what mechanism is in place?
Does the technician guarantee the exclusion of your data from the model?

Fornasier (2022) demonstrated that the absence of answers to these questions in
Brazilian research protocols create a zone of legal opacity that compromises both
The rights of participants regarding the ethical validity of the research.

The Ethics Committee (CEP) assesses whether the researcher understands what they are doing with the data.
People who entrusted the healthcare system with the most intimate information about their bodies.
This assessment requires a new set of questions. Chapter 8 of this guide, with the Matrix...
Risk Assessment in Artificial Intelligence in Research with Humans (MARIAH), and
The following sections of this chapter offer this repertoire.

The starting point, however, is the recognition that health data in a model
AI occupies its own ethical category. Treating it like any other data is the first and
the most serious mistake a protocol can make.

7.2 AGENTS, ROLES AND RIGHTS IN DATA-BASED RESEARCH

Data governance in AI research begins with accurately identifying who...
who. Resolution CNS No. 738/2024, in its articles 3, 9 to 15 and 16 to 18, names three
legal figures that structure the entire chain of responsibility for the database: the
Controller, Operator and Holder (Brazil, 2024b). These figures correspond to the
LGPD and, in the Resolution, gain specific operationalization for research.

The Controller is the natural or legal person responsible for decisions regarding
to data processing. In the context of research, this could be the sponsor, the researcher
The person responsible for the protocol is the person in charge of the biobank or a person designated by them.

Assigned responsibilities are inalienable, although they can be shared.

through joint control, but they do not allow its transfer.

The Operator performs the processing on behalf of the Controller and is responsible for its use. in accordance with the purpose approved by the CEP. The Data Subject is the individual related to the data, whether as a research participant or citizen whose information was collected in context. healthcare services and are now used in studies with existing databases.

This tripartite division has a direct practical consequence for the protocol. The protocol must to explicitly identify who the Controller is and, if applicable, the Operator. The absence of this Identification is not a formality: it indicates that the researcher has not mapped the chain of responsibility for the data you intend to use. In multicenter protocols or with For external partners, joint controlling must be formalized in a Terms of Agreement. Institutional, in accordance with article 3, XVI and paragraph 2 of article 12 of CNS Resolution No. 738/2024.

The research participant is also a Data Subject within the meaning of the LGPD (Brazilian General Data Protection Law) and the... CNS Resolution No. 738/2024. This overlapping of categories expands the range of rights. which are recognized to him. The rights of the Data Subject established in articles 16 to 18 of the Resolution. They converge on a fundamental principle, enshrined in Article 4: self-determination. Informative, understood as the right of each individual to control and protect their data. personal.

From this principle stem four specific guarantees: access to information. stored at any time; correction or updating of erroneously stored information inserted; partial or total removal of your information, with the withdrawal being valid from the date of entry; Express declaration; and compensation in case of damage resulting from misuse or breakage. security.

Law No. 14,874/2024 adds to this set the inalienable right of withdrawal from consent at any stage of the research (Brazil, 2024a). Added to this are the rights of correction of incomplete, inaccurate, or outdated data, as guaranteed by article 18, III, of Law No. 13.709/2018 (LGPD), with direct repercussions on the training bases of The model. The feedback of the results is also part of this set. At the end of the research, the The participant has the right to receive information about the study results and any potential consequences. findings that impact your health condition or your health advice, in form foreseen by the protocol. In conventional studies, this right is materialized by Exclusion of the participant's record from the databases and the interruption of any use.

The future of your information. In research that uses health data to train models.

In terms of AI, this mechanism doesn't work. The data already incorporated into the model's weights doesn't have...

Locable address. Cannot be deleted by administrative command.

Software engineering has developed a technical answer to this problem: the Machine unlearning . The technique consists of the selective removal of influence of a specific dataset on an already trained model, without the need to retrain the entire system from scratch, a computationally intensive procedure. prohibitive in large-scale models (Cota et al., 2024).

One of the most promising architectures to enable this research process in The health model is SISA (Sharded, Isolated, Sliced, Aggregated). In this architecture, the data of The training sessions are divided so that the influence of each participant is contained within a specific area. a specific fragment of the model, allowing for the withdrawal of consent technically executed and audited through retraining of only that fraction.

The existence of this technology has a direct consequence for the design of the protocol. The promise to delete the server data, sufficient in conventional searches, does not satisfies the right of withdrawal in protocols involving AI model training. The protocol It needs to describe the technical mechanism by which a participant's data influences the outcome. It may be removed from the model if consent is withdrawn. The protocol that brings A generic promise of exclusion, without describing the mechanism, offers a guarantee that... The system architecture cannot comply.

Article 20 of the LGPD guarantees the data subject the right to request a review of decisions. taken based exclusively on automated processing of personal data (Brazil, (2018). In the context of AI-based health research, this right takes on additional weight. Algorithms can define eligibility for experimental treatments, establish Risk stratification or guiding differential diagnoses. When a decision with this The magnitude of the issue stems from an opaque model; the right to review needs to translate into something more. more than the formal possibility of filing a request.

The loophole lies in the wording. The Brazilian LGPD does not require the review of decisions. automated tasks must necessarily be performed by a human being. Mariotto (2024) This demonstrated that this flexibility opens the door for one machine to audit another machine. unlike the European General Data Protection Regulation (General Data Protection Regulation) The Data Protection Regulation (GDPR), which expressly establishes the right to intervention.

human in the review process. For Fornasier (2022), this difference makes the Brazilian framework excessively permissive in the context of high-impact clinical applications, and transfers to CEP has the responsibility to supply, as required by protocol, what the law does not stipulate.

The protocol should define the specific mechanism for meaningful human supervision. (*human-in-the-loop*) for any decision with clinical consequences for the participants. Meaningful human supervision does not equate to the formal presence of a healthcare professional in the approval chain. It equates to that professional's actual ability to understand the The criteria of the algorithmic decision, challenge them based on clinical evidence, and override their... judgment based on the model's outcome. The protocol that provides for human supervision without guaranteeing This actual capacity offers nominal protection, not effective protection.

The right to know completes the set. The participant has the right to be informed. about what data was collected, for what specific purpose it was used, if it was shared with third parties and under what conditions this sharing occurred. In Research into AI gives this right dimensions that conventional models of consent is not achieved. The participant needs to know not only that their data has been used, but which were used to train a model that can operate at scale, be transferred to other institutions or serve as a basis for commercial products (WHO, 2021).

In other words, for this right to know to be exercised effectively, the Transparency needs to be incorporated into the very instrument that formalizes participation: the consent form. This is where the protocol should be translated, in clear language and verifiable, the intended use of the data and the model, as well as the terms of use, Sharing and withdrawal.

The informed consent form (ICF) for a protocol involving AI must include compatible additional information. with the characteristics and risks of the AI system.

- Description of the type of model to be trained and the specific purpose of the system. resulting.
- Identification of all recipients of the model, including business partners and international institutions.
- time frame for using the data and the model.
- Technical mechanism for withdrawing consent after training, with a description of The machine unlearning method adopted.
- Policy regarding the sharing of the model with third parties and applicable contractual conditions.

- An accessible channel through which participants can exercise their rights at any time during and after the research (Fornasier, 2022; WHO, 2021).

The AI protocol without these six pieces of information in the consent form is a incomplete protocol. The absence is not a formal detail. It indicates that the researcher did not map it. With precision, what will you do with the data of the people who entrusted the health system with their services? More sensitive information about your existence. Informed consent in research. With AI, you need to be as specific as the technology you intend to govern.

7.3 RESEARCH DATABASES: CONSTITUTION, FUTURE USE AND WAIVER OF CONFIDENTIAL INFORMATION

AI research for healthcare typically utilizes databases, with models. trained on large sets of clinical records, either from their own studies or obtained from databases. existing ones. CNS Resolution No. 738/2024 regulates this practice precisely and establishes a two-pronged approach that the protocol must follow before any merit assessment.

The first path, regulated in Chapter VII of the Resolution, deals with the establishment of Databases in the context of research. When a protocol calls for the creation of a database. During the execution of the research, the committee requires four descriptions: identification of Controller or Controllers; list of data and information to be collected; mechanisms that guarantee the confidentiality and security of information, including access and storage strategies; and criteria for sharing and data transfer. Any intention to share data or to anonymize it. Irreversible action requires specific justification, with documented risk assessment and benefits to participants (Brazil, 2024b, art. 19).

The second option, outlined in Chapter VIII, deals with the use of existing banks. constituted. This is, in practice, the most common situation in AI research in healthcare. The A protocol that intends to use data from an existing database faces a larger set of... requirements, which vary depending on the bank's origin. If the bank comes from another In the research, the protocol must include authorization from the Controller, a description of the data, confidentiality mechanisms, identification of the original protocol, Terms of Signed Data Use Agreement and Informed Consent Form (Brazil, 2024b, art. 24). If the bank is external to the research (such as hospital records, SUS databases or private collections), the

The protocol should include contextualization of the data's origin and the Consent Form.

Institutional statement signed by the head of the institution (Brazil, 2024b, art. 27).

This distinction has concrete implications for the protocol. A protocol that intends

To train a model on data from the Brazilian Unified Health System (SUS) or a private hospital without presenting a Terms of Reference.

Institutional approval is incomplete by definition, even if all the others

The documents must be correct. The Consent Form is not a formality: it is the act by which the

The institution that holds the data acknowledges the intended use outlined in the protocol and assumes responsibility.

Joint responsibility for the chain of custody.

Four institutional instruments structure this chain of custody in research.

with databases. The Institutional Agreement formalizes the joint controllership in

banks formed with the participation of more than one institution, with sharing criteria and

Destination of data in case of dissolution. The Institutional Consent Form attests that the

The institution holding the data authorizes its use in the protocol. The Commitment Agreement of

Data Use is the formal statement in which the principal investigator and their team agree to use the terms of their data.

They commit to secrecy, confidentiality, and the use of data for the approved purpose.

The Information Transfer Agreement formalizes the transfer of data between banks, and

assumes responsibility for safeguarding and respecting confidentiality. The absence of any of these

Terms, when required by the protocol's nature, transform data governance.

as a decorative figure.

Article 20 of CNS Resolution No. 738/2024 submits to SINEP all research that

intends to create a database or use a database created for other purposes, and

It establishes, in its paragraphs, the regime of consent. The rule in § 1 is consent.

prior consent from the participant. There are two situations in which new consent may be

Not applicable: when future use has already been authorized in the Registry or in the Informed Consent Form of the original research.

(§ 2) and when the data provided is anonymized by the Controller (§§ 3 and 5).

Paragraph 5 is crucial for AI research: in databases established outside the scope of research.

— medical records, administrative files and databases, the most frequent source of data for

Anonymization by the Controller is the only way to obtain a waiver. For banks, training is already available.

Established within the scope of the research, Article 25 details this system and lists five

Exemption situations: data originally collected without participant identification;

Prior consent from the participant for the use of data in future research;

sharing of anonymized data by the Controller; irreversible anonymization of

data in the original protocol (Brazil, 2024b). The scope of article 25 does not extend to banks. constituted outside the scope of the research, governed by article 27, whose exemption follows the rule of article 20, § 5. The distinction is operational: the postal code first classifies the origin of the bank. and only then does it verify the applicable normative hypothesis. This is crucial for AI research. Because most models are trained on previously collected data.

The exemption scenarios outlined in the applicable regulations require evaluation. individualized. The protocol must indicate the normative hypothesis invoked and present Ethical and methodological justification, with a description of the origin and purpose of the database, of The risks involved and the safeguards adopted. The bank's origin guides the identification of the applicable requirements, but do not, in isolation, constitute grounds for exemption. The decision It is up to the Ethics Committee and must be expressly justified.

When consent cannot be withheld and was not authorized in the research. Originally, the protocol must present a new Informed Consent Form (ICF), or that of the legal guardian, for each participant whose data will be reused. In large banks, this can be operationally unfeasible. The correct response to this unfeasibility is not to disregard the This is a requirement, but the research question needs to be reformulated to work only with data. originally anonymized or with prior consent for future use.

7.4 Data Sovereignty and International Transfer

Data sovereignty in healthcare expresses a political choice. It means that the State and The citizens of a country determine who accesses, processes, and benefits from the data. produced by its population. In the field of AI, this choice takes on a strategic dimension. Additionally: the health data of a population constitute the raw material for the models that, Once trained, they go on to guide diagnoses, screenings, and clinical decisions regarding this. same population (Santos *et al.*, 2025). Releasing this data without safeguards is tantamount to surrendering the ability to shape the intelligence that will act upon the health system itself.

Health data do not operate as a neutral input. Recent literature identifies a structural pattern of data extraction from the Global South by platforms concentrated in the North. globally, with direct effects on technological dependence and asymmetry of health power. Santos *et al.* (2025) describe this process as digital colonialism: peripheral countries They provide raw data, while global platforms concentrate infrastructure and capacity.

processing and capturing value. The model was trained with data from the Brazilian population. It returns as a foreign product. The public system loses the capacity to govern. knowledge produced from its own base.

Brazil has built a regulatory framework that recognizes this strategic importance. LGPD classifies health data as sensitive personal data and requires a robust legal basis for any international transfer (Brazil, 2018). Law No. 14,874/2024 and Decree No. Law 12.651/2025 reinforces institutional responsibility for data used in research. including when treatment takes place outside the national territory (Brazil, 2024a; Brazil, 2025a).

Matsushita and Almeida (2025) situate this normative set within a movement International regulatory convergence around common principles: purpose Specific; data minimization; transparency; security; accountability. The power This framework depends on a condition that no single standard guarantees: protocols. research that accurately identifies where the data goes, on whose behalf, and with what purpose. What guarantees?

Two episodes, one international and one national, illustrate what happens when this... The condition is not met. In 2015, Google DeepMind signed an agreement with the National Service. British National Health Service (NHS) to access clinical data of 1.6 million of patients from three London hospitals. The data included admission records, Test results, diagnoses, and medical records.

They were transferred without the patients' consent, without proper legal basis and without the owners knowing that their information would fuel the development of an AI system from one of the world's largest technology corporations (Hern, 2017). A The British data protection authority has declared that the Royal Free NHS Trust has violated the law. by failing to adequately inform patients about the use of their information. Powles and Hodson (2017) described the case as a warning for the algorithmic era in healthcare: a absence of independent ethical oversight and the speed with which the agreement was signed They created a precedent of transferring public health data to private interests. without accountability.

The second episode is Brazilian. In November 2020, in the midst of the pandemic, a Spreadsheet with data from approximately 16 million people diagnosed with or suspected of having COVID-19. Version 19 was accidentally made available in a public GitHub repository by an employee. outsourced private hospital that provided services to the Ministry of Health.

The spreadsheet contained full names, identification documents, addresses, clinical diagnoses and risk classifications, including data from the President of the Republic, of government ministers and their families. The exposure lasted at least six months before it was... detected by independent researchers and reported in the news.

Therefore, the ANPD initiated an investigative procedure, but the case was closed. without a substantiated sanction, in part because the Agency was still operating in a transitional phase. without applying administrative sanctions (Brazil, 2026d). The episode showed that the exposure of Brazilian clinical data is collected both through international agreements and routinely within the healthcare system. custody, including third parties, consultancies, repositories and environments of development. The chain of responsibility, when not explicitly outlined, is filled in by chance.

The two cases, read together, point to the same lesson. The pattern identified It is structural. Public and private institutions handle extensive clinical databases under a logic of... project, with fragile chains of custody, lacking independent ethical oversight dedicated to given and without sufficient contractual instruments. The public system loses control over the knowledge produced from its own base (Santos *et al.*, 2025) and about security daily use of the data that it should protect.

The Brazilian Unified Health System (SUS) produces one of the largest collections of clinical data in the world every day. development. This collection represents almost four decades of service to the entire Brazilian population, with epidemiological, ethnic and regional diversity that none Global private platform reproduces. Research protocols that provide access to these Data provided by international partners or technology corporations need to demonstrate, with precision, which neither the DeepMind agreement nor the 2020 spreadsheet exposure will be repeated in Brazil.

The research protocol that provides for the international transfer of health data needs to answer a minimum set of questions before any analysis of scientific merit:

- What is the legal basis that supports the transfer, under the terms of the LGPD (Brazilian General Data Protection Law)?
- Does the recipient country or organization offer an adequate level of data protection, recognized by the ANPD (Brazilian National Data Protection Authority), or does the protocol include specific contractual safeguards?
- Can the model trained with the Brazilian data be reused, transferred to third parties, or incorporated into commercial products?

- How will the Brazilian institution exercise its custodial responsibility over the data after the transfer, given that Law No. 14,874/2024 assigns a full duty of protection regardless of the place of processing (Brazil, 2024a)?

A protocol that doesn't answer these questions is incomplete. Without these...

In response, what is approved is what has not been evaluated.

Two technical alternatives emerge in the literature as ways to reduce the need for international transfer of raw health data. The first is the use of Synthetic data: artificially generated datasets that replicate statistical patterns from a real database without directly exposing patients' identifiable information. A second is transfer learning : models trained on national environment that exports only the learned parameters, not the original data, allowing adaptation to other contexts without circulating from the clinical base.

Santos *et al.* (2025) identify these strategies as alternatives to extractivism. data, by prioritizing the sharing of learning over raw data. Both They align with the principles of purpose, necessity, and prevention of the LGPD (Brazil, 2018).

These alternatives have limitations that the researcher needs to be aware of. Synthetic data They reproduce population patterns, but may not accurately capture rare subgroups. infrequent conditions or specific epidemiological profiles of populations such as Indigenous peoples and quilombola communities. A model trained on synthetic data. that underrepresent these groups reproduces, in the synthetic version, the same bias that the real data already shows. contained. The transfer of learning, in turn, assumes that the source model was trained on data sufficiently representative of the target population. When that The condition is not met; the transferred model carries the biases of the original context into the... new environment, with the added complication that these biases become more difficult to identify. (Borges *et al.*, 2026). The use of these techniques reduces risks, but does not eliminate them.

Data sovereignty in AI research is resolved through ethical and political decisions. prior to the technique. Evaluation of a protocol with international data transfer. That act defines whether clinical data produced by the SUS will cross borders under protection equivalent to that guaranteed within the national territory. This decision precedes any other. Choose between anonymization methods, model architectures, or tools. Contractual. Data sovereignty in research begins with the committee's opinion.

7.5 Federated Learning and Privacy Architectures

Federated learning addresses a specific problem of

Governance: How to train AI models with data from multiple institutions without...

none of them relinquish custody of their clinical records. The technical response reverses the

Conventional logic. Instead of the data migrating to the model, the model visits the data.

Each participating institution trains a version of the algorithm locally using its own...

collection. What circulates between the institutions and the central server are only the parameters.

Mathematical knowledge, not patient records. Clinical data remains where it was.

produced (Rieke *et al.*, 2020).

The model learns from various contexts without accessing individual data directly.

For health research in Brazil, this architecture has direct implications. A protocol that

Utilize federated learning; it can bring together, in a single model, the clinical learning of

university hospitals in the South, basic units in the Northeast and reference laboratories of

The Amazon, without any record crossing the boundaries of the institution that generated it.

The epidemiological diversity of the Brazilian Unified Health System (SUS) becomes a technical asset, not a logistical obstacle.

Brazilian reality, however, reveals a worrying asymmetry between potential

of the technique and its effective use. Borges *et al.* (2026) conducted a scoping review on

AI applications in Brazilian public health and identified, among 349 studies retrieved,

only six that explicitly used federated learning or transfer of

learning. The authors conclude that these techniques remain marginal in practice.

everyday life, despite its ability to enable AI solutions with computational resources.

limited and aimed at preserving data sovereignty. The adoption of these approaches occurred

mainly in international multicenter consortia, which indicates that collaboration

Interinstitutional collaboration and technical training are essential conditions for its dissemination within the public system.

Brazilian.

Federated learning represents a significant ethical advancement. A protocol that...

Adopting this approach reduces the risk of improper transfer of sensitive data and strengthens governance.

national regarding the clinical collection of the SUS (Brazilian Public Health System). The committee that receives a protocol with this

Architecture should recognize this merit and, at the same time, verify whether the technique has been described.

with sufficient precision to ensure that its benefits are real and not merely rhetorical.

Federated learning reduces the risk of clinical data exposure, but not the eliminates. This distinction needs to be clearly stated in the protocols. The parameters Mathematicians who move between institutions and the central server carry information about the data that generated them. Under specific conditions, researchers have demonstrated that These parameters can be reversed to reconstruct fragments of the original data, in a a procedure called a gradient inversion attack . (Geiping *et al.*, 2020). The data that did not leave the hospital may, therefore, be partially Reconstructed from what was removed.

The practical severity of this risk depends on factors that the protocol should describe: the size of the data batch used in each round of local training, the frequency of updates sent to the central server and the adoption or not of complementary techniques of protection. Differential privacy *introduces* calibrated noise into parameters before transmission, reducing the accuracy of potential reversal attacks. A *Secure aggregation* prevents the central server from accessing the parameters. individual data from each institution, processing only the aggregate result. Rieke *et al.* (2020) They indicate that federated learning only reaches its full protective potential when combined with these additional safeguards.

There is also a risk of another nature, less visible and equally relevant. Federated learning assumes that the data from each participating institution are sufficiently representative to contribute useful learning to the global model. When an institution concentrates patients with a specific demographic profile, the model learns disproportionately from this profile. The aggregate result may show inferior performance precisely in the populations underrepresented in the local data of each network node (Borges *et al.*, 2026). The problem of algorithmic bias does not disappear with the Federal architecture, migrate into it.

Homomorphic encryption addresses the point of An additional vulnerability that federated learning leaves open: the server of aggregation. Homomorphic cryptography is a set of mathematical techniques that allow perform operations on encrypted data without needing to decrypt it at all. moment of the process. The aggregation server receives encrypted parameters, operates on them. and returns an encrypted result to the institutions. It never accesses the content. The data remains protected even from the system that processes it.

Even if the clinical data remains within the institutions, the server needs to process the received parameters to generate the global model. A compromised server, controlled by a malicious actor or subject to a court order in a foreign jurisdiction, You can explore these parameters.

The application of this technique to Brazilian public health already has national evidence to support it. Lunkes, Lunkes and Oliveira (2024) demonstrated the feasibility of homomorphic cryptography in Analysis of dengue epidemiological data using the TensorSeal library with Python. execute models on encrypted data without decryption at any stage of the process. processing.

The work, developed at the National Laboratory for Scientific Computing (LNCC), linked to the Ministry of Science, Technology and Innovation, indicates that Brazil has installed technical capacity to advance in this direction in research with sensitive data of health. The computational cost still limits large-scale applications, and the technique remains at the forefront of research for routine clinical use. This horizon has practical implications. Immediate: Protocols that combine federated learning with homomorphic cryptography. They offer the highest level of protection currently available for clinical research data. collaborative with AI, and deserve special recognition in ethical evaluation.

Five questions structure the evaluation of a protocol with federated learning.

First: who controls the aggregation server and under what jurisdiction? A server Controlled by a foreign partner, it partially recreates the problem of sovereignty that... The second question: what complementary techniques does federated architecture aim to solve? Protective measures have been adopted, such as differential privacy, secure aggregation, or encryption. The third: the participating institutions. The consortium's actions adequately represent the epidemiological diversity of the target population. including peripheral regions, indigenous peoples, and quilombola communities? The fourth: the The model added at the end of the research may be transferred to third parties, licensed, or incorporated into commercial products? Fifth: the Informed Consent Form informs participants about this Distributed training architecture, in accessible language and without omitting the model's destination. resulting?

The protocol that answers these five questions accurately provides a basis. sufficient for a well-founded assessment. The protocol that omits them or answers them in a way that

The generic form indicates that the researcher did not map the risks with the depth that...

The technique requires it.

The asymmetry identified by Borges *et al.* (2026), that is, federated learning marginal in Brazil and adopted mainly in international consortia, reinforces the responsibility to describe the technique accurately: the less widespread the technique, The less familiar researchers are with its ethical requirements, the greater the likelihood of protocols that invoke federated learning without fulfilling the conditions that make it so. effectively protective.

7.6 WHAT THE PROTOCOL SHOULD CONTAIN AND WHAT THE CEP VERIFIES

The Personal Data Protection Impact Assessment (DPIA) is the instrument provided for. in article 38 of the LGPD to document the processes of processing personal data that They can generate risks to civil liberties and fundamental rights of the holders (Brazil, 2018).

The ANPD (National Data Protection Authority) published specific guidelines on the RIPD (Data Protection Impact Assessment), which establishes that the document must...

It should contain, at a minimum, a description of the types of data collected; the methodology used for the collection and security of information; and the controller's analysis with respect to the measures, safeguards and risk mitigation mechanisms adopted (Brazil, 2026d). In Research using AI that utilizes health data, the RIPD ceases to be a recommendation and becomes... Obligation. The processing of sensitive data on a large scale for training purposes. The algorithmic nature and potential for re-identification perfectly illustrate the high-risk scenario that... The law aims to map the damage before it occurs.

The RIPD for AI research has a different structure and requirements than the RIPD. conventional. The ANPD (Brazil, 2024d), in its Technological Radar on artificial intelligence Generative engineering acknowledges that large-scale language models present risks. Specific memorization and re-identification techniques not found in traditional models. impact assessment.

A suitable RIPD for AI research should address four dimensions that... Conventional models do not consider: the risk of sensitive data being stored by model during training; the risk of re-identification of participants from Combinations of clinical variables even after pseudonymization; the risk of data drift. (*data drift*) when the model is applied to populations different from the training set; and the risk

Improper transfer of the model to third parties after the research is completed. The absence of Any of these dimensions in the RIPD submitted to the CEP indicates that the document was prepared without considering the technical and ethical specificities of AI research.

The RIPD submitted with an AI research protocol should map the flow of Data from start to finish. This mapping encompasses at least six stages: data collection. clinical practices and their legal basis; pseudonymization or anonymization and the technical method adopted; the storage and security measures applied; model training and risks Memory skills are assessed; sharing with partners and contractual safeguards. applicable; and the fate of the data and the model after the research is completed.

A RIPD that describes only the initial collection and storage steps, without Mapping what happens to the resulting model covers the least risky phase of the cycle and omits the most critical phase. The document must accompany the data until the end of its useful life.

The search is conducted, and the postal code verifies if it does so.

Evaluating an AI research RIPD does not require in-depth technical knowledge in AI. Critical, objective questions linked to the interests of the participant and the AI are sufficient. Society: Does the document describe the complete data lifecycle? Does it identify the risks? Specific re-identification criteria for the studied population? Assess the risk of memorization by The model? Describe the technical safeguards adopted and demonstrate their suitability to the risks. Have the risks been identified? Have the residual risks been quantified after the safeguards? Have responsibilities been defined? Clear guidelines in case of a security incident? A RIPD that responds affirmatively to these Questions with documentary evidence meet the minimum standard for AI research in health. A RIPD that ignores them or responds to them generically signals that the researcher He did not understand the risks of what he is proposing.

In addition to RIPD, the AI research protocol that uses health data requires a a specific set of documents that conventional submission templates do not They contemplate. This set articulates four technical documents with the four Terms. institutional guidelines of CNS Resolution No. 738/2024. The first technical document is the plan of Data management, which describes the complete lifecycle of collected information: origin, format, interoperability standard, storage location, access controls, deadline retention and destination after the research ends. The plan should cover not only the raw data, but the trained model and the intermediate parameters generated during the learning process.

The second document is the Institutional Agreement Term, which is required whenever the protocol involving more than one institution in the joint establishment or use of the bank data. The agreement defines the respective responsibilities, the sharing criteria and the Destination of data in case of future dissolution. In protocols with AI, the instrument It takes on an additional dimension because sharing can include the model itself. resulting data, not just the source data. CNS Circular Letter No. 1/2021 already provided guidance on Ethics Committees on the need to verify data sharing instruments in research. collaborative (Brazil, 2021); Resolution No. 738/2024 consolidated this orientation in formal requirement.

The third document is the Institutional Consent Form, which is mandatory in protocols. that use databases created outside the scope of the research, as per article 27 of Resolution No. 738/2024. Approval is granted by the head of the institution from which the data originates. are sourced from and attest to the authorization for the intended use, and identify the position, function and Signature of the responsible party.

The fourth document is the Data Usage Commitment Agreement, by which the The lead researcher and their team are committed to maintaining confidentiality and secrecy. data and its use for the purpose foreseen in the research. The Term is added, when applicable. Information Transfer Agreement, required when data is transferred between banks or institutions.

The scope of this collection of documents is proportional to its volume and sensitivity. from the research data. Research that processes large volumes of health data, such as population databases, image repositories, medical records at scale or monitoring Continuous use of devices requires the complete set described above. In research on For smaller volumes or less sensitive data, the documentation requirements are proportional, without Reducing the rigor of the analysis. Proportionality grades the requirements demanded by the protocol. Not the ethical principles of evaluation, which remain the same for all research.

The Informed Consent Form (ICF) with the specific clauses for AI research described in section 7.2. It remains mandatory when the exemption does not apply under any of the circumstances. provided for in articles 20, paragraphs 2, 3 and 5, and 25 of Resolution No. 738/2024. WHO (2021) recommends that protocols using AI include adapted consent forms that Describe, in accessible language, the specific risks of using personal data in

Machine learning systems. In a protocol with AI, a generic informed consent form sets up a Consent that does not inform.

When the protocol provides for international transfer of data or models, a This legal instrument for contractual safeguards is added to this set. The LGPD allows three Mechanisms for the international transfer of sensitive personal data: the transfer to countries with a level of protection recognized as adequate by the ANPD; the adoption of clauses standard contractual agreements approved by the ANPD; or the development of global corporate rules. approved by the competent authority (Brazil, 2018). The protocol must indicate which If the mechanism applies, present the corresponding instrument and demonstrate that its The clauses cover the use of the model after the research is completed. An agreement that protects the Raw data, but silence about the trained model protects the raw material and releases the... product.

Governing an AI protocol applied to healthcare requires more than just documentation; It demands transparency regarding decision-making responsibilities at each stage of the life cycle of research. The Wilton Park report (2026), produced in association with WHO and Co-Develop, proposes a practical framework of responsibilities for digital health infrastructures that... This applies directly to the ethical evaluation of protocols using AI.

The chart organizes four distinct functions: who is directly responsible for execution; who is accountable for the results; who should be consulted before the decisions; and who should be informed about the progress. In protocols with AI, this Mapping needs to cover at least six dimensions: control over the data; Accountability in case of security incidents; validation of the technical architecture; supervision of model results; authorization for integration with external systems; and Accountability to participants and the implementing institution. The protocol that does not map These responsibilities operate without real governance, regardless of sophistication. technique of the proposed system.

The same report identifies vendor lock -in as Persistent risk in digital health systems. Proprietary platforms offer immediate convenience, but they create long-term dependence and limit auditing. They reduce the State's ability to correct course (Wilton Park, 2026). This risk has Precise ethical address. A platform-dependent AI research protocol. closed foreign domain, without access to the source code, without auditable documentation of the architecture.

and without a data and model portability clause, approves along with the research a
A structural limitation to any future ethical oversight. The protocol must provide for an architecture.
auditable, open standards, and an exit mechanism that preserves institutional sovereignty over
The data produced by the research is analyzed, and the Ethics Committee verifies whether these guarantees are present.

Data governance in AI research doesn't begin with contracts.
Transfer, in storage servers, nor in technical protection mechanisms.
It begins the moment you read a protocol and ask yourself: does this researcher know what?
Are you asking people to participate in this research? Do you know what you will do with the data?
Theirs? Do you know where they will go? The chapter that the reader concludes here offers the tools.
Conceptual, legal, and practical aspects are needed to provide a well-founded answer to this question. The SUS (Brazilian Public Health System)
It serves the entire Brazilian population. The data generated by this service belongs to
to these people. The Ethics Committee exists to ensure that AI research treats them with dignity.
that this affiliation requires. Chapter 8 presents the Intelligence Risk Assessment Matrix.
Artificial, a practical instrument that translates this governance into a proportional risk assessment.

8 MARIAH: ARTIFICIAL INTELLIGENCE RISK ASSESSMENT MATRIX (AE VERSIONS)

B)

8.1 The Construction Process and New Issues for Ethics Committees in SEARCH

Chapter 8 can be consulted sequentially or by section, as appropriate.
reader's need.

The construction of an ethical evaluation instrument for research protocols that
The use of artificial intelligence requires, before any methodological decision, a choice of
Stance: Does the instrument exist to support the judgment of the Ethics Committee or to replace it?
This guide answers that question unambiguously. The matrices presented in
Parts III (Version A) and IV (Version B) of MARIAH are tools to SUPPORT deliberation. They
These are not new rules, and they do not create obligations beyond those already established within the System.
The National Ethics Committee for Research (SINEP) guidelines are not binding. Their adoption is a
institutional choice of each postal code. What they offer is precision: a path that makes the
a more systematic, better documented, and more justifiable assessment in the eyes of the participants
research and society.

This stance is not a concession, but a direct consequence of what...
International experience has demonstrated this. None of the eight regulatory models evaluated
managed to create a universally effective instrument for all institutional contexts.
planned.

The Canadian Algorithmic Impact Assessment model is robust, but it was
Designed for government public services, not for research involving human subjects.
(Canada, 2024). The guidelines of the Australian National Health and Medical Research Council.
(National Health and Medical Research Council — NHMRC) are accurate in the treatment of
life cycle, but they do not address the specific legal aspects of data protection in
Middle-income countries (Australia, 2023). Guidelines of the Indian Council of Medical Research.
(Indian Council of Medical Research — ICMR) are the only ones created to guide Ethics Committees in
evaluation of protocols with AI, but they were developed for a different context than
Brazilian (ICMR, 2023). Each model offered real contributions and presented real gaps.
MARIAH is the synthesis of these contributions, anchored in Brazilian law and guided by...
context of SUS.

The journey that led to the creation of the two versions that will be presented at the following sections were deliberately lengthy. The analysis considered five axes of evaluation internationally recognized and, during the research, included contributions such as the concept Opportunity to Override, which expresses the actual condition of the Professionals can challenge the system's outcome, based on the Korean experience (MFDS, 2022). The right The participant's right to request the deletion of their data from the model's training stems from... Indian model (ICMR, 2023). The Algorithm Change Protocol (Algorithm Change Protocol), which deals with the researcher's responsibility to predict how the model will perform. retrained, it stems from Indian and Japanese regulatory experience (CDSCO, 2025; PMDA, 2023). The concept of context of use as a mandatory preliminary step to risk classification. This is derived from guidelines from the U.S. Food and Drug Administration (FDA, 2025). Each one These elements expanded and improved the final instrument.

The result of this process is two versions of the same instrument, with the same ethical foundation and the same set of essential questions.

The corresponding operational material is in Part II of MARIAH.

The ethical evaluation of AI protocols cannot be reduced to a mere formality. Neither version of the matrix acknowledges this risk.

AI absorbs biases from training data and uses opaque logic. and generates impacts on participants that no other contemporary method produces with Equal scale and speed. Evaluating a protocol with AI is not evaluating a technology, but the impact that this technology can have on real human beings, in contexts concrete, with concrete vulnerabilities. The matrices in this guide exist to make that... more rigorous evaluation. The responsibility for the evaluation remains entirely with the

Postal codes.

8.2 FUNDAMENTALS, SOURCES AND DESCRIPTION OF MATRICES

This section presents the theoretical, normative, and methodological foundations that support the Two versions of MARIAH. Each design choice is based on experiences documented regulatory frameworks within the Brazilian legal framework. This applies to the axes, the Questions, weights, and logic for arriving at the risk level. The description of the two versions is... organized into independent subsections. Section 8.2.1 deals with the Qualitative Version A.

Section 8.2.2 deals with the Quantitative Version B. Both versions share the same ethical basis and... same set of essential questions. What distinguishes them is the logic of operation. That The distinction determines the audience and context for which each version is most suitable.

MARIAH is a proportional gradation instrument. It operates on a spectrum that... It ranges from low risk to critical risk and acknowledges that most AI-based protocols occupy Intermediate positions that allow for safeguards, conditions, and mitigations. This Gradation has a limit that the instrument cannot overcome. Three dimensions are not gradable. by MARIAH: direct and potentially irreversible harm to the participant without a mechanism documented interruption and repair; lack of effective human supervision in context in which the system influences critical clinical decisions; and technical uncertainty does not documented information about the system's behavior in the stated context of use.

Given these dimensions, the principle of non-maleficence and the precautionary principle, established by Law No. 14,874/2024, prevail over any result that the sum of blocks produce (Brazil, 2024a). The matrix structures the CEP judgment and does not exhaust its Ethical limits. These limits are set by principles established by law, and no one... The assessment tool has the capacity to relativize.

8.2.1 Qualitative Version A

The ethical evaluation of research protocols that use AI requires instruments that translate the technical complexity of algorithmic systems into measurable ethical impact. about the participant. The qualitative MARIAH Version A was constructed from this premise, based on eight international regulatory experiences and anchored in Brazilian legal framework.

The choice for a qualitative approach

Version A of the Matrix adopts a structured qualitative logic: response questions. binary, organized into five thematic axes, traversed sequentially by means of a flowchart. This choice is based on the experience of the Australian model, which demonstrated that Ethical evaluation instruments are most effective when they guide the evaluator's judgment. without replacing it. The evaluation does not require mastery of the system's technical architecture, but rather...

understanding the impact it can produce. Qualitative logic preserves this.
centrality of ethical judgment, while ensuring systematicity and
traceability to the evaluation process (Australia, 2023).

The binary question format, in which any risky answer increases the...
The classification on the axis was adapted from the Canadian Algorithmic Impact Assessment model.
(Algorithmic Impact Assessment — AIA). The Canadian AIA has demonstrated, in consolidated use
Since 2019, objective and well-formulated questions have been effective tools for...
evaluators without technical training in computer science. The Brazilian version adopts this
It applies logic and expands it to the research scenario involving human beings. It directs the
attention to the provision of government public services for the protection of participants
from these studies (Canada, 2024).

The input filter and the characterization of the usage context.

Step 0 (input filter) distinguishes routine uses from uses that require
in-depth evaluation. This distinction is essential. Without an initial filter, the committees analyze
indiscriminately ranging from bibliographic organization tools to systems
autonomous diagnostic support. This compromises the efficiency and proportionality of
process. The FDA guideline from the United States of America (USA), published in January of
In 2025, it was demonstrated that no algorithmic risk assessment begins without defining...
It requires context of use. This refers to the specific role of the system and the scope of the question.
that it was developed to respond to (FDA, 2025). The Matrix incorporates this requirement as
A mandatory preliminary step for all five axes.

The five axes and their sources

Axis 1, dedicated to the nature and autonomy of the system, articulates two concepts.
complementary. The first is the influence of the model: how much AI weighs in the final decision,
in relation to other sources of evidence (FDA, 2025). The Ministry of Food Safety and
The South Korean Ministry of Food and Drug Safety (MFDS) defines two
Criteria (MFDS, 2022): Potential for Malfunction, that is, the chance of algorithmic failures.
cause clinical harm; and Opportunity to Contest, which is the real possibility of

The professional questions the system's outcome. The combination of these two dimensions allows to evaluate not only what the system does, but what happens when it makes a mistake. The most important right now is knowing who has the power to intervene.

Axis 2 examines the impact on the participating individual. Its basis is found in Law No. 14,874, of May 28, 2024, and in Decree No. 12,651, of 2025, which establish the fundamental ethical principles of research with human beings in Brazil (Brazil, 2024a; 2025a). The axis includes a specific question about vulnerable populations: children, pregnant women, the elderly, indigenous peoples, and people in situations of socioeconomic vulnerability. This inclusion addresses a gap in the international models analyzed. None of them was designed for the Brazilian epidemiological and social context, marked by inequalities pronounced regional characteristics and the ethnic and cultural diversity of the population served by the Unified Health System (SUS). The axis also incorporates questions about exclusion mechanisms. Model training data. This question is based on Indian guidelines. Council of Medical Research (ICMR), which establishes the right of the participant to request deletion of their data from the repository used by the AI at any time (ICMR, 2023). None of the Western models analyzed provide for this right.

Axis 3, on data sensitivity and governance, is the axis most directly... anchored in Brazilian law. The LGPD classifies health data as personal data. sensitive and requires a specific legal basis, stated purpose, and identification of the Data Protection Officer. Data for its processing. This set of requirements is more protective than legislation. data protection laws of all countries analyzed in this guide, with the exception of Regulation General Data Protection Regulation of the European Union (GDPR).

The contrast becomes even more evident in relation to the Data Protection Law. India's Digital Personal Data Protection Act of 2023, which eliminated the category of sensitive data was removed from the legal text and health data began to be treated like any other data. other personal data (Burman, 2023). Equating medical records, genetic data and registries Comparing psychiatric data to current commercial data exposes research participants to risks of Discrimination without legal basis for differentiated protection.

Brazilian law, LGPD, makes the opposite choice. Health data constitutes a sensitive category. with enhanced protection. This distinction is one of the pillars of Axis 3 of the Matrix. The axis includes still specific question about risk of re-identification. The MRCT Center and WCG (2025)

They documented that large language models are able to infer identities from Data that appears to be anonymized. This risk does not disappear with formal anonymization.

Resolution CNS No. 738, of February 1, 2024, operationalizes the LGPD in specific context of research with databases involving human subjects. Defines the legal figures of the Controller, the Operator and the Holder (art. 3, items IV, XIII and XX) and It establishes the four institutional terms that structure the data chain of custody: Agreement, Consent, Commitment to Use and Transfer (Article 3, XVI to XIX). The Resolution It also distinguishes two situations for the use of databases — those created within the scope of research (Chapter VII) and those already established outside of that scope (Chapter VIII) — with requirements distinct documents, and regulates the future use of data and the hypotheses for exemption from informed consent (art. 20, §§ 2, 3 and 5, and art. 25). MARIAH brings together these requirements in two dedicated subsections to databases — Axis 3.b in qualitative Version A and Block 6.b in quantitative Version B — which incorporate the elements into three groups of operational questions: the protocol Does it identify the database controller? Does it present the institutional terms required by... nature of the bank (especially the Institutional Consent Form, mandatory in banks) outside the scope of the research)? Presents, when a request for exemption from informed consent is made, the normative hypothesis invoked and sufficient ethical and methodological justification (art. 20, § 5, or art. 25)? (Brazil, 2024b).

The corresponding operational material is in Part II of MARIAH.

Axis 4, on representativeness and transferability, addresses one of the main... Ethical gaps identified in the international literature: trained models predominantly with data from high-income populations and the northern hemisphere They consistently perform worse when applied to populations of different ethnic, socioeconomic and geographic characteristics (WHO, 2024). The guide The Australian and Korean MFDS guidelines documented the absence of local validation. This is one of the main ethical risk factors in protocols that import developed systems. in other contexts (Australia, 2023; MFDS, 2022). The axis also incorporates questions about *Data drift*. Machine learning models degrade their performance as... Real-world data diverges from training data. This phenomenon requires Continuous monitoring throughout the entire system lifecycle. The question has based on the life cycle maintenance concept developed by the FDA (2025) and

operated by the IDATEN system of the Medical Devices and Products Agency

Pharmaceuticals of Japan (Pharmaceutical and Medical Devices Agency — PMDA, 2023).

Axis 5 examines human oversight and explainability. Its central premise is... the principle of irrevocable medical autonomy as stated by the Indian ICMR (2023). Artificial intelligence serves to enhance diagnosis. It does not replace clinical judgment. The principle is formulated for clinical practice. In research, it focuses on the design of the supervision described in the protocol: what the Ethics Committee evaluates is not the future care act, but whether the protocol ensures human review and challenge before the execution of each supported decision. The protocol ensures human review and challenge before the execution of each supported decision. through the system. Sovereignty over exiting the system, in this context, rests with the researcher. responsible, without prejudice to the responsibility of the assisting professional when there is an act. The clinical setting is the next stage of care. The principal investigator is responsible for verifying the findings. Human responsibility for all results generated by AI. This principle is linked to the concept of The *human-in-the-loop* element present in the Australian model (Australia, 2023).

It aligns with the FDA's requirement that, when the context of use involves a Humans are involved in the decision-making process; the system's performance evaluation should consider the outcome. from the human-AI team, not just the isolated performance of the model (FDA, 2025). The axis It also includes a question about the model's retraining protocol. This question has based on the Indian Algorithm Change Protocol, which requires manufacturers to submit a plan in advance describing how the model... This will be updated. This requirement was established by the Central Control Organization of Indian Drug Standards (CDSCO, Central Drugs Standard Control Organization) (India, 2025). In the context of this guide, this principle translates to It is the researcher's responsibility to inform the Ethics Committee how and when the model will be implemented. modified during the study.

The convergence of some axes around the clinical impact on the participant is intentional, not inadvertent overlap. The axes of nature and autonomy of the system, of The impact on the participant and the transferability and performance assessments are dimensions... distinct aspects of the same reality: what the system can do to the person who participates in research. The nature axis examines how the system operates. The impact axis examines the what the system produces on the participant. The transferability axis examines whether what the A system that produces results in one context will remain valid in another. These are different questions about...

same risk because risk has multiple dimensions that a single axis cannot capture. One
A protocol that presents a high risk in all three areas simultaneously is not penalized.
Repeatedly, the same problem occurs. This demonstrates that the risk is consistent, structural, and unmitigable.
for a single safeguard. The convergence between axes in a high-risk outcome is evidence.
of gravity, not a methodological artifact.

The logic of the sequential flowchart and the critical rule.

The sequential flowchart was adopted for two reasons. The first is pedagogical.
The sequence guides the evaluation in a progressive manner, from the simplest to the most complex, and avoids
leaps that compromise the consistency of the analysis. The second is ethical. The level of risk only
It goes up, it never goes down. This logic is consistent with the precautionary principle that underlies the...
Ethics in research with human subjects (Brazil, 2025a). A single axis at Level IV is sufficient.
to elevate the entire protocol to the highest level of rigor, regardless of
Performance in other areas. The absence of human supervision in a system that
It influences therapeutic decisions, for example, but is not compensated for by good responses in
data axes or representativeness.

The critical rule originates directly from the Canadian Impact Assessment model.
Algorithmic (Canada, 2024). The model demonstrated that allowing mitigations to reduce the
The level of risk creates a perverse incentive. Researchers are now declaring safeguards for
They obtain more favorable ratings. They fail to demonstrate that the risk was effectively...
controlled. The rule solves this problem with a clear separation. The risk level is
determined by the characteristics of the system and the protocol. The required specifications are
adjusted based on the guarantees demonstrated. This separation preserves the integrity of
The instrument and the clarity of communication between the researcher and the committee are key.

8.2.2 Quantitative Version B

Version B of MARIAH uses questions with varying numerical weights in seven blocks.
thematic. The sum of the points determines the risk of the protocol on a scale of I to IV. This
This approach makes your decisions more traceable and auditable. The scoring identifies areas of
This text discusses risk concentration and shows how the researcher's choices affect that risk.

The choice for a quantitative approach

The quantitative logic is inspired by the Canadian Impact Assessment model. Algorithmic (Canada, 2024). The model operates with 65 risk questions, maximum score of 169 points, and 41 mitigation questions, maximum score of 75 points, distributed into four risk levels.

The Canadian experience demonstrated two concrete advantages of the model. Quantitative. The first is precision: the score identifies in which dimensions the risk is present. The second is the one that is focused and guides the committee's requirements proportionally. Consistency: protocols evaluated at different times and by different members of the Ethics Committee. They become comparable, which strengthens the consistency of institutional decisions (Canada, 2024). Version B incorporates this logic and adapts it to the context of research with living beings. Humans in Brazil. The focus shifts from governmental public services to...

Protection of research participants and compliance with the LGPD (Brazilian General Data Protection Law).

This guide proposes a structure in seven blocks: Project, System, Algorithm, Decision, Impact, Data, and Mitigation. The blocks with the greatest ethical and regulatory weight receive questions with a higher individual value. This choice makes explicit the hierarchy of priorities of The Impact and Data blocks together account for 139 of the 275 maximum points. from the matrix. They represent 50.5% of the total score. This concentration expresses a choice deliberate. In research with human subjects, what matters most is what the system can do. to the participating individual and how their data is handled.

The input filter and the characterization of the usage context.

Version B retains Step 0 and the characterization of the usage context from Version A. as mandatory steps prior to scoring. The FDA (USA) guideline demonstrated that No algorithmic risk assessment begins without a precise definition of the context of use. the specific role of the system and the scope of the question it was developed to answer. (FDA, 2025). Characterization does not score. It delimits the object of the evaluation and prevents the The system should be evaluated in the abstract, detached from the specific protocol submitted for analysis.

The seven blocks and their sources

Block 1 assesses the protocol's suitability for the use of AI. It examines transparency. in the system description, the team's qualifications, the clarity of the methodological rationale, and the Adequacy of the Informed Consent Form. The maximum score is 22.

Points. Each question is worth three points, except for the question about technical consultation. External, which is worth four. The block operates as a document completeness verification.

The protocol should describe the system's development stage: discovery. Exploratory (*Discovery*), clinical translation (*Translation*) , or deployment (*Deployment*). This This requirement is based on the MRCT Center and WCG (2025) framework, which demonstrated that ethical risks vary depending on the stage of the AI lifecycle in which it operates. One A system in the exploratory discovery phase presents different risks than a system already in operation. Clinical implementation. This distinction should be visible in the protocol submitted to the Ethics Committee.

Block 2 assesses the technical nature of the system: autonomy, adaptability, Explainability and traceability. The maximum score is 17 points. Each question is worth Two or three points. The block articulates two concepts developed from different frameworks. The first is the influence of the model: how much AI weighs in the final decision compared to other factors. sources of evidence. This concept was developed by the FDA (2025). The second is that of potential for failure (*Malfunction Potential*) and opportunity to contest (*Opportunity to Override*): the probability that algorithmic failures will cause clinical harm and the existence of Real conditions that allow the responsible professional to question the system's results.

This concept was developed by the Ministry of Food and Nutrition Security. South Korean medicines (MFDS, 2022). Direct risk questions are worth three points. Each. They deal with situations without mandatory human intervention, adaptive systems, and... Opaque systems. Their proximity to direct risk to the participant justifies the greater weighting. Questions regarding lack of guarantee are worth two points. They deal with documentary deficiencies that... Although serious, they are more easily corrected.

Block 3 assesses the quality, origin, and representativeness of the data. Training and testing. The maximum score is 14 points. Each question is worth two points. The bloc is directly based on the Australian guide and the guidelines of the Korean MFDS, which They documented that the absence of local validation is one of the main ethical risk factors in protocols that use systems developed in other contexts (Australia, 2023; MFDS,

2022). The block includes a question about independence between training and test data. A FDA guidelines state that test data must be strictly independent of development data so that the model's performance evaluation is valid (FDA, 2025). It also includes a question about representation by population subgroups. A The World Health Organization has identified that models trained predominantly with Data from high-income populations and the Northern Hemisphere show good performance. systematically lower when applied to populations with different characteristics ethnic, socioeconomic and geographic (WHO, 2024). This gap is especially relevant for the Brazilian context.

Block 4 assesses the consequence of the decision. The maximum score is eight points. The first two questions are worth three points each. This asymmetry is deliberate. The questions This section deals with the situations of greatest potential severity throughout the instrument: the The system as the sole determinant of a decision without human review and its irreversibility. the damage in case of error. The concentration of points on a few high-weighted questions reflects the logic of the consequence of the decision developed by the FDA (2025). When the model operates with high human influence and low human oversight, the severity of the potential adverse event This should be the dominant factor in risk assessment. The maximum score for this block, by itself, You can move a Level II protocol to Level III. This move expresses the The seriousness of the situations described in these questions.

Block 5 assesses the impact on the participant. The maximum score is 62. points. Each question is worth five or six points. The questions with the greatest direct impact are worth more. six points: influence on clinical decisions, risk of physical harm, risk of psychological harm or social engagement and involvement of vulnerable populations. The bloc is based on Law No. 14,874, of May 28, 2024, and in Decree No. 12,651 of 2025, which establish the ethical principles. fundamentals of research with human beings in Brazil (Brazil, 2024a; 2025a). The block includes The question concerns the inference of sensitive features not explicitly collected. The MRCT Center and WCG (2025) documented that large language models are able to infer Sensitive identities and attributes can be identified from seemingly neutral data. This risk does not... disappears with formal anonymization. The block also includes a question about the mechanism of Exclusion of data from model training. Guidelines of the Indian Council of Medical Research institutions establish the participant's right to request the deletion of their data from...

repository used by AI at any time (ICMR, 2023). None of the Western models

The analyses provided for this right. This guide incorporates it as an ethical requirement.

Block 6 assesses data sensitivity and governance. The maximum score is...

77 points. Each question is worth between five and seven points. Questions directly linked

The General Law on the Protection of Personal Data is worth seven points each. The block expresses a

The central choice of the guide. The LGPD is the most robust protective feature of the Brazilian framework in
in relation to all international models analyzed.

The law classifies health data as sensitive personal data and requires a legal basis.

specific, stated purpose and identification of the Data Protection Officer for its processing.

This level of protection surpasses that of the Chinese Personal Information Protection Act (PIPL).

of the Japanese Personal Information Protection Act (APPI) and, in particular, the Act of

Indian Digital Personal Data Protection Act of 2023

which completely removed the category of sensitive data from its wording and equated

health records and any other personal data (Burman, 2023).

The question about transferring data to servers outside of Brazil is worth seven.

points. Depending on the receiving country and the contractual guarantees established, this transfer

This could completely compromise the protection afforded by the LGPD to participants in

research. The question about the risk of re-identification incorporates the documented concern.

with the ability of AI systems to infer identities from seemingly unknown data

anonymized (MRCT Center; WCG, 2025).

Subsection Block 6.b of Version B brings together five specific points for protocols that

They use databases, with weights calibrated by operational centrality in the Resolution.

CNS No. 738/2024 (Brazil, 2024b): nominal identification of the bank's Controller — P6.b.1

(seven points); presentation of the Institutional Consent Form when required by the originating entity.

from the bank — P6.b.2 (seven points, ELIMINATORY); presentation of the Commitment Agreement

Data Usage — P6.b.3 (five points); reasoned framing of any

exemption from informed consent in the applicable case according to the origin of the bank — P6.b.4 (five points); and

Formalization of joint control in multicenter research, by means of a Terms of Agreement.

Institutional — P6.b.5 (five points).

Block 6.b adds 29 points to the total score when activated by the filter in Step 0.

The absence of an affirmative response in P6.b.2, when applicable, renders the protocol unevaluable.

on the merits of the critical rule (item "The critical rule in Version B", below), regardless

of the total score. The subsection also includes three due diligence items — P6.b.4.1 (need and proportionality of access to identifiable data), P6.b.6 (cumulative criteria for waiver of informed consent in AI research) and P6.b.7 (Controller's statement) — which does not. They don't score points nor add up to the ceiling: they operate as a cumulative *gate*, and the absence of any. This criterion implies the return of the protocol.

Block 7 assesses mitigation measures. It is the only block with a bidirectional logic: questions of Risk factors add points to the total, and mitigation questions subtract. The maximum score is... 75 points. The reduction capacity through mitigation measures reaches 65 points — to which are added up to 23 points of evidence from sub-block 7C. The block operationalizes an inherited distinction from the Canadian model. Mitigation measures do not alter the risk level determined by the others. blocks, when evaluated in isolation, reduce the total score and, therefore, may to move the protocol to a lower level on the consolidated scale (Canada, 2024). This logic. It is more flexible than Version A, where mitigations do not affect the level in any way. hypothesis. It reflects the understanding that, in Version B, the scoring already incorporates the quality of the safeguards adopted by the researcher as an integral part of the risk assessment. Sub-block 7C, concerning transparency evidence, operates under a distinct, one-off regime: a Documented evidence subtracts points, while its absence adds no points.

The block is divided into three sub-blocks. Sub-block 7A (queries) is worth ten points. Assess whether the researcher sought external technical guidance before submission, including Consultation with Anvisa regarding the classification of the system as SaMD. Sub-block 7B (measures (risk reduction, *de-risking*) is worth 65 points. It assesses the quality of the mechanisms of Human supervision, explainability, lifecycle monitoring, and failure management. The sub-block includes a question about monitoring data drift, that is, the Degradation of model performance over time. This question is valid. in the life cycle maintenance concept developed by the FDA (2025), operationalized by the IDATEN system of the Japanese regulatory agency (PMDA, 2023) and by the Protocol of Indian Algorithm Change Protocol (CDSCO, 2025). Sub-block 7C (evidence of transparency) records documentary evidence — standardized documentation of the model and database, algorithmic impact assessment or Impact Report Protection of Personal Data and provision of source code — which can reduce scores by up to 23 points. when present and documented, not counting when absent.

Level ranges and score calibration

The four level ranges were calibrated based on two principles. The first is... proportionality. The Level I range is narrow, from zero to 58 points, because protocols with Truly low-risk uses of AI are rare and should be rigorously classified. The second factor is risk sensitivity. The Level IV range covers 67 points, from 209 to 275, because protocols at this level exhibit combinations of severe characteristics in multiple blocks. The width of the band reflects the variation in gravity within the critical level itself. When the filter in Step 0 activates Block 6.b (database), the matrix ceiling exceeds 275. for 304 points and the bands are recalibrated proportionally: Level I from 0 to 64 points, Level Level II: 65 to 141 points, Level III: 142 to 230 points, and Level IV: 231 to 304 points (ceiling). Evaluable score of 297 points, excluding the elimination question P6.b.2). Recalibration preserves the relative distance between levels and prevents the 29 additional points of Block 6.b from distorting the classification. The ranges will be empirically tested during the usage and testing period of the Version 1.0; any adjustments will be considered in version 2.0, to be submitted to Inaep.

The corresponding operational material is in Part II of MARIAH.

The critical rule in Version B

In Version B, the critical rule has a different formulation than in Version A. The level The score determined cannot be altered by the evaluator's subjective decision. The score is the systematic result of applying the instrument. Mitigation measures are already in place. incorporated into the scoring of Block 7. They constitute the only legitimate way to reduce the final score. This architecture preserves the integrity of the instrument. It prevents that undocumented declarations of intent replace evidence of effective control of risk. The Canadian model demonstrated that this principle is essential for the credibility of instrument for use by researchers, sponsors and society (Canada, 2024).

The gradation described so far coexists with four safeguards that are not graded. These are the special rule of Axis 3.b, the Ethical Precedence Clause, the chain of offense elimination. custody and the diligence preventing new consent, all described in the Supplement to Technical Note on Premises. They operate automatically, without depending on the judgment of... who applies the matrix. The first two determine the risk level on their own. The last two

They suspend the merit analysis and return the protocol to the researcher for further investigation.

None of them allow for compensation for declared mitigations, because they all deal with conditions that precede the risk gradation. The triggers and the procedure for each are in Part II to MARY.

The corresponding operational material is in Part II of MARIAH.

9 SPECIFICITIES BY NATURE OF THE RESEARCH

The previous chapters laid the groundwork: the concepts, the context. Brazilian, the international panorama, scientific integrity, data governance and the Risk Assessment Matrix in Artificial Intelligence. This chapter applies these fundamentals. to two territories with specific risk profiles that require questions that the instruments General guidelines do not cover this. Health research with systems classified as devices Software-based medical practice involves dual regulation, by the Ethics Committee (CEP) and the National Health Surveillance Agency (Anvisa), with mandates. complementary aspects that the protocol must satisfy in a distinct way. Research in Sciences Humanities and Social Sciences with behavioral analysis algorithms, data mining and Predictive models produce political, social, and epistemic risks that the biomedical framework does not. It reaches. Identifying the territory before applying the instrument is what allows for evaluation with the The precision that the field demands.

9.1 HEALTH RESEARCH: ASSISTANCE IMPACT, ANVISA AND SAMD INTERFACE

9.1.1 When the AI system is also a medical device

Resolution of the Board of Directors No. 657, of March 24, 2022, of the Agency The National Health Surveillance Agency (Anvisa) establishes the criteria that must be known beforehand. from any other analysis of a protocol with an AI system in healthcare: the purpose of The system determines its regulatory classification, not its technical architecture (Anvisa, 2022).

A deep learning algorithm that analyzes tomography images to Identifying pulmonary nodules is a computer program used as a medical device. same algorithm, with the same architecture, trained to optimize shift schedules Hospital-related, right? The distinction isn't in the code. It's in what the system does with the data. clinical aspects and the consequences of this operation on decisions that affect patients.

Brazilian Regulatory Decree No. 657/2022 classifies computer programs as medical devices. for the purpose of diagnosis, prevention, monitoring, prediction, prognosis, treatment or compensation for illness, injury or disability. The following are expressly excluded. programs with purely administrative, billing, or communication purposes. general or management support unrelated to individual clinical decisions. This distinction produces

an operational question that the protocol must answer before merit evaluation: the

Does the AI system described in this protocol qualify as a SaMD according to RDC No. 657/2022?

The researcher must explicitly state the classification and justify it. The absence
this statement in protocol with diagnostic, predictive or support system

Monitoring is information that prevents a complete ethical assessment of the risk posed to...
participants.

Identification does not depend solely on the researcher's statement. Two
The indicators in the protocol allow for independent reading.

The first is a description of the system's output: if the system produces classifications,
Risk scores, therapeutic recommendations, or clinical alerts that reach the professional.
Regarding health or patient care, classification as SaMD is likely.

The second is the description of the usage flow: whether the system output integrates, influences, or
By replacing steps in a care process, the classification is practically certain. The CEP
who identifies these indicators and cannot find a corresponding statement in the protocol should
Request clarification before any further analysis.

9.1.2 The CEP-Anvisa sequence: complementary, non-concurrent mandates

A system classified as SaMD in research simultaneously occupies two
Regulatory territories with distinct mandates. The Ethics Committee evaluates whether the research it develops or
The system tests whether it respects the rights and safety of participants. Anvisa evaluates whether the...
The product resulting from this research is safe and effective for widespread clinical use. The two
The mandates are complementary and sequential, not concurrent. For an AI system...
To obtain commercial registration from Anvisa, the developer needs to present evidence of
accuracy and clinical safety produced in research with human subjects. This research, before
To be conducted, it needs approval from the Ethics Committee, whose mandate is to protect the participants.
During the evidence generation phase, Anvisa certifies that the evidence is sufficient for widespread use.
Neither instance replaces the other (Anvisa, 2022; Barucci *et al.*, 2026).

The stage the system is in at the time of the research determines what it needs.
to be evaluated with greater attention. Barucci *et al.* (2026) and the MRCT Center and the WCG (2025)
They propose a three-phase taxonomy with distinct regulatory and ethical obligations: training,
Testing under real-world conditions and post-entry monitoring in use.

During the training phase, the system is under development and clinical data from Participants contribute to the learning of the model. The impact on the care of The involvement of participants is indirect, but the risk of exposure of sensitive data is at its highest. At this stage, the The protocol should detail data governance and the demographic composition of the training data. and the guarantees of consent.

In the real-world testing phase, the already trained system is evaluated against data from participants in a real clinical setting, yet without direct interference in decisions. healthcare. Data risk and the risk of automated decision-making coexist.

In the post-entry-into-use monitoring phase, the system is already operating clinically and is monitored regarding the maintenance of its performance. The risk of data drift and of Silent degradation of accuracy is the most relevant factor at this stage.

For the protocol, this taxonomy has a direct consequence. The researcher must state which phase the system is in at the time of the research and describe the specific ethical obligations of that phase.

A protocol that describes a system as a "diagnostic support tool in "validation phase" without specifying whether this corresponds to the training phase, the testing phase in Real-world conditions or post-market monitoring provide insufficient information. to assess the risk imposed on participants.

9.1.3 What the SaMD protocol should contain — and what the CEP verifies

The protocol with SaMD must contain two statements: the regulatory classification and the Phase declaration; it's the first point that the CEP verifies.

The first is the regulatory classification statement: the researcher states explicitly state whether or not the system qualifies as a SaMD according to RDC No. 657/2022. with justification based on the purpose of the system and not just on its technical architecture.

The second is the phase statement: the researcher identifies which of the three phases of the process the researcher is in. The system's life cycle, the stage it is at during the research phase, and its ethical obligations are described. specific to that phase.

If any of these statements are missing in a protocol that uses systems In terms of diagnostic, predictive, or monitoring support, it is not possible to correct this deficiency simply by

with a generic complement. This is because such information is fundamental for to define which ethical issues should be considered in the protocol.

The second requirement is a description of the potential healthcare impact on the Participants during the research phase. A system in the training phase that processes data. clinical history without interfering in decisions about active research participants has A risk profile focused on data governance and consent. But a system in the testing phase that already guides or influences clinical practices on active participants has risk profile that additionally includes automation bias, actual human supervision and possibility of harm resulting from algorithmic error.

The protocol should accurately describe the system's usage flow during the process. Research: who accesses the system's exit, at what time, and with what level of influence. regarding the clinical decision and what appeal mechanism is available to the researcher. Without Given this description, the impact on participant care remains uncertain.

The third requirement is the transition plan between ethical approval of the research and the eventual registration with Anvisa. When the research ends, the system remains. The data from Participants who contributed to the training or validation of the model remain included. within its parameters.

The model can then proceed to the registration process with Anvisa, or be transferred to a partner. commercial, published in an open repository, or discontinued. Each of these destinations has ethical implications for the participants who consented to the research, not for the outcome. posterior of the system.

Closure also creates a duty to provide feedback. The protocol should stipulate how. Participants will be informed about the product developed and any results. with an impact on your health condition or your health advice. The protocol It must explicitly describe what will happen to the system after the termination of the research indicates which participant data will remain incorporated into the model that will follow. for regulatory registration, and whether the consent obtained covers this subsequent use or if new Consent will be required. Without this description, a destination for the is implicitly approved. A system that has not been evaluated.

Studies that clinically evaluate a decision support system rely on international reference for reporting. The DECIDE-AI guideline, published simultaneously by Nature Medicine and the BMJ in 2022, organizes what is a phase 1 clinical evaluation study.

The initial report should describe the system, the data, the results, and the observed failures (Vasey *et al.*, 2022). A protocol in the initial clinical evaluation phase that organizes its plan according to this reference reaches the committee with the completeness that verification requires, and the subsequent report of the results achieve a standard recognized by the scientific community. The adoption of the guideline is voluntary and does not replace the declarations required in this section. It provides them with a format internationally established.

When the SaMD protocol uses databases — either its own or from another research project — it utilizes databases. or from an external institution, a more common situation in trained diagnostic support systems. Regarding hospital records, the operational requirements of CNS Resolution No. [number] apply. 738/2024 (Brazil, 2024b): identification of the Controller, presentation of the Terms institutional requirements according to the bank's origin, and eligibility for any exemption. of the Informed Consent Form in the applicable case according to the origin of the bank (art. 20, § 5, or art. 25).

Chapter 7 develops these figures, and Axis 3 of MARIAH (Chapter 8) translates them into operational questions. For protocols with SaMD, these requirements constitute a second verification axis, parallel to the regulatory filter of RDC No. 657/2022, and they do not replace each other.

They do not cancel each other out, nor do they cancel each other out.

9.2 RESEARCH IN HUMAN AND SOCIAL SCIENCES (HSS)

9.2.1 A territory with its own risks

Research in the Humanities and Social Sciences (HSS) using artificial intelligence (AI) does not produce clinical diagnoses nor does it guide therapeutic conduct. It produces something different and equally consequential: inferences about behavior, identity, beliefs, condition social and mental health of individuals and groups.

A sentiment analysis algorithm applied to public social media posts. It can infer emotional states that the participant never declared. A predictive model. Those trained on information consumption patterns can infer political affiliation, status socioeconomic or psychological vulnerability based on data that the participant considers trivial. A pattern recognition system for urban mobility data can reveal routines, relationships, and living conditions that the participant never authorized to be made public. visible information. AI gathers information from what people say and also infers from what is not said.

The risks that this capacity produces are not clinical. They are political, social and epistemic.

The political risk is that of surveillance: AI systems in CHS could be used for to monitor dissent, identify activists, or map social organization networks with consequences that go beyond the academic sphere.

Social risk, on the other hand, is that of discrimination and stigmatization: inferences about groups. Vulnerabilities identified in research can migrate to application contexts where They provide the basis for decisions regarding access to services, credit, employment, or housing.

Epistemic risk lies in the production of knowledge that justifies exclusion: when A model trained on historical inequality data confirms patterns of Inequality is treated as if it were a natural regularity; research transforms injustice. structural evidence in scientific terms (WHO, 2021; Soares; Chiavegatto Filho, 2026).

This change in scale, which goes from potentially individual harm to the effect on communities and social categories require that ethical evaluation also change its focus, from lens and instrument.

Evaluate CHS protocols with AI using only instruments designed for Biomedical research employs tools that are not appropriate for this context. specific. For example, the question about impact on clinical research does not have... direct correspondent in CHS.

A question about human oversight in diagnostic decisions does not translate. automatically for contexts where the decision is about publishing a discourse analysis. or train a model on data from a quilombola community. Chapter 9 proposes that The committee should learn to formulate these principles with the right questions for this territory. specific.

Resolution No. 510, of April 7, 2016, of the National Health Council (Brazil, (2016), was built for a world of questionnaires, interviews, focus groups and participant observation.

Their categories were formulated with qualitative research and ethnography in mind. The participant was someone the researcher encountered, approached, and about whom they took... conscious inclusion decisions. The data was something the participant produced through interaction. with the researcher, or that existed in institutional records with regulated access.

In that world, the distinction between public and private data was relatively stable, and the category of research that uses publicly available information exempt from Submission to SINEP made precise operational sense.

AI has dissolved that stability. Publicly available data on platforms. digital data, when processed by machine learning algorithms, produce inferences that are not public and that the participant never authorized.

For example: a social media post is public. The inference about the state of The mental health of its author, as determined by a model trained on millions of posts. Similar, isn't it? A comment in an open forum is public. The behavioral profile constructed by algorithmically cross-referencing this comment with hundreds of other traces. It is not the same user's digital fingerprints.

Resolution No. 510/2016 treated this data as minimal risk because it did not anticipated what AI does with them (Brazil, 2016). Applying the consent waiver of The solution to protocols using data mining algorithms on social networks is to apply the correct norm for the wrong context.

The resolution remains valid as a subsidiary reference to Law No. 14.874/2024 and its amendments. The principles remain relevant. What needs reinterpretation are their categories. operational in light of what AI does with data that the standard treated as minimal risk.

Three categories require this reinterpretation. The waiver of consent for research using public data: when will this data be processed by algorithms that They produce sensitive inferences, so dismissal is not automatically justified. The classification of minimal risk: research using social media data that appears to pose minimal risk. individually they can produce a significant collective risk for the groups represented in data. And the definition of participant: when an algorithm processes data from millions of People with no direct interaction, who is a participant and what are their rights?

These questions do not have a consolidated answer in the Brazilian legal system. Using them with precision contributes to building the ethical jurisprudence that the field still lacks.

Research in Human Sciences with AI presents a risk profile that is organized into three dimensions to be considered before evaluating the scientific merit of the protocol: the nature of inferences, scale, and collective risk.

The three dimensions are distinct from those that structure the evaluation of protocols. biomedical and require specific questions that conventional assessment instruments do not. Ethics are not taken into account.

In the first dimension, the nature of the inferences generated by the system stands out: the Machine learning algorithms used in CHS tend to produce inferences about attributes considered sensitive data under Brazilian law, such as mental health, sexual orientation, political affiliation, socioeconomic status, as well as ethnic origin or racial. These inferences are produced from data that the participant never associated with. These attributes, in isolation, would not reveal any sensitive information.

The algorithmic cross-referencing of seemingly trivial digital traces can produce profiles that violate privacy more profoundly than any direct questionnaire. on the same topics.

The General Data Protection Law (LGPD) classifies these attributes as data. sensitive and requires a specific legal basis for its treatment (Brazil, 2018). The protocol should Identify the sensitive inferences that the system will produce and the legal basis that authorizes this. treatment.

The second dimension is scale. Research in CHS with AI often operates about data from thousands or millions of people, most of whom never knew that participated in a research study and will never have the opportunity to exercise any rights regarding The data that the system processed.

This scale creates an ethical asymmetry that conventional instruments of protection... The number of participants cannot be reached: individual consent is operationally impossible, the The identification of participants is incomplete and the monitoring of damages is structurally flawed. limited. The evaluation of these protocols cannot demand the impossible, but the protocol must demonstrate that the safeguards adopted are proportionate to the scale of the risk produced. (MRCT Center; WCG, 2025).

The third dimension is collective risk. The inferences produced by AI systems in CHS often affects not only the individuals whose data was processed, but the social groups that these individuals represent in the data.

A model trained on data from a marginalized community can produce inferences that stigmatize all residents of that community, including those who do not participants in the research. A discourse analysis algorithm trained on messages of

A religious group can produce representations that affect everyone's public perception.
the members of that group.

Collective harm in CHS with AI occurs without the need to identify individuals.

The protocol should identify the affected groups, assess the risk of collective stigmatization, and...

To present safeguards for this risk, which goes beyond protecting the participants.

Individuals.

9.2.2 Consent in CHS with AI: what changes

Resolution No. 510/2016 waived the requirement for submission to SINEP, established by Law No.
Law 14.874/2024, for research that used publicly available information.

This exemption rested on an implicit assumption: that the use of public data
In research, no risks are introduced beyond those that the participant has already accepted by making them public.
Whoever publishes a post on social media, grants an interview to a media outlet.
communication or recording of a demonstration in a public space accepts that this information
be read, quoted and analyzed.

Conventional qualitative research operated within this assumption with
Reasonableness: the researcher reads what has been published, interprets what has been said, and produces an analysis.
which remains on record as to what the participant made accessible.

AI breaks this assumption structurally. The algorithm doesn't read what was...
Published. It cross-references, combines, and extrapolates what has been published with thousands of other records.
to produce inferences that the participant never made public and that he frequently
He is ignorant about himself.

Consent in CHS with AI cannot be resolved by data category.
origin. It needs to be resolved by the category of inference produced. A given origin.
public that feeds an algorithm capable of inferring psychiatric diagnosis, sexual orientation
economic vulnerability is not treated in an ethically equivalent way to a given fact.
public analyzed by a human researcher.

The inference produced by the algorithm is sensitive data according to the LGPD (Brazilian General Data Protection Law).
regardless of whether the source data is public (Brazil, 2018). Dispense with
consent by Resolution No. 510/2016 without considering the possible inferences of
The system applies the standard only to the original data and disregards the data that was actually generated.

by the research (Brazil, 2016). The collection is approved and the production of is implicitly accepted.

Sensitive information about people who never consented to this production.

Three concrete situations illustrate the problem of consent in CHS with AI and the questions that the protocol must answer in each of them.

The first is research that uses publicly available social media data to Training a mental health analysis model.

The source data is public. The posts were made voluntarily by their users. authors on open platforms. Resolution No. 510/2016 would classify this research as candidate for exemption from submission to SINEP (Brazil, 2016).

The algorithm, however, will produce inferences about psychiatric state and suicide risk. or substance use based on these posts. These inferences are health data and data Sensitive data according to the LGPD (Brazilian General Data Protection Law). The waiver of consent does not apply automatically. because the inference produced belongs to a different category than the source data. The postal code should The question is: Does the protocol identify the sensitive inferences that the model will produce? What is the basis? What legal authorization is required to process this sensitive data? The authors of the posts will be... Have they been informed that their data was used to produce this type of inference?

The second is research that uses recordings of interviews obtained with consent. preliminary training for a speech pattern recognition model.

Original consent was obtained for conducting the interview and for the use of your information. content in qualitative analysis. It did not anticipate the use of the recording to train an algorithm. Machine learning, which will process not only the semantic content of speech, but also its... Prosodic patterns, pauses, variations in tone, and vocal characteristics. These patterns can... to reveal an emotional state, health condition, or regional origin that the participant does not intended to disclose. The CEP should ask: does the original consent cover the use of the recording? For algorithmic training? If not, has new consent been obtained or will it be obtained? The protocol describes which attributes the model infers from the vocal characteristics of the Participants?

The third is longitudinal research, which reuses data collected in previous studies. previous data was collected to train a predictive model about behavior. with consent obtained for a specific and distinct purpose. Reuse for AI model training constitutes secondary use which, according to Law No. 14.874/2024, regulated by Decree No. 12,651/2025, requires specific ethical evaluation (Brazil, 2024a;

2025a). The CEP should ask: did the original consent authorize secondary use for Development of AI systems? If it wasn't authorized, the protocol demonstrates that new consent was obtained or its waiver is adequately justified based on the circumstances. Are the exemptions permitted by the CNS resolutions, given that Law No. 14.874/2024 does not provide for them? A model trained on this data will produce inferences about the original participants or Are these just general population patterns?

Informed consent in CHS with AI is consent about what the system... it will conclude, not only about what the researcher will collect. This distinction reorganizes what The protocol and informed consent form should be required in research in this field.

Therefore, three requirements structure this reorganization:

The protocol should describe not only the origin of the data, but also the inferences that... The system will produce and explicitly assess whether consent exists or whether a waiver is granted. The expected data covers these inferences. A protocol that accurately describes the data that will be... collected, but remains silent on what the algorithm will infer from this data. Incomplete information about the actual risk of the research. The relevant question is not only the Origin of the data. This refers to what the system will produce with that data and who will be affected by it. production.

When the inference produced by the system is about a sensitive attribute, the Consent must be obtained even if the source data is public. Advertising The source data does not transfer to the participant the responsibility for inferences that he/she makes. did not anticipate and did not authorize. Accept the waiver of consent based on advertising. The algorithm applies the criterion based on the data without verifying the nature of the inferences produced. Correct to the wrong element of the research. The legal basis for the processing of sensitive data. requires specific justification that the protocol must present and the committee must evaluate (Brazil, 2018).

Informed consent in CHS with AI should inform participants about the type of inference that will be used. produced, not just about the type of data that will be collected. A term that describes the collecting public posts without mentioning that these posts will be used to infer "State of mental health" is not a complete term. A term that describes the recording of interviews without mentioning that the vocal patterns will be processed algorithmically for Identifying characteristics that the participant did not declare is not a transparent term.

Law No. 14,874/2024 requires that consent cover research procedures.
(Brazil, 2024a). Algorithmic data processing for the production of sensitive inferences.
It is a research procedure with consequences that the participant has the right to know.
before deciding whether to consent.

9.2.3 Collective risk and vulnerable groups in CHS

The risk in biomedical research has a precise individual address: the participant is
exposed to an intervention that could cause harm to their health, and the ethical protection system.
It was built to protect that specific person who consented, who can withdraw the consent.
consent and the right to be monitored during and after the research.

The risk of research in the humanities and social sciences (HSS) that mobilizes AI operates on a different scale. A model
Training on data from residents of a peripheral community can produce inferences.
which stigmatize all residents of that community, including those who did not participate.
from the research, they don't know it was conducted and will never have the opportunity to contest it.
The algorithm's conclusions about itself.

A discourse analysis system trained on messages from a social movement.
It can produce representations that affect the public perception of all its members.
including those who never posted any messages. The collective damage in CHS with AI does not
It requires individual identification to occur. It occurs at the group, community, or national level.
social category that the data represents, and its effects are distributed among people who...
The research protocol never named it.

This characteristic creates a structural gap in conventional instruments of
Ethical protection in research. SINEP was designed to protect participants: people who
They have consented, that they can be identified, that they have exercisable rights, and that the researcher
They have an obligation to follow up.

The collective participant in a research study in CHS with AI does not have these characteristics.
It is a social group whose members may never know they have been affected, they have no idea.
collective consent mechanism recognized by the Brazilian legal system and are not
covered by the protection instruments provided for in Law No. 14,874/2024 with the same
clarity that individual participants (Brazil, 2024a).

Evaluating research in CHS with AI therefore faces a challenge that goes beyond Regarding the application of existing instruments: it requires formulating questions about collective risk for which the system has not yet built consolidated responses, and it is necessary to do so based on principles that research ethics has accumulated, applied to a territory where these principles They did not explicitly anticipate this.

Populations that already suffer from structural discrimination are the most exposed to risk. Collective research with AI in CHS. The reason is twofold and reinforces itself.

First, these populations are systematically underrepresented in the data that They train the models: the databases available for research tend to reflect the populations that had greater access to services, institutional records, and platforms digital, which are not the populations in the most vulnerable situations.

Second, the social categories of these populations are systematically super-represented in the historical data of exclusion, criminalization, and stigma that the models They learn through patterns.

The algorithm, trained on data that carries centuries of injustice, learns this. Injustice as a regularity. It reproduces as a scientific finding what is a historical product of discrimination (Soares; Chiavegatto Filho, 2026; Fiske *et al.*, 2025).

For Black and Brown, Indigenous, Quilombola, LGBTQIA+ populations, deprived of liberty, residents of the outskirts of cities, and people in vulnerable situations. Economically, this mechanism has consequences that extend beyond the academic sphere. Inferences produced by algorithms about these groups can support decisions about access to public services, eligibility for social benefits, priority in service of health or crime risk attributed to entire territories.

When research in CHS with AI produces models that will eventually be applied in public policy contexts, approving it without assessing the collective risk regarding It is also important to approve the chain of consequences that the model will carry for vulnerable groups. Beyond the academic environment. The principle of non-maleficence, as outlined in Resolution No. 466/2012. (Brazil, 2013) established as a foundation of ethics in research and that Law No. 14.874/2024 It preserves, and is not limited to, the protection of the individual participant during the study. It reaches the groups that the data represent and the contexts in which the inferences will be used (Brazil, 2012; Brazil, 2024a).

9.2.4 What the AI-powered CHS protocol should contain — and what the CEP verifies

The previous sections covered the complete diagnosis of research in CHS with AI: The risks are not clinical, but they are real; Resolution No. 510/2016 needs reinterpretation. In light of what algorithms do with data that the standard treated as low-risk; the Consent is organized by the inference produced, not by the source data; and the risk Collective action affects groups that the protocol never named (Brazil, 2016).

This section provides the operational tool that transforms this diagnosis into... assessment. Four blocks of questions structure the reading of any CHS protocol with IA.

The first section is about the origin and nature of the data. Where does the data come from that... The system will process: prospective data collection with consent, institutional data with access. regulated, publicly available data on digital platforms, or data reuse. collected in previous research? Each source has distinct implications for the consent and the legal basis for processing. The protocol accurately describes the origin of each type of data used and the legal basis that authorizes its processing by AI system?

The second section is about the inferences produced. The protocol describes what the What inferences will the system make from the collected data? The inferences produced involve attributes. Sensitive data under the LGPD (Brazilian General Data Protection Law), such as mental health, sexual orientation, political affiliation, and condition. Socioeconomic or ethnic origin? If so, what is the legal basis that authorizes the treatment of these individuals? Does the consent obtained or the waiver provided specifically cover sensitive data? Does the consent obtained or the waiver provided specifically cover these? inferences, and not just the source data? The informed consent form informs participants about the type What type of inference will the system produce in accessible and precise language?

The third section is about collective risk. The protocol identifies the social groups that groups will be affected by the system's inferences, including those not represented. directly in the data but sharing the social categories of the participants? The protocol assesses the risk of collective stigmatization and describes specific safeguards for it. How does this risk differ from safeguards for individual protection? The researcher demonstrates that considered the potential use of the inferences produced in contexts that amplify Structural discrimination against vulnerable groups?

The fourth section is about the model's fate. The protocol describes what will happen to the model after the research is completed? When the intended destination includes publication in Whether used as an open repository or for transfer to partners, the protocol assesses the risk of using the model. in discriminatory contexts that the participants did not anticipate and did not authorize? The Informed Consent Form (ICF) honestly informs participants about the model's intended use. What are the possible secondary effects and collective risks that these uses may produce?

The distinction between an incomplete protocol and an inadequate protocol applies in CHS with AI with the same logic as section 6.5, adjusted to the specific territory. The incomplete protocol. It does not describe the inferences that the system will produce, nor does it assess whether the dismissal of Does the consent provided for in Resolution No. 510/2016 cover these inferences or does it not identify the groups? affected beyond formal participants (Brazil, 2016).

These gaps can be filled by requesting additional information with specific criteria. explicit. The inadequate protocol presents public data processed by an algorithm that produces sensitive inferences such as minimal-risk research that does not require consent, without any analysis of the inferences produced or the legal basis for their treatment. This This case requires a reasoned refusal based on the principles of Law No. 14.874/2024 and the requirements of... LGPD, with arguments demonstrating to the researcher why the risk classification is necessary. The minimum does not hold up in light of what the system actually produces (Brazil, 2024a; Brazil, 2018).

This chapter concludes with a statement that runs throughout the entire guide: Specificity is not the exception. It's the rule. Every AI protocol has a specific nature that... determines the set of relevant questions. Biomedical research with a system classified as SaMD and research in Human and Social Sciences with algorithms of Behavioral analysis is not a matter of variations on the same ethical problem with different intensities. different. They are distinct territories that require distinct instruments, distinct questions and Different evaluation criteria. Identifying the territory before applying the instrument does not create Obstacles to AI research: Is the evaluation mandate exercised with the precision that... The complexity of the field demands it, and the people participating in this research have the right to... to be expected.

10. RELATIONSHIP WITH THE GUIDE AND MARIAH

The documentary collection now consists of three complementary pieces:

- Guide to the Ethical Use of Artificial Intelligence in Research with Human Subjects (Guide INAEP) — institutional guidance document for ethical evaluation, with recommendatory language and focus on the criteria that should be considered by researchers and ethics committees.
- MARIAH — matrix, checklists and application tools for identification, classification, verification and documentation of risks.
- Reference Notebook — a foundational document that preserves the construction theoretical, normative, methodological and comparative approach developed by the Group of Work.

In case of differences in wording between this Reference Booklet and the INAEP Guide, the Current institutional guidelines should be consulted in the Guide and applicable regulations.

The Reference Notebook explains the rationale and records the technical trajectory that underpins these guidelines.

11 INTERNATIONAL REGULATORY UPDATE — 2024–2026

The rapid evolution of international regulatory frameworks demands that the rationale...
The ethics of research with artificial intelligence should be interpreted as a dynamic framework.
Between 2024 and 2026, the European Union and drug regulatory authorities
They deepened the connection between AI regulation, drug development,
Data protection and secondary use of health data. The developments below.
These complement the international overview presented in this Notebook and do not represent a transposition.
Automatic foreign requirements for the National Ethics System in Research with Beings
Humans.

11.1 European Artificial Intelligence Regulation (AI ACT)

Regulation (EU) 2024/1689 (European Union, 2024) established an architecture
Risk-based research for artificial intelligence systems. For the research, your contribution
The methodological aspect lies in requiring that the framework consider the intended purpose and the context.
of use and the system's influence on decisions and rights. The mere presence of AI in a
The protocol alone does not determine that the system is high-risk. The framework
It depends on the assumptions and criteria set out in the Regulation.

Implication for ethical evaluation: it is recommended to start from the context of use, from
potential consequence of the decision, the degree of autonomy, human supervision and of
potential effects on participants and groups, avoiding automatic classifications based
only in terms of the type of technology.

11.2 EMA AND THE MEDICINE LIFE CYCLE

In September 2024, the European Medicines Agency (EMA) adopted the *Reflection paper*.
on the use of Artificial Intelligence (AI) in the medicinal product lifecycle
(EMA/CHMP/CVMP/83833/2023). The document considers applications of AI and machine learning.
The machine is used throughout the drug's lifecycle and guides the user to understand how to use it.
in conjunction with European requirements on medicines, artificial intelligence and
data protection.

For this Notebook, the document reinforces a life cycle approach: provenance and suitability of the data, performance in the intended context, documentation, Human supervision, change control, and monitoring remain relevant to as the system and the usage environment evolve.

11.3 Joint EMA–FDA Principles of Good AI Practices in Development OF MEDICINES

In January 2026, the EMA and the US Food and Drug Administration (FDA) published ten guiding principles for good AI practices in development are collectively outlined. medications. The document covers AI used to generate or analyze evidence in the phases non-clinical, clinical, post-marketing and manufacturing and establishes a common basis for the Development of future best practices.

Convergence reinforces elements incorporated into MARIAH: a focus-centered approach. human being, clearly defined context of use, governance and data quality, performance appropriate to the purpose, assessment proportionate to the risk, transparency, documentation, lifecycle management and monitoring.

11.4 European Health Data Area (EHDS)

Regulation (EU) 2025/327 establishing the European Health Data Space (EHDS), It adds a relevant layer to health data governance in Europe. Among its... The objective is to establish harmonized conditions for the secondary use of certain Categories of electronic health data, including for research, innovation and activities. regulatory, subject to the safeguards provided for in the Regulation.

The EHDS experience demonstrates that data governance for research is not limited to... to obtaining consent. It involves purpose, legal basis, controlled access, minimization, security, institutional responsibilities, traceability and conditions for Reuse. This is a comparative reference, not a standard directly applicable to Brazil.

11.5 EDPB AND SCIENTIFIC RESEARCH — GUIDELINES 1/2026

In April 2026, the European Data Protection Board (EDPB) submitted it for consultation. pública as *Guidelines 1/2026 on processing of personal data for scientific research purposes*. A The consultation ended on June 25, 2026. The document addresses legal bases and roles. controller and operator, rights of data subjects and principles applicable to data processing. Personal data for scientific research.

Since this document has been submitted for consultation, its use in this Notebook is... It serves as a reference point and preserves that *status*. The material is especially relevant for distinguishing consent for participation in research on the legal basis for treatment of Personal data, related issues, but not necessarily equivalent.

11.6 EDPB AND ANONYMIZATION — GUIDELINES 02/2026

In July 2026, the EDPB published *Guidelines 02/2026 for consultation*. *Anonymization*. As of the date of this update, the consultation remains open until October 30th. 2026. The document, therefore, is not presented here as European guidance. definitively consolidated.

Its relevance to AI research stems from the need to distinguish anonymization of pseudonymization and assessing the risk of re-identification in the context of treatment. In systems trained with high-dimensional data, the simple statement The fact that a dataset is 'anonymous' does not replace the technical analysis of the risk of re-identification.

11.7 British Recommendations for the Regulation of AI in Healthcare

In September 2026, the National Commission into the Regulation of AI in Healthcare published recommendations for a future British regulatory framework, in an initiative supported pelo Department of Health and Social Care e pela Medicines and Healthcare products Regulatory Agency (MHRA) (United Kingdom, 2026). This document constitutes a comparative reference. and it does not produce normative effects in Brazil.

Your contribution to ethical review lies in the articulation between purpose, context, Use, performance, equity, and responsibilities throughout the lifecycle. The evaluation does not... It ends with the initial validation: updates, recalibrations, retraining, replacement of

The model, changes in the data, or expansion of the user population can alter the profile of risk and require a new assessment.

According to the INAEF Guide and MARIAH, the recommendations reinforce the need for: Identify the system version and configuration; justify the suitability of the evidence to the context. Brazilian; evaluate performance and potential differences between subgroups; define supervision. Effective human resource management; document change control; establish monitoring criteria. interruption and incident response; and clearly distribute responsibilities among researchers, institutions, developers, suppliers and other stakeholders.

Equity should also be treated as a dimension of security. A system can to present satisfactory aggregate performance and yet produce inferior results. or disproportionate harm to underrepresented groups. For the Humanities and Social Sciences, This perspective broadens the analysis to include inferences about communities, stigmatization, discrimination, power relations, and affected individuals who do not formally fit into the system. as participants.

11.8 CONSEQUENCES FOR THE BRAZILIAN REFERENCE FRAMEWORK AND FOR MARIAH

Developments in 2024–2026 reinforce five conclusions: (i) risk must be assessed based on the context of use and potential consequences; (ii) data governance and algorithmic governance should be examined in an integrated manner; (iii) validation, supervision and monitoring accompany the life cycle; (iv) secondary use of data requires own analysis of purpose, legal basis, access, security and rights; and (v) International convergence does not eliminate the need for contextualization of norms. Brazilian, to population and institutional diversity and to the characteristics of the Health Sciences. and the Humanities and Social Sciences.

These elements complement, without replacing, the fundamentals of the INAEF Guide and the MARIAH. Foreign benchmarks are comparative sources of regulatory learning; a Ethical evaluation in Brazil remains guided by the applicable national legal and ethical framework.

REFERENCES

- 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](https://doi.org/10.1109/PGEC.1967.264666). Disponível em: <https://doi.org/10.1109/PGEC.1967.264666>. Acesso em: 24 jul. 2026.
- NATIONAL HEALTH SURVEILLANCE AGENCY. Collegiate Board Resolution - RDC No. 657, of March 24, 2022. Provides for the regulation of software as a medical device (Software as a Medical Device — SaMD). Official Gazette of the Union: section 1, Brasília, DF, edition 61, March 30, 2022. Available at: https://anvisa.gov.br/legis/datalegis.net/action/ActionDatalegis.php?acao=abrirTextoAto&tipo=RDC&numeroAto=00000657&seqAto=000&valorAno=2022&orgao=RDC/DC/Anvisa/MS&codTipo=&desltem=&desltemFim=&cod_menu=9434&cod_modulo=310&pesquisa=true. Accessed on: April 19, 2026.
- AUSTRALIA. National Health and Medical Research Council (NHMRC). Guide for assessing research involving AI. Canberra: NHMRC, 2023. Available at: <https://www.nhmrc.gov.au/about-us/resources/guide-assessing-research-involving-ai>. Accessed on: March 10, 2026.
- BRAZIL. Ministry of Justice and Public Security. National Data Protection Agency. Data Protection Impact Assessment Report (DPIA). Brasília, DF: ANPD, September 23, 2026. Available at: https://www.gov.br/anpd/pt-br/canais_atendimento/agente-de-tratamento/relatorio-de-impacto-a-protecao-de-dados-pessoais-ripd. Accessed on: September 24, 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>. Accessed on: May 24, 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. Available at: <https://doi.org/10.15420/ecr.2025.59>. Accessed on: April 19, 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. Available at: <https://doi.org/10.3390/ijerph23010081>. Accessed on: May 24, 2026.

BRAZIL. National Data Protection Authority. Generative artificial intelligence. Brasília, DF: ANPD, 2024d. (Technological Radar, 3). Available at: <https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/radar-tecnologico-ia-generativa-anpd.pdf>. Accessed on: March 23, 2026.

BRAZIL. Brazilian Institute of Geography and Statistics. Estimates of the resident population for municipalities and for the federative units with a reference date of July 1, 2026. Rio de Janeiro: IBGE, 2026e. Published in the Official Gazette of the Union on August 28, 2026. Available at: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html>. Accessed on: September 24, 2026.

BRAZIL. Ministry of Science, Technology and Innovation. AI for the good of all: Brazilian Artificial Intelligence Plan. Brasília, DF: Ministry of Science, Technology and Innovation; Center for Management and Strategic Studies, 2025c. Available at: <https://www.cgee.org.br/documents/37878/157505/000149+-+CGEE+Brazilian+Artificial+Intelligence+Plan+%28VN%29+8+Electoral+Closed+Season.pdf/0a05e153-87fb-0ac1-bfa3-af27471204dc?version=2.0&t=1784645088790>. Accessed on: September 24, 2026.

BRAZIL. Ministry of Justice and Public Security. **Brazilian user guide to artificial intelligence**. Brasília, DF: MJSP, 2026c. Available at: <https://www.gov.br/mjsp>. Accessed: Apr. 2026.

BRAZIL. Ministry of Health. National Health Council. Circular Letter No. 1, of March 3, 2021. Provides guidance to Ethics Committees on data sharing in collaborative research. Brasília, DF: CNS, 2021.

BRAZIL. Ministry of Health. National Health Council. Resolution No. 466, of December 12, 2012. Approves the following guidelines and regulatory standards for research involving human beings; Revokes (RES. 196/96); (RES. 303/00); (RES. 404/08). Official Gazette of the Union: section 1, Brasília, DF, no. 12, p. 59, June 13, 2013. Available at: <https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2012/resolucao-no-466.pdf/view>. Accessed on: June 24, 2026.

BRAZIL. Ministry of Health. National Health Council. Resolution No. 510, of April 7, 2016. This Resolution establishes the rules applicable to research in Human and Social Sciences whose methodological procedures involve the use of data directly obtained from participants or identifiable information, or that may entail risks greater than those existing in everyday life, as defined in this Resolution. **Official Gazette of the Union:** section 1, Brasília, DF, no. 98, pp. 44-46, May 24, 2016.

Available at: <https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2016/resolucao-no-510.pdf/view>. Accessed on: April 19, 2026.

BRAZIL. Ministry of Health. National Health Council. Resolution No. 738, of February 1, 2024. Provides for the use of databases for scientific research involving human beings. Brasília, DF: CNS, 2024b. Available at: <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>.

Accessed on: April 19, 2026.

BRAZIL. Ministry of Health. Ordinance GM/MS No. 3,232, of March 1, 2024. Amends Consolidation Ordinance GM/MS No. 5, of September 28, 2017, to establish the SUS Digital Program. **Official Gazette of the Union:** section 1, Brasília, DF, edition 43, p. 52, March 4, 2024c.

Available at: <https://www.in.gov.br/en/web/dou/-/portaria-gm/ms-n-3.232-de-1-de-marco-de-2024-546278935>.

Accessed on: March 23, 2026.

BRAZIL. Ministry of Health. **Ordinance No. 1,434, of May 28, 2020.** Establishes the National Health Data Network. Brasília, DF: Ministry of Health, 2020a. Available at: <https://www.gov.br/saude/pt-br/composicao/seidigi/rnds>. Accessed on: March 23, 2026.

BRAZIL. Ministry of Health. Secretariat of Information and Digital Health. Public Notice No. 1/2026 — Public Call for Proposals for the Inova SUS Digital Laboratory. Brasília, DF: Ministry of Health/Seidigi, January 7, 2026a.

BRAZIL. Ministry of Health. Executive Secretariat. Department of Informatics of the SUS (Unified Health System). Digital Health Strategy for Brazil 2020–2028. Brasília: Ministry of Health, Department of Informatics of the SUS, 2020b.

BRAZIL. Presidency of the Republic. Decree No. 11,358, of January 1, 2023. Approves the Organizational Structure and the Demonstrative Chart of Commissioned Positions and Trust Functions of the Ministry of Health and reallocates commissioned positions and trust functions.

Official Gazette of the Union, Brasília, DF, special edition. January 1, 2023b. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/decreto/d11358.htm. Accessed on: April 19, 2026.

BRAZIL. Presidency of the Republic. Decree No. 11,798, of November 28, 2023. Approves the Organizational Structure and the Demonstrative Chart of Commissioned Positions and Trust Functions of the Ministry of Health and reallocates and transforms commissioned positions and trust functions. Official Gazette of the Union: extra edition, Brasília, DF, November 28, 2023c. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/decreto/d11798.htm. Accessed

Published: September 24, 2026.

BRAZIL. Presidency of the Republic. Decree No. 12,651, of October 7, 2025. Regulates Law No. 14,874, of May 28, 2024, which provides for research with human beings and establishes the National System of Ethics in Research with Human Beings. **Official Gazette of the Union:** section 1, Brasília, DF, edition 192, p. 3, Oct. 8, 2025a. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/decreto/d12651.htm. Accessed on: June 19, 2026.

BRAZIL. Presidency of the Republic. **Law No. 13.709, of August 14, 2018.** General Law on the Protection of Personal Data (LGPD). **Official Gazette of the Union:** section 1, Brasília, DF, August 15.

2018. Available at: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Accessed on: April 19, 2026.

BRAZIL. Presidency of the Republic. Law No. 14,510, of December 27, 2022. Amends Law No. 8,080, of September 19, 1990, to authorize and regulate the practice of telehealth throughout the national territory, and Law No. 13,146, of July 6, 2015; and repeals Law No. 13,989, of April 15, 2020. Official Gazette of the Union: section 1, Brasília, DF, Dec. 28, 2022. Available

available at: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2022/lei/l14510.htm. Accessed on: April 19, 2026.

BRAZIL. Presidency of the Republic. Law No. 14,874, of May 28, 2024. Provides for research with human beings and establishes the National System of Ethics in Research with Human Beings. Official Gazette of the Union: section 1, Brasília, DF, no. 103, p. 6, May 29, 2024a.

Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/l14874.htm.
Accessed on: March 23, 2026.

BRAZIL. Presidency of the Republic. Law No. 15,211, of September 17, 2025. Provides for the protection of children and adolescents in digital environments (Digital Statute of Children and Adolescents). Official Gazette of the Union: section 1, Brasília, DF, extra edition, September 17, 2025b.

Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm.
Accessed on: June 29, 2026.

BRAZIL. Federal Senate. Bill No. 2,338 of 2023. Provides for the use of Artificial Intelligence. Brasília, DF: Federal Senate, 2023a. Available at: <https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>. Accessed on: April 12, 2026.

BURMAN, A. Understanding India's new data protection law. Washington, DC: Carnegie Endowment for International Peace, Oct. 2023. Available at: <https://carnegieendowment.org/research/2023/10/understanding-indias-new-data-protection-law>. Accessed on: Mar. 11, 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. Available at: <https://open.canada.ca/data/en/dataset/5423054a-093c-4239-85be-fa0b36ae0b2e>. Accessed on: 24 Sep. 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. Accessed on: June 23, 2026.

CHINA. National Health Commission (NHC). *Measures for the ethical review of life sciences and medical research involving human subjects*. Beijing: NHC, Feb. 2023.

Federal Council of Medicine. Resolution CFM No. 2,454, of February 11, 2026. Standardizes the use of artificial intelligence in medicine. Official Gazette of the Union: section 1, Brasília, DF, edition 39, p. 158, Feb. 27, 2026. Available at: https://sistemas.cfm.org.br/normas/arquivos/resolucoes/BR/2026/2454_2026.pdf. Access

Published: September 22, 2026.

NATIONAL COUNCIL FOR SCIENTIFIC AND TECHNOLOGICAL DEVELOPMENT. Ordinance No. 2,664, of March 6, 2026. Establishes the Policy of Integrity in Scientific Activity and establishes obligations regarding the use of generative artificial intelligence in research. Official Gazette of the Union: section 1, Brasília, DF, p. 4, March 11, 2026. Available at: http://memoria2.cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/23142775. Accessed on: March 13, 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. Available at: 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. Accessed on: June 19, 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, July 9, 2025. Available at: <https://cset.georgetown.edu/publication/south-korea-ai-law-2025/>. Accessed on: March 11, 2026.

COTA, DOC; JESUS, DCS; ROCHA, AAA. Machine Unlearning and the LGPD: from privacy to the right to be forgotten. In: BRAZILIAN SYMPOSIUM ON COMPUTER NETWORKS AND DISTRIBUTED SYSTEMS, 42nd, 2024, Niterói, RJ. Mini-courses [...]. Porto Alegre: Brazilian Computer Society, 2024. Chapter 4, p. 143-185. Available at: <https://books-sol.sbc.org.br/index.php/sbc/catalog/view/154/661/1202>. Accessed on: May 24, 2026.

DOURADO, DA; AITH, FMA. The regulation of artificial intelligence in healthcare in Brazil begins with the General Law on the Protection of Personal Data. *Revista de Saúde Pública*, São Paulo, v. 56, 80, 2022. DOI: <https://doi.org/10.11606/s1518-8787.2022056004461>. Available at: <https://doi.org/10.11606/s1518-8787.2022056004461>. Accessed on: May 24, 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. Available at: https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en. Accessed on: April 18, 2026.

FIGUEIREDO, K.; AGUIAR, RAT. Artificial Intelligence: promises and responsibilities. In: VIANNA SOBRINHO, L. *et al.* (org.). **Artificial Intelligence and bioethics: perspectives**. Boca Raton: CRC Press, 2025. Chapter 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. Available at: <https://doi.org/10.1016/j.landig.2025.01.003>. Accessed on: Apr. 19, 2026.

FORNASIER, MO. 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. Available at: https://www12.senado.leg.br/ril/edicoes/59/233/ril_v59_n233_p201. Accessed on: June 19, 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. Available at: <https://doi.org/10.1007/BF00344251>. Accessed on: July 24, 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. Generative artificial intelligence tool used for creating scientific figures. Available at: <https://deepmind.google/technologies/gemini>. Accessed: Apr. 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. Available at: <https://doi.org/10.1145/3605764.3623985>. Accessed on: May 24, 2026.

HADDAD, AE *et al.* Evolution of Brazilian telehealth regulation: from the National Telehealth Program to the Digital SUS. *Revista de Direito Sanitário*, São Paulo, v. 25, n. 1, e0008, 2025. DOI: 10.11606/issn.2316-9044.rdisan.2025.238280. Available at: <https://doi.org/10.11606/issn.2316-9044.rdisan.2025.238280>. Accessed on: May 24, 2026.

HERN, A. Royal Free breached UK data law in 1.6m patient deal with Google's DeepMind. *The Guardian*, London, 3 July 2017. Available at: <https://www.theguardian.com/technology/2017/jul/03/google-deepmind-16m-patient-royal-free-deal-data-protection-act>. Accessed on: June 19, 2026.

INDIA. Central Drugs Standard Control Organization (CDSCO). *Draft guidance on medical device software*. New Delhi: CDSCO, Oct. 2025. Available at: <https://community.nasscom.in/communities/public-policy/cdsco-issues-draft-guidance-medical-device-software>. Accessed on: Mar. 11, 2026.

INDIAN COUNCIL OF MEDICAL RESEARCH. Ethical guidelines for application of artificial intelligence in biomedical research and healthcare. New Delhi: ICMR, 2023. Available at: <https://www.icmr.gov.in/ethical-guidelines-for-application-of-artificial-intelligence-in-biomedical-research-and-healthcare>. Accessed on: March 11, 2026.

INDIA. Ministry of Law and Justice. The Digital Personal Data Protection Act, 2023. The Gazette of India Extraordinary, New Delhi, no. 22, August 11, 2023. Available at: <https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf>. Accessed on: March 11, 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. Available at: <https://www.icmje.org/recommendations/>. Accessed on: April 18, 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.

JAPAN. Pharmaceutical and Medical Devices Agency (PMDA). Report on AI-based software as a medical device (SaMD). [Tokyo]: PMDA, August 8, 2023. Available at: <https://www.pmda.go.jp/files/000266100.pdf>. Accessed March 11, 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](https://doi.org/10.1353/tech.1999.0128). Disponível em: <https://doi.org/10.1353/tech.1999.0128>. Accessed on: July 24, 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](https://doi.org/10.3390/ijerph23010012). Disponível em: <https://doi.org/10.3390/ijerph23010012>. Acesso em: 24 maio 2026.

LUNKES, ALZ; LUNKES, AFZ; OLIVEIRA, FB Application of homomorphic cryptography in the analysis of dengue data. *Caderno Pedagógico*, [S. l.], v. 21, n. 8, p. e7263, 2024. DOI: [10.54033/cadpedv21n8-278](https://doi.org/10.54033/cadpedv21n8-278). Available at: <https://ojs.studiespublicacoes.com.br/ojs/index.php/cadped/article/view/7263>. Accessed on: May 24, 2026.

MARIOTTO, C. Brazilian legal framework on automated decision-making. International Bar Association, 9 June 2024. Available at: <https://www.ibanet.org/Brazilian-legal-framework-on-automated-decision-making>. Accessed on: 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. Available at: <https://mrctcenter.org/resource/framework-for-review-of-clinical-research-involving-ai/>. Accessed on: March 23, 2026.

WORLD HEALTH ORGANIZATION. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO, 2024. Available at: <https://www.who.int/publications/i/item/9789240084759>. Accessed on: March 23, 2026.

WORLD HEALTH ORGANIZATION. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization, 2021. Available at: <https://www.who.int/publications/i/item/9789240029200>. Accessed on: March 23, 2026.

PAN AMERICAN HEALTH ORGANIZATION; WORLD HEALTH ORGANIZATION. Artificial Intelligence in Public Health: Preparedness Assessment Toolkit. Washington, DC: PAHO/WHO, 2024. Available at: <https://www.paho.org/es/documentos/inteligencia-artificial-salud-publica-kit-herramientas-evaluacion-preparacion>. Accessed on: March 23, 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/>. Accessed on: May 24, 2026.

POWLES, J.; HODSON, H. Google DeepMind and health in an age of algorithms. *Health and Technology*, v. 7, n. 4, p. 351-367, 2017. DOI: 10.1007/s12553-017-0179-1. Available at: <https://doi.org/10.1007/s12553-017-0179-1>. Accessed on: May 24, 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>. Available at: <https://doi.org/10.1038/s41746-020-00323-1>. Accessed on: May 24, 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>. Available at: <https://doi.org/10.1038/s41467-019-10933-2>. Accessed on: May 24, 2026.

RUSSIA. Federal Law No. 123-FZ, of April 24, 2020. Experimental legal framework for digital innovation in the municipality of Moscow. Moscow: Federal Duma, 2020. Available at: <https://cis-legislation.com/document.fwx?rgn=124089>. Accessed on: March 11, 2026.

RUSSIA. Federal Law No. 152-FZ, of July 27, 2006. On personal data. Moscow: Federal Duma, 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, MA; MORENO, GMM; SANTOS, WP. Explainable artificial intelligence. In: VIANNA SOBRINHO, L. *et al.* (org.). **Artificial Intelligence and bioethics: perspectives**. Boca Raton: CRC Press, 2025. Chapter 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, MQ; CHIAVEGATTO FILHO, ADP Artificial intelligence in Brazilian public health: potentialities, challenges and ethical implications for the Unified Health System. *Brazilian Journal of Epidemiology*, São Paulo, v. 29, e260006, 2026. DOI: <https://doi.org/10.1590/1980-549720260006>. Available at: <https://doi.org/10.1590/1980-549720260006>. Accessed on: April 19, 2026.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. Adopted on November 23, 2021. Paris, FR: UNESCO, 2022. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Accessed on: July 12, 2026.

EUROPEAN UNION. European Parliament; Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *Official Journal of the European Union*, L 119, p. 1–88, 4 May 2016. Available at: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

Accessed on: April 14, 2026.

EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 establishing 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 Regulation) (Text with EEA relevance). **Official Journal of the European Union**, L series, 12 July 2024. Available at: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=OJ:L_202401689. Accessed on: April 14, 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. Available at: <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>. Accessed on: September 10, 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. Available at: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological>. Accessed on: March 10, 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

Published in BMJ, v. 377, e070904, 2022. DOI: 10.1038/s41591-022-01772-9. Available at: <https://doi.org/10.1038/s41591-022-01772-9>. Accessed on: July 24, 2026.

VASWANI, A. *et al.* Attention is all you need. arXiv:1706.03762 [cs.CL], 2017. Available at: <https://arxiv.org/abs/1706.03762>. Accessed on: April 2, 2026.

WILTON PARK. Leveraging digital public infrastructure for sustainable health system transformation. [S. l.]: Wilton Park, Feb. 2026. Available at: <https://www.wiltonpark.org.uk/reports/leveraging-digital-public-infrastructure-for-sustainable-health-system-transformation/>. Accessed on: Mar. 25, 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. Available at: <https://doi.org/10.1001/jamanetworkopen.2024.32482>. Accessed on: May 24, 2026.

RECOMMENDED READINGS

ALBUQUERQUE, A.; REGO, S. Ethical Opinion No. 1/2026, CFM Resolution No. 2,454, of February 11, 2026, regulates the use of artificial intelligence in medicine. [S. l.]: Brazilian Society of Bioethics, 2026. 18 p. Available at: 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. Accessed on: September 11, 2026.

BOLLOS, RH AI in medicine and the CFM resolution: advances and gray areas. Nexô Políticas Públicas, São Paulo, March 11, 2026. Point of View. Available at: <https://pp.nexojournal.com.br/ponto-de-vista/2026/03/11/ia-na-medicina-ea-resolucao-do-cfm-avancos-e-zonas-cinzentas>. Accessed on: July 26, 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>.

BRAZIL. National Council for Scientific and Technological Development. CNPq establishes a Policy of Integrity in Scientific Activity, which sets standards and best practices for action. Brasília, DF: CNPq, March 11, 2026. Available at: <https://www.gov.br/cnpq/pt-br/assuntos/noticias/cnpq-em-acao/cnpq-publica-portaria-que-institui-politica-de-integridade-na-atividade-cientifica>. Accessed on: March 13, 2026.

BRAZIL. Ministry of Justice and Public Security. Guide to the ethical use of artificial intelligence: civil society and public authorities expand contributions to the MJSP. Brasília, DF: Ministry of Justice and Public Security, April 10, 2026b. Available at: <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>. Accessed on: April 12, 2026.

BRAZIL. Presidency of the Republic. Law No. 8,080, of September 19, 1990. Provides for the conditions for the promotion, protection and recovery of health, the organization and operation of the corresponding services and other measures. Official Gazette of the Republic. Union: Section 1, Brasília, DF, September 20, 1990. Available at: https://www.planalto.gov.br/ccivil_03/leis/l8080.htm. Accessed on: June 29, 2026.

BRAZIL. Presidency of the Republic. Law No. 15,378, of April 6, 2026. Establishes the Statute of Patient Rights. Official Gazette of the Union, Brasília, DF, April 7, 2026. Available at: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/lei/l15378.htm. Accessed on: July 26, 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/10.1177/00131644600200010>.

Brazilian Internet Steering Committee; Brazilian Internet Bureau (NIC.br). Sectoral studies: artificial intelligence in healthcare. São Paulo: CGI.br; NIC.br, 2024.

CORRÊA, NK *et al.* Worldwide AI ethics: a review of 200 guidelines and recommendations for AI governance. Patterns, New York, NY, v. 4, n. 10, art. 100857, 13 Oct. 2023. DOI: 10.1016/j.patter.2023.100857. Available at: <https://doi.org/10.1016/j.patter.2023.100857>. Accessed on: 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>. Available at: <https://doi.org/10.1007/s43681-024-00469-8>. Accessed on: June 18, 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/10.1561/04000000042>.

NATIONAL SCHOOL OF PUBLIC HEALTH SÉRGIO AROUCA. Ordinance No. 69, of June 1, 2026. To regulate the use of artificial intelligence and generative artificial intelligence tools and establish guidelines for verifying similarity and academic integrity within the scope of ENSP/Fiocruz. Rio de Janeiro: ENSP/Fiocruz, 2026. 8 p. Process No. 25388.000033/2026-21, SEI no. 6277864. Available at: <https://informe.ensp.fiocruz.br/assets/anexos/5ed838d9f860fa3f9f09e181d68d5069.PDF>. Accessed on: September 11, 2026.

HADDAD, AE; LIMA, NT Digital Health in the Unified Health System (SUS). Interface – Communication, Health, Education, Botucatu, v. 28, e230597, 2024. DOI: <https://doi.org/10.1590/interface.230597>. Available at: <https://doi.org/10.1590/interface.230597>. Accessed on: Apr. 9, 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, AVC. China and the Reach Zone. 2025. 151 p. Thesis (Doctorate in Public Policies, Strategies and Development) – Institute of Economics, Federal University of Rio de Janeiro, Rio de Janeiro, 2025. Available at: <https://www.ie.ufrj.br/images/IE/PPED/Teses/2025/A%20China%20e%20a%20Zona%20de%20Alcance.pdf>. Accessed on: September 22, 2026.

RUMELHART, DE; HINTON, GE; WILLIAMS, RJ. Learning backpropagation error representations. Nature, v. 323, n. 6088, p. 533–536, October 9, 1986. Available at: <https://doi.org/10.1038/323533a0>. Accessed on: April 2, 2026.

RUSSIA. Ministry of Health. Order No. 686n: Amendment of risk classification rules for medical software. Moscow: Ministry of Health of the Russian Federation, 2020. Available at: <https://medicaldevicesinrussia.com/2020/09/>. Accessed on: March 11, 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>. Accessed on: April 9, 2026.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. Adopted on November 23, 2021. Paris, FR: UNESCO, 2022. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Accessed on: July 12, 2026.

REGULATORY UPDATE REFERENCES 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. I.]: EDPB, 2026. Disponível em: https://www.edpb.europa.eu/public-consultations/guidelines-022026-on-anonymisation_en#no-back. Accessed on: September 22, 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. I.]: EDPB, 2026. Disponível em: https://www.edpb.europa.eu/public-consultations/guidelines-12026-on-processing-of-personal-data-for-scientific-research_en. Accessed on: September 22, 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. Available at: <https://www.ema.europa.eu/system/files/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle-en.pdf> . Accessed on: September 22, 2026.

EUROPEAN MEDICINE AGENCY; U.S. FOOD AND DRUG ADMINISTRATION. Guiding principles of good AI practice in drug development. [St. I.]: EMA; FDA Jan. 2026. Available at: https://www.ema.europa.eu/en/documents/other/guiding-principles-good-ai-practice-drug-development_en.pdf . Accessed on: September 22, 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/eng> Acesso em: 22 set. 2026.

MALTBY, J.; LIGUORI, L.; STEFANINI, E. Navigating Europe's AI regulatory labyrinth: A strategic guide for US pharma innovators. [St. I.]: Pharmaphorum, 31 Aug., 2026. [Secondary source used for theme identification; regulatory foundations were verified in official sources]. Available at: <https://pharmaphorum.com/market-access/navigating-europes-ai-regulatory-labyrinth-strategic-guide-us-pharma-innovators>. Accessed on: September 22, 2026.

APPENDIX A — CORRESPONDENCE MAP WITH GUIDE AND MARIAH

This map will guide future editorial cross-referencing. The definitive numbering. This may be updated after the edition goes to press.

- Ethical foundations and central questions — underpin scope, proportionality, and General criteria of the Guide and the initial axes of MARIAH.
- Critical concepts and properties — underpin context of use, explainability, Supervision, bias, life cycle, and validation.
- Brazilian context and international panorama — underpin the adaptation of instrument applied to Brazil and comparative methodology.
- Scientific integrity and authorship — underpin the recommendations on declarations of Use of AI, human verification, and accountability.
- Data governance — establishes the origin, purpose, and sharing of data. Re-identification, transfer, and destination of data and models.
- Health and SaMD — provide the basis for specific recommendations and risk factors from the Sciences. Health.
- Human and Social Sciences — provide the basis for sensitive inferences, consent, and data. public issues, collective risk, stigmatization, and secondary uses.