Responsible Artificial Intelligence (AI): Keys to applying ethical principles in AI solutions in the health field.

Technical document

Authors: Santiago Esteban,Rosa Angelina Pace,Velén Pennini, Adrián Santoro, Adolfo Rubinstein and Cintia Cejas

Hub publication date: 12/06/2024

The relationship between artificial intelligence (AI) in the healthcare domain and ethics is a topic of growing interest and debate. AI is defined as the field of study and development of systems and technologies capable of simulating human intelligence to carry out complex tasks autonomously [1]. In the context of healthcare, AI has become a promising tool with the potential to improve diagnosis, treatment, and disease management, as well as the analysis of large-scale medical and health data. However, the application of AI in healthcare poses a series of ethical challenges that must be addressed carefully and thoughtfully. The main issues to consider are the associated risks, primarily related to data handling and protection, as well as the biases that could occur or worsen, placing minorities from various backgrounds at a disadvantage and exacerbating existing disparities, such as those related to gender and others. Throughout this document, the concept of ethics and its associated principles are defined; the role of ethics in AI solutions is discussed; what biases are and why they are so important in the development of AI models, especially in the healthcare field, and finally, the document addresses, with examples, the application of ethical principles throughout the lifecycle of AI-based solutions: problem selection and definition, planning and design, development and validation, deployment and implementation, and operation and monitoring.

CLIAS

CENTRO DE INTELIGENCIA ARTIFICIAL Y SALUD PARA AMÉRICA LATINA Y EL CARIBE

Responsible Artificial Intelligence (AI): Keys to Applying Ethical Principles in AI Solutions in the Healthcare Field.

TECHNICAL DOCUMENT 3

September 2023

CIPS IMPLEMENTACIÓN E INNOVACIÓN EN POLÍTICAS DE SALUD

IECS INSTITUTO DE EFECTIVIDAD CLINICA Y SANITARIA

IDRC · CRDI International Development Research Centre Centre de recherches pour le développement international

Canada

Contenido

Team

Santiago Esteban: Family Doctor (Universidad Austral). Master's in Public Health, focusing on data at the T.H. Chan School of Public Health, Harvard University. Master in Business Administration from the University of San Andrés. Staff researcher in AI at the intersection of epidemiology, public health, machine learning, causal inference, data science, and information systems at CIIPS-IECS.

Rosa Angelina Pace: Medical Surgeon (UNNE) and Master in Bioethics from the Complutense University of Madrid (UCM). Coordinator of the Bioethics Center at the Italian Hospital of Buenos Aires (HIBA) and Director of the Department of Human and Social Sciences at the University Institute Italian Hospital (IUHIBA). Member of the Ethics Council in Medicine National Academy of Medicine. Received awards in Bioethics, Velasco Suarez OPS. Consultant CIIPS-IECS.

Velen Pennini: Bachelor of Anthropology (UNLP), specialist in Field Epidemiology by the training program in service of the Ministry of Health of the Nation, and specialist in Statistics applied to Health (FCEN-UBA). Researcher at CIIPS-IECS.

Adrián Santoro: Bachelor of Sociology (UBA) and Master in Generation and Analysis of Statistical Information (UNTREF). Works in the field of research in epidemiology, demography, and health statistics—expert in programming and development of mathematical models at CIIPS-IECS.

Adolfo Rubinstein: Family Doctor (UBA). Master in Clinical Epidemiology from the Harvard TH Chan School of Public Health, Diploma in Health Economics from the University of York. Doctor in Public Health (UBA). Regular Full Professor of Public Health (UBA). Certificate of implementation of public policies from the Harvard Kennedy School. Minister of Health of Argentina (2017-2019). Director of the Center for Implementation and Innovation in Health Policies (CIIPS-IECS).

Cintia Cejas: BA in Political Science (UCA) and Master in Social Sciences and Health (FLACSOCEDES). Specialist in health project management. Coordinator of the Center for Implementation and Innovation in Health Policies (CIIPS-IECS) and the Center for Artificial Intelligence in Health for Latin America and the Caribbean (CLIAS

This work was carried out thanks to the assistance of a grant awarded by the International Development Research Centre, Ottawa, Canada. The opinions expressed here do not necessarily represent those of the IDRC or its Board of Governors

1. Presentation

This document, prepared by the Center for Implementation and Innovation in Health Policies (CIIPS) of the Institute for Clinical Effectiveness and Health Policy (IECS), is part of a series of Technical Documents on Artificial Intelligence and Health (https://clias.iecs.org.ar/publicaciones/).

These documents aim to contribute to the region's knowledge, addressing different relevant axes and perspectives in the analysis of this subject.

Targeted at healthcare teams, health programs, policy formulators, decision-makers at all levels, and the general public, with a particular interest in the digital transformation of the health sector and its connection to sexual, reproductive, and maternal health (SRMH), this series of documents on AI that we are developing complement the activities carried out by CLIAS (Center for Artificial Intelligence in Health for Latin America and the Caribbean) developed at CIIPS, with the support of the International Development Research Centre (IDRC). For more information about CLIAS, visit http://clias.iecs.org.ar.

This document addresses the use of artificial intelligence (AI) in health from the responsibility perspective. Responsible AI refers to ethically developing, implementing, and using AI systems to minimize the risks and negative consequences associated with their application. This involves considering a series of principles and practices to ensure that AI benefits society as a whole and does not cause harm.

2. Key Messages of the Document

- In the context of health and healthcare, AI has become a promising tool with the potential to improve diagnosis, treatment, disease management, and the analysis of large-scale medical and health data.

- However, it is crucial to consider the associated risks, primarily related to data handling and protection, as well as biases that can occur or worsen, putting vulnerable groups at a disadvantage and exacerbating existing disparities, such as gender disparities or the exclusion of minorities, among others.

- The harms resulting from the application of AI can be both material, including security breaches (personal data leakage) and damage to patient's health (diagnostic errors), and immaterial, such as loss of privacy and dignity, limitations on freedom of expression, and discrimination in access to job opportunities, among other aspects.

- Consequently, there is a need to take the incorporation of complementary perspectives in the production and evaluation of AI applications in the healthcare field exceptionally seriously, including approaches from areas not limited solely to the technical development field. These additional approaches, especially those from the humanities and social sciences, and particularly from ethics experts, should participate from the initial stages of projects to try to mitigate biases in algorithms and programs conceived solely by technologists but which may overlook some aspects that contribute to the widening of inequalities and the neglect of other human values (integrated ethics).

- The principles of "safety," "self-determination," "benevolence," and "universalism" not only aim to ensure the responsible design of AI technologies but also pave the way for more equitable, inclusive, and beneficial solutions for society as a whole.

- Biases are errors or systematic deviations, or inclinations in the decisions or predictions of an AI model that can lead to unfair or inequitable outcomes. One such bias occurs when the data used to train the model does not adequately represent the diversity or variability of the target population. It can also happen when a database has problems regarding its structure, such as binary gender coding and erasing other gender identities into grouped categories.

- Therefore, efforts should be made to ensure that the dataset represents the target population as accurately as possible. In this sense, it is necessary to emphasize the importance of developments permeated by plurality, context, and intersectionality from their origins.

- Addressing these ethical aspects effectively requires collaboration among healthcare professionals, data scientists, computer engineers, lawmakers, decision-makers, and ethics experts.

- Therefore, it is imperative to establish solid regulatory frameworks that guide the development and implementation of AI in healthcare settings, ensuring that benefits are maximized while potential harms are minimized.

- Pursuing ethical solutions in AI guarantees the integrity and reliability of emerging technologies. It promotes a responsible and sustainable approach to technological innovation to benefit society as a whole.

3. Introduction

The relationship between artificial intelligence (AI) in the healthcare domain and ethics is a topic of growing interest and debate. AI is defined as the field of study and development of systems and technologies capable of simulating human intelligence to carry out complex tasks autonomously [1]. **In the context of healthcare, AI has become a promising tool with the potential to improve diagnosis, treatment, and disease management, as well as the analysis of large-scale medical and health data**. However, the application of AI in healthcare poses a series of ethical challenges that must be addressed carefully and thoughtfully. **The main issues to consider are the associated risks, primarily related to data handling and protection, as well as the biases that could occur or worsen, placing minorities from various backgrounds at a disadvantage and exacerbating existing disparities, such as those related to gender and others**.

Throughout this document, the concept of ethics and its associated principles are defined; the role of ethics in AI solutions is discussed; what biases are and why they are so important in the development of AI models, especially in the healthcare field, and finally, the document addresses, with examples, the application of ethical principles throughout the lifecycle of AI-based solutions: problem selection and definition, planning and design, development and validation, deployment and implementation, and operation and monitoring.

4. What is ethics?

Ethics is a discipline or knowledge that guides action, making it a form of practical knowledge. **It is a type of knowledge intended to guide rational behavior**. Within this category of helpful expertise, which focuses on directing action to achieve a tangible result, as in the realm of technology or art, ethics pursues a broader goal by attempting to reflect on and guide efforts toward rational, correct action [2], without ignoring the circumstances that always condition us, appealing us to choose wisely and rationally.

Different authors [3] argue that morality is closely linked to human nature and is essential to it. They also suggest that intelligence and, therefore, morality have roots in biology, and their purpose is to ensure the survival of our species. Intelligence functions as a form of adaptation, where human decisions are not solely determined by natural selection but include conscious and responsible choices. In the human species, we must justify the decisions we make.

This moral structure prevents us from being amoral. That is, we all must give content to our morality. We can act immorally, but we cannot wholly lack morality. The quality of our lives, actions, and projects will depend on how we fill that moral structure with content. Therefore, any human activity should be viewed in the light of positive values, aiming to improve the quality of life and minimize possible harm caused by its activity. In any of its applications, artificial intelligence is not exempt from this analysis.

Thinking about artificial intelligence ethically means being extremely careful to ensure that it is beneficial for humans and the environment, capable of empowering us in every way, especially in refining strategies for a healthy life, and avoiding any harmful actions such as direct harm or errors, discrimination, or injustice in its outcomes.

To achieve this, as already demonstrated, strategies for acceptable outcomes from an ethical point of view must be considered from the very conception of projects and include complementary perspectives, specifically technological ones, to try to prevent and mitigate unwanted outcomes.

5. Why talk about IA ethics?

According to Klaus Schwab, the world has embarked on its fourth industrial revolution, and the changes have acquired an unimaginable speed [4]. Schwab argues that while the digital revolution began in the mid-20th century, blurring the boundaries between the physical, biological, and computational realms through a fusion of technologies, the notable acceleration that has led to this fourth revolution has seen the emergence of artificial intelligence, robotics, blockchain, nanotechnology, biotechnology, among others, giving rise to cyber-physical systems. These cyber-physical systems are characterized by being a virtual representation of the physical universe, operating digitally and decentralized, and interacting through the "Internet of Things," a network of interconnected devices that can collect and share data over the Internet, enabling communication and automation between physical objects and digital systems.

The distinctive features defining the present revolution lie in its accelerated pace of advancement, its encompassing scope, and tangible impact in the physical realm.

In this context, it is imperative to recognize the relevance of ethical responsibility, which must be observed and critically evaluated during the development of these processes. Such ethical responsibility involves genuine reflection and deep dialogue, leading to a thorough understanding of humanity's obligations within the framework of this technological revolution, which can hardly be stopped nor should be stopped.

These introspective considerations must translate into concrete and practical actions in which **decisions regarding the creation of artificial intelligence (AI) tools are forged from the initial stages of design to their complete implementation and subsequent monitoring**. Caution and foresight in building these tools are crucial to ensure a responsible and ethical approach that safeguards the values inherent to human dignity and collective well-being. Various groups are working in the same vein including international organizations such as the WHO [5] and UNESCO [6].

Artificial intelligence (AI) has the potential to significantly transform society, and is a promising means to favor human prosperity, thus improving individual and social well-being, benefitting the common good, as well as supporting progress and innovation [7]. However, it is crucial to recognize that the implementation of AI also entails certain risks and challenges that must be addressed appropriately and proportionally. **Among these risks, highlights include the opacity in the functioning of systems, the widening of gender gaps, the exclusion of minorities, intrusion into individuals' privacy, financial speculation, and misuse in criminal activities or wars**.

The harms resulting from the application of AI can be both material, including damage to security

(personal data leakage) and people's health (diagnostic errors), and immaterial, such as loss of privacy, limitations on freedom of expression, dignity, and discrimination in access to job opportunities, among other aspects [6,7].

There has been some consensus regarding the principles that should govern the development and implementation of AI-based systems. However, many have expressed concerns about the inadequacy of these principles to properly guide actions, arguing that they are too generic in the face of actual and potential harm [8]. Consequently, **there is a need to take the incorporation of complementary perspectives in the production and evaluation of AI applications in the healthcare field exceptionally seriously,** including approaches from areas not limited solely to the technical development field. **These additional approaches, especially from the humanities and particularly from ethics experts, should participate from the initial stages of projects to try to mitigate biases in algorithms and programs conceived solely by technologists, which may overlook some aspects that contribute to the widening of inequalities and the neglect of other human values.**

In this direction, it is crucial to talk about **integrated ethics** or "embedded ethics"[9] that seek to address the need for major ethical guidelines for AI to be addressed and respected so that they can anticipate, identify, and mitigate these issues during the development of AI-based solutions. Thus, McLennan et al. [9] **propose a development model that integrates ethics from the beginning of projects, especially for healthcare professionals.** This model primarily promotes **integrated work between computer development teams, teams with thematic knowledge, and ethicist teams from the beginning of the project, ensuring transparency to the extent that it does not compromise confidentiality and intellectual property.**

This involves establishing coordinated objectives, the sought-after impact, and the methods used with regular exchanges and clear and explicit theoretical frameworks.

Solanki et al. [8] propose a series of human values mapped with ethical principles to guide teams in developing AI-based tools. These proposed guidelines emphasize the importance of ethics in developing AI tools, highlighting crucial aspects such as **safety**, **self-determination**, **benevolence**, and **universalism**. These principles ensure the responsible design of AI technologies and pave the way for more equitable, inclusive, and beneficial solutions for society as a whole. The following concepts are summarized and discussed:

Security	Safe AI systems are those that do not generate damages, dangers, risks, or threats as a consequence of their use. This encompasses both the psychological sphere and safeguarding against physical and social harms, as well as issues such as preserving privacy, integrity, and security. Primacy is given to the principle of **non-maleficence**, which implies a firm obligation to prevent harm, prevailing over the intention to promote good.
Self-determination	Consideration for individuals' intrinsic dignity, preservation of their autonomy, safeguarding of their fundamental freedoms, and unrestricted respect for their **right to informed consent** hold paramount significance concerning participation or submission to any AI system in the field of healthcare. Informed consent entails a complete understanding of the inherent implications of the proposal in question. While in the case of specific AI tools, characterized by their opacity resembling "black boxes" where the interior cannot be seen, complete comprehension may be hindered, it is undeniable that individuals increasingly perceive the inherent uncertainty underlying the healthcare field. Concomitantly, the right to privacy, recognized as a fundamental prerogative of human beings, imposes responsibilities on individuals and, especially, on medical personnel, among which **safeguarding confidentiality** stands out.[10]
Benevolence	The importance of contemplating, in the first instance, the imperative of beneficence, which refers to the obligation of healthcare professionals to act for the **benefit of patients and seek their well-being**; therefore, any AI-based tool used must be applied with this as one of its central objectives. Furthermore, **transparency** and "**explainability**" are fundamental. They refer to making the results of AI systems understandable by explaining how they were obtained to gain users' trust in these tools. These principles relate to understanding how the tools work, interpreting what they produce, comprehending their trends and deviations, and understanding how to interact with them. However, the discussion on the implications of explainability and its necessity remains relevant.[11] Finally, the authors emphasize the importance of including ideas of **solidarity**. Specifically, they suggest that it is essential to always consider, during the creation of these tools, the groups of people most vulnerable to the effects of social biases embedded in the data that may eventually result in markedly inferior performance. Potentially serious consequences could harm these groups, which differ significantly from the effects anticipated for the general population.
Universalism	The question arises from this concept: **Is it plausible that such a solution could equitably confer an advantage to all individuals?** (principle of justice). In that context, there is a need to elucidate what measures and mechanisms should be developed to counteract the inequalities manifested in areas such as the digital divide. Likewise, questions are raised about facilitating access for vulnerable groups and mitigating this solution's potential impact on underrepresented groups in the digital sphere (a focal point on equity). All these considerations become particularly relevant when addressing the issue of gender, as **inherent biases in technological fields, which often have a predominantly Caucasian and male composition, end up permeating the developed algorithms**. It is difficult to anticipate that these algorithms possess, from their genesis, a perspective in line with modern and inclusive perspectives, which advocates for the formation of heterogeneous development teams capable of reflecting the necessary diversity of approaches.

6. What are biases in AI models?

AI systems based on machine learning aim to improve prediction performance by optimizing a loss function[1], or in other words**, minimizing the error of those predictions**. However, errors come in different types. There is a random error that is intrinsic to AI systems and cannot be eliminated entirely but instead minimized. This error usually comes from various sources, such as sample size, data variability, and variations in training processes. On the other hand, **there is a non-random error known as bias**.

Biases are systematic errors or inclinations in an AI model's decisions or predictions that can lead to unfair or inequitable outcomes or simply incorrect results.

When working on developing AI tools, biases can occur in different ways or at different stages. One occurs when **the training data does not adequately represent the diversity or variability of the target population**. When this happens, the model may struggle to generalize, meaning it may not perform similarly with new or unseen data. A dataset can be biased from its design, underrepresented, or not adequately representing a particular population, which can lead to discriminatory behaviors. Even when well-represented**, a database may have problems regarding its structure**, such as encoding gender in binary form (female-male), thus obscuring other gender identities in grouped categories.

An important goal we must consider when working with these developments is to ensure that the dataset is the best possible representation of the target population.

This can be achieved through validation studies. However, the data training process may also incorporate biases into our results. Numerous studies have pointed out that, on a large scale, the problem of biases in AI comes from universities and companies that develop these technologies, mainly composed of white men with high socioeconomic status and very technical orientations. In response to this situation, it is proposed to move towards the development of collaborative AI projects involving social disciplines and engaging communities and civil society organizations. In this regard**, the importance of having developments influenced by plurality, context, and intersectionality from their origins is emphasized again.**

As explained, the notion of bias is complex, and humans also have biases in their practice. However, it is possible, and therefore ethically necessary, to design AI systems that help compensate for cognitive biases and thus lead to fairer and more equitable outcomes. An interesting scenario may arise when there is a disagreement between the expert professional and the AI tool designed to support decision-making. If the AI tool has been developed ethically and rigorously tested in different scenarios, it could have the ability to identify potential biases present in the professionals themselves. This ability of AI to pinpoint biases in experts is especially relevant, as these biases can be challenging to address through traditional strategies, such as case supervision or exchanging experiences with other professionals, since professionals in a similar area share biases.

The most relevant examples of using AI to address bias in decision-making come from the field of human resources. In the technology industry, where women and minorities are underrepresented, using AI in hiring decisions can lead to less biased decision-making and increase the promotion of women and minorities in these positions. By exploring the potential of AI in this regard, we can expand our understanding of the underlying biases in human decisions and work to mitigate them.

1 A loss function is a mathematical measure that evaluates how well a model based on automated learning fits the training data by quantifying the discrepancy between the model's predictions and the observed actual values.

7. Application of ethical principles of AI-based solutions

According to the Organization for Economic Cooperation and Development (OECD), the lifecycle of projects involving artificial intelligence could unfold in the following stages:

1. Problem Selection and Definition
2. Planning and Design
3. Development and Validation
4. Deployment and Implementation
5. Operation and Monitoring

Potential sources of bias during the lifecycle of AI-based solutions

1. Problem selection and definition:
    - Bias related to prevalence
    - Bias related to the history of access to the healthcare system
2. Planning and Design

- Bias related to unforeseen use
- Bias in addressing ethical and clinical limitations
- Bias related to algorithm design

3. Development and Validation
    - Bias related to training datasets

      Different access to healthcare

      Bias in exclusion/inclusion criteria
    - Bias relating to labeling

      Use of biased proxies for clinical outcome

      Bias in clinical evaluation
4. Deployment and implementation
    - Bias due to lack of model generalization
5. Operation and monitoring
    - Evaluation measures of equity from the patient and population perspective:

      Equitable access and outcomes

      Performance over time

      Inequitable performance among population groups

Adapted from: Abràmoff MD, Tarver ME, Loyo-Berrios N, Trujillo S, Char D, Obermeyer Z, Eydelman MB; Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D.C.; Maisel WH. Considerations for addressing bias in artificial intelligence for health equity. NPJ Digit Med. 2023 Sep 12;6(1):170. doi: 10.1038/s41746-023-00913-9. PMID: 37700029; PMCID: PMC10497548.

Based on the description provided previously, it is crucial to embed ethical considerations so that they are integrated into all phases of the development process of an AI tool. Below are these phases and the main ethical challenges involved:

7.1 Problem Selection and Definition

The decision-making process regarding what to develop in healthcare AI is complex and often involves various stakeholders. However, it is essential to ensure that this process is transparent and inclusive, satisfying the needs of all actors, taking into account:

- Potential Benefits and Risks: The benefits of AI in healthcare can be significant, but there are also potential risks. These risks include the possibility of AI having significant biases, being used for malicious purposes, or causing job loss in specific trades or professions.

- Ethical and Legal Implications: The use of AI in healthcare raises a series of ethical and legal issues. Among them are the right to privacy, the right to informed consent, and the possibility of reproducing certain social biases embedded in the data.

- Data Availability: Developing such solutions requires access to a large amount of data. This data can be challenging to obtain, and ensuring that it is accurate and representative of the target population while being treated confidentially is essential. The challenge lies in finding a balance between the availability of open data to promote the development of AI-based tools and the rigorous preservation of privacy and confidentiality of personal information to ensure that these developments are carried out ethically and responsibly.

- Development and Implementation Costs: Given the costs involved in developing and deploying AI-based healthcare solutions, it is essential to ensure that these solutions are affordable and accessible to all users and patients.

The inclusion of the perspectives of all key stakeholders from the outset of the project is essential to ensure products that add value to the healthcare chain while avoiding harm. This includes:

- Government Agencies: Governments regulate the development and use of AI in healthcare. For example, agencies regulating the administration of drugs and healthcare technology play a key role in approving AI-based medical devices.

- Healthcare Organizations: Service-providing institutions, both public and private, are responsible for implementing AI-based solutions in their clinical environments. They also play an essential role in collecting and managing the data used to train the models.

- Research and Development Teams of AI, Including Ethicists: These teams are responsible for creating and maintaining AI-driven healthcare solutions. They work with healthcare organizations to understand the needs of physicians and patients and to develop solutions that meet those needs.

- Patients and Users: For most use cases, patients are the end-users or the final recipients of decisions based on AI tools. They have a role in ensuring these solutions are safe, effective, and accessible.

- Research Institutions, Universities, and Academic Institutions: These institutions research the use of AI in healthcare. They develop new AI-based solutions and work to understand AI's ethical and legal implications in healthcare.

- Civil Society Associations (Non-Governmental Organizations, Scientific Societies): Sector

associations work to promote the responsible development and use of AI in health. They develop standards and guidelines for the use of AI and also provide education and training to healthcare professionals.

- Health Technology Industry: The health technology industry plays a crucial role in the responsible development of artificial intelligence. It is not only developers and funders of technologies, but its expertise in regulation and compliance with ethical standards gives it an advantageous position to promote good practices and even influence agendas. However, it is important to consider the clear conflicts of interest that such an actor presents.

7.2 Planning and design

Embedded Ethics

As previously mentioned, ethical considerations emerge as a fundamental imperative during the design process of AI-based systems. This includes the composition of the design team, the selection of data sources, and the contemplated use cases.

Once the ethical aspects to consider have been defined**, it is essential to identify those responsible for embedding these aspects into development**. In this regard, Miller [14] identifies two crucial issues when defining responsibilities for incorporating ethical aspects and bias control in the development processes of AI-based applications: **first, the multiplicity of profiles enrolled under the category of developers** (whom he considers ethically responsible actors in terms of developments [14-18]). Here, technicians, designers, financiers, etc., can converge. Secondly, he emphasizes that, within the framework of the projects within which the solutions are developed, there are often changes in the work teams, which means that professionals with different profiles enter and exit throughout these projects.

This requires a broad view of the role of actors or "stakeholders," traditionally defined as the players involved in a project who, in turn, are impacted by its results. In this sense, Miller [14] incorporates the figure of the passive stakeholder, which includes other actors who, without actively participating in the process, may end up being affected. This implies generating a balance that includes the evaluation of the community (and all the ethnic, cultural, and social problems that this implies) and the environment [19].

Own Development or Reuse

Part of the discussion during the design phase involves deciding on the possibility of reusing tools already developed. This is particularly relevant in the case of developing countries seeking to adopt technologies developed and implemented in developed countries. The inconveniences associated with the adoption of solutions based on other ethnic, cultural, and social contexts have been mainly highlighted [20-22], from the application of models trained originally in different languages to

applications that are assumed to be "plug & play," that is, for direct use without the need for local adaptation, without considering the existing differences in processes. In that context, the parties' role in each part of the process and their skills are decisive.

Governance

Another crucial aspect is data governance, defined as the **"organizing logic of data management: collection, storage, processing, use, exchange, and destruction."** Janssen et al. propose a framework for data governance to ensure that the correct data is shared securely and reliably and that the exchange complies with regulations [23]. This framework also promotes controlled opening of data and algorithms to allow external scrutiny, reliable information exchange within and between organizations, risk-based governance, system-level controls, and data control through (self-sovereign)[2] identities and shared ownership.

2 A "self-sovereign identity" refers to a system of personal data management in which an individual has complete and autonomous control over their own information, allowing them to selectively and securely share specific identity details online.

7.3 Development and Validation

The literature distinguishes **three key ethical challenges** facing the implementation of AI in medical and healthcare practice. These include **potential biases in AI models, protection of patient privacy, and the trust of physicians, users, and the general public** in incorporating AI in medical and healthcare [24,25].

Data Generation

Constructing databases or "datasets" for machine learning models constitutes an essential process in developing intelligent and automated systems. [26]. However, **it is crucial to recognize and address the inherent biases that may arise in such datasets**.

Biases in datasets can manifest in various forms. One of the most common is **selection bias**, which arises when the collected data does not adequately represent the diversity and variability present in the actual population [27]. For example, in applications such as facial recognition, data sets that underrepresent minority ethnicities can lead to unsatisfactory results for those groups [28]. For example, the same has been described in applications for identifying skin lesions [29,30]. These biases can be amplified during the training process, as algorithms tend to learn patterns from the provided data, regardless of whether they are appropriate or not.

On the other hand, **labeling biases** can be introduced when annotations are subjective or reflect cultural and social perceptions. This arises because annotators who label or classify the training data for a model are individuals embedded in society and bring their own biases to this task. For example, when classifying messages on social networks according to their polarity, annotators may condition the classification depending on how they interpret the gender of the message's author. This is more frequent when the annotation task is complex (assigning polarity to a text, detecting sarcasm or

irony, detecting hate speech, or medical diagnoses). This can be mitigated with detailed and field-tested annotation3 manuals that ensure adequate levels of standardization in annotation.

Effectively addressing these challenges requires a combination of approaches.

1. Firstly, a comprehensive analysis of the data for potential biases must be conducted. This involves assessing demographic distributions, gender relationships, ethnic characteristics, and other relevant variables to ensure that the dataset reflects the diversity of the population [32].

2. Subsequently, preprocessing strategies4 should be implemented, such as sample reweighting (which adjusts the weights of classes) or synthetic data generation (which creates artificial examples similar to existing ones for underrepresented groups) [32]. These techniques are necessary when there is an imbalance in the datasets concerning any of the variables of interest. Balancing or equalizing classes through one of the mentioned techniques significantly improves the model's ability to generalize when there are underrepresented classes or biases in our datasets. Incorporating feedback and review by domain experts is also essential at this stage to ensure adequate identification and mitigation of biases [32,33].

3 A detailed document that provides specific guidelines and instructions to human annotators on correctly labeling and annotating training data for a machine learning model.

4. Refers to the set of techniques used to prepare and clean data before training a model, typically to achieve better performance and generalization.

Furthermore, **transparency and detailed documentation of the dataset construction process are crucial**. Research teams should record all decisions made, from selecting data sources to cleaning and labeling methods. This allows for external critical evaluation and facilitates the early detection of potential unnoticed biases [33]. Additionally, **interdisciplinary collaboration should be encouraged, involving experts in ethics, diversity, and sociology, along with machine learning engineers, to ensure a comprehensive perspective and deep understanding of potential social and ethical impacts from the project's inception [8].**

Data Privacy

In the field of AI and health, it is common for developed tools to be trained with or use a considerable amount of personal and clinical data from patients as input. This means there is an inherent risk of this sensitive information being "hacked" or compromised somehow. Therefore, from our various roles, it is crucial to ensure that these developments are accompanied by respect for privacy and confidentiality.

Fortunately, **standards and guidelines have been established to manage the use of personal data in the context of AI**. For example, the Ibero-American Data Protection Network, which includes 22 data protection authorities from countries such as Portugal, Spain, Mexico, and others in Central America, South America, and the Caribbean, has published recommendations for treating personal data in artificial intelligence [34]. Some of these recommendations include designing appropriate

governance schemes for processing personal data in organizations conducting AI developments, ensuring the quality of personal data, using anonymization tools, and increasing trust and transparency with data subjects. Such initiatives are essential to ensure that AI advances do not compromise personal data privacy and security.

Data Preprocessing

The data preprocessing phase, aimed at rectifying inaccuracies from their original sources, holds significant importance in safeguarding the accuracy and fairness inherent to a given model. **The work in this area demands the ability to clearly discern vulnerable groups within the data in question (such as those derived from gender or ethnic identity variables), allowing us to measure disparities in data integrity and completeness among these groups**. The work by Larrazabal et al. [37] explores this aspect in machine learning models applied to X-ray images to predict different diagnoses. In this work, the researchers trained different models using training datasets with varied levels of gender imbalance, such as 25% men and 75% women or 0% and 100%. Then, they evaluated how these models detected pathologies in images of individuals of both sexes separately. They demonstrated that when gender imbalance in the training data was high, the model's performance decreased significantly in the underrepresented group, and even in cases of intermediate imbalance, such as 25% men and 75% women, the model's performance in the minority group was negatively affected. This has significant implications because the lack of consideration of this imbalance could lead to the generation of false positives or false negatives in predictions of diseases that particularly affect the underrepresented group.

Complex data-cleaning procedures are proposed in more challenging situations where datasets present more severe problems. An example of this is the MLClean preprocessing strategy proposed by Tae et al. [36], which aims to integrate data cleaning (removal of duplicates, correction of erroneous values), bias mitigation through data reweighting based on variables related to these biases, and data sanitization, i.e., the removal of confidential or sensitive information, into a single preprocessing pipeline.

Training and Model Validation

Performance metrics in an artificial intelligence model should be carefully selected according to their purpose. For example, in classification problems, evaluation can focus on determining whether it is more relevant to focus on false positives (e.g., regular X-rays misclassified by the model as "pneumonia") or false negatives (using the same previous example, X-rays labeled as "pneumonia" that are classified by the model as normal), depending on the context. However, when evaluating an AI model, it is also crucial **to use metrics that fairly reflect its performance across different groups or segments of the population, especially those considered vulnerable**. Thus, using global metrics can be misleading and biased, as it could conceal discrimination or underrepresentation issues in minority groups. A more suitable alternative would be to evaluate metrics for each specific group, such as different ethnic groups, genders, or other sensitive characteristics. For example,

standard metrics like the F1 score5 or even accuracy6 can be appropriate metrics if disaggregated by gender or race. This can significantly mitigate problems that arise when including variables in a model that improve overall performance but worsen it within specific groups.

In cases where models show poor metrics performance, whether for classification or regression, in any of the evaluated subgroups, it is advisable to analyze the examples within that subgroup and consider the possibility of retraining the model. However, accumulations (clusters) of similar examples may still have a high variability of characteristics, making their summary and interpretation difficult. Therefore, **it is essential to find an effective technique to detect subpopulations where performance metrics are deficient and simultaneously allow the identification of subsets of data that are easy to understand [8].**

Identifying problematic subsets through tools like Slice Finder [41] helps users improve the fairness of the model and provide more reliable and responsible decision-making results by identifying interpretable subsets of data where the model performs poorly. Alternatively, periodic audits can be conducted to identify subgroups with low performance, thus iteratively improving the model's performance for the target subgroups. Some biases can even be corrected with techniques involving re-labeling, such as Reinforcement Learning from Human Feedback, used in the popular chatbot chatGPT [42].

Realistic recognition by teams must accompany equity-addressing practices. Whenever possible, **the AI algorithm should be tested in multiple healthcare institutions, socioeconomic groups, and age ranges** [8,24].

The continuous search for approaches to address equity issues in artificial intelligence is crucial for advancing towards more inclusive and fair technological development.

5 A metric that combines sensitivity (or "recall" in English) and positive predictive value (or "precision" in English) of a model into a single measure, providing a balanced evaluation of performance by considering both false positives and false negatives.

6 A metric that measures the proportion of correct predictions made by a model out of the total predictions made.

7.4 Deployment and Implementation

Model Explainability

The notion of explainability refers to an artificial intelligence system's inherent ability to reconstruct the underlying process by which it arrives at certain predictions or specific outcomes

[43,44]. This attribute holds fundamental significance in adapting such systems and within the framework of the imperative ethical evaluation that governs their operation.

Different categories of AI-based systems pose heterogeneous challenges regarding their level of explainability. For instance, in the case of expert systems or symbolic artificial intelligence systems widely employed in the healthcare domain, clinical or medical knowledge is encoded using rule-based algorithms for decision-making. These systems are adapted by adjusting or modifying rules established by the scientific community and applying them to reference cases. This confers upon such systems a highly elevated degree of explainability, as each rule is explicitly encoded. Thus, each algorithmic decision can be traced back to a rule or combination thereof.

In contrast, **machine learning algorithms aim to discover intrinsic patterns in data to achieve an optimal level of generalization**. These algorithms "learn" by adjusting their parameters based on training data, optimizing a loss function (i.e., a function quantifying the discrepancy between the model's predictions and the actual training values) to solve specific tasks. **In the context of deep neural networks, the multitude of distributed computation layers between inputs and outputs obscures the underlying procedure, resembling a "black box.**" Consequently, comprehending their predictions is challenging since it is often not easy to "delineate" or unravel the flow of information numerically or in a visual representation, unlike simpler models such as logistic regression or decision trees [45].

This implies that **as the inherent complexity of the algorithm increases to improve prediction performance, the difficulty in accurately elucidating which rule or set of rules was instrumental in generating the prediction also intensifies** [45].

The trade-off between complexity, interpretability, or the level of explainability emerges as one of the predominant challenges in adopting such tools in the healthcare domain.

Explainability in AI can be intrinsic to the algorithm to be used (for example, linear regressions and decision trees) [46], or it can be an approximation made by other methods (for instance, LIME [47] or SHAP [48]) external to the model. This differentiation can help understand the common designation of some AI methods as "black boxes." It is essential to highlight that the explainability inherent to an algorithm will typically be more precise than other approximate explainability methods. However, it usually also performs less, as is the case with linear or logistic regression compared to convolutional or multi-layer neural networks.

Thus, **a trade-off between explainability and model performance must be considered during the development and validation of our tool, mainly when used to support clinical decisions**. Considering the growing preference for high-performance methods and the need for explainability in the healthcare domain, **an approach should be prioritized where multiple stakeholders critically evaluate and address explainability methods, prioritizing plurality, intersectionality, and interdisciplinarity.**

From a development perspective, explainability is essential for validating models based on coherence, not just performance. An example of a medical application reflecting this problem is

described by Zech et al. [49]. In their work, they describe the impact of confounding variables (e.g., image metadata embedded in X-ray plates) on model performance, which is clearly a problem for generalization. In this regard, it is necessary to work together with data scientists and engineers to understand the type of explainability or interpretability needed for the particular tool being developed.

The development of algorithms will be different if local explainability is sought (where a particular prediction needs to be explained, which variables weighed in that decision, or what rules were applied to reach it) or global (where the aim is to understand the model, such as the weights of the included variables or to approximate the model to a set of rules or decision tree) [50]. Thus, **developing models that seek to satisfy the need for explainability that users have and using human-centered approaches to keep people informed [50] will allow addressing the ethical and social issues associated with using AI in healthcare**.

7.5 Operation and Monitoring

Once these models have been deployed in production, meaning they are already being used within real-world processes, it is essential to **conduct continuous evaluations over time** to identify any early signs of performance deterioration. This becomes relevant **due to the possibility that changes in the population composition over time may differ significantly from the conditions initially considered during the initial deployment phase of the model**. Feng and colleagues [51] provide a detailed description of the monitoring process and **suggest areas for improving the quality of such systems within healthcare institutions**. Carrying out this monitoring process is a fundamental pillar, requiring a comprehensive consideration of the previously outlined elements. For example, it is imperative to conduct rigorous monitoring of the model's performance in contexts involving vulnerable groups and changes in the distribution of its variables.

8. Conclusions

The rapid evolution of artificial intelligence in recent years has posed a series of ethical challenges that demand rigorous attention and deep reflection. Advances in AI have demonstrated significant potential to transform various industries but have also raised concerns about privacy, algorithmic bias, accountability, and the substitution of human tasks. Effectively addressing these ethical aspects requires collaboration among healthcare professionals, data scientists, engineers, policymakers, and experts in bioethics.

Establishing robust regulatory frameworks that guide the development and implementation of AI is imperative. This will ensure that benefits are maximized while potential harms are minimized.

The pursuit of ethical solutions in AI ensures the integrity and reliability of emerging technologies and promotes a responsible and sustainable approach to technological innovation, which benefits society as a whole.

9. References

1. European Group on Ethics in Science and New technologies. RTD:Directorate-General for Research, Innovation, corporate-body. ETHI:European Group on Ethics, New Technologies. Statement on artificial intelligence, robotics and “autonomous” systems : Brussels, 9 March 2018. Publications Office of the European Union; 2018. Recuperado: https://data.europa.eu/doi/10.2777/531856

2. Cortina A, Orts AC. Etica de la empresa: claves para una nueva cultura empresarial. Trotta; 1994.Recuperado: https://play.google.com/store/books/details?id=IUIQSgAACAAJ

3. Shermer M. Morality is real, objective, and natural. Ann N Y Acad Sci. 2016;1384: 57–62. doi:10.1111/nyas.13077

4. Schwab K, World Economic Forum. The Fourth Industrial Revolution: what it means and how to respond. En: World Economic Forum [Internet]. 14 de enero de 2016 [citado 11 de agosto de 2023]. Recuperado: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolutionwhat-it-means-and-how-to-respond/

5. Ethics and governance of artificial intelligence for health. World Health Organization; 28 de junio de 2021 [citado 11 de agosto de 2023]. Recuperado: https://www.who.int/publications/i/item/9789240029200

6. Recommendation on the Ethics of Artificial Intelligence. [citado 2 de agosto de 2023]. Recuperado: https://unesdoc.unesco.org/ark:/48223/pf0000380455

7. White Paper on Artificial Intelligence: a European approach to excellence and trust. En: European Commission [Internet]. [citado 11 de agosto de 2023]. Recuperado: https://commission.europa.eu/publications/white-paper-artificial-intelligence-europeanapproach-excellence-and-trust_en

8. Solanki P, Grundy J, Hussain W. Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI and Ethics. 2023;3: 223–240. doi:10.1007/s43681-022-00195-z

9. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics. 2022;23: 6. doi:10.1186/s12910-022-00746-3

10. Davis SLM. The Trojan Horse: Digital Health, Human Rights, and Global Health Governance. Health Hum Rights. 2020;22: 41–47. Recuperado: https://www.ncbi.nlm.nih.gov/pubmed/33390691

11. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3: e745–e750.

doi:10.1016/S25897500(21)00208-9

12. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med. 2018;378: 981–983. doi:10.1056/NEJMp1714229

13. Mulya MODP, Ali M. Artificial Intelligence crime within the concept of society 5.0: Challenges and opportunities for acknowledgment of Artificial Intelligence in Indonesian Criminal Legal System. International Journal of Law and Politics Studies. 2023;5: 07–15. doi:10.32996/ijlps.2023.5.1.2

14. Miller GJ. Stakeholder roles in artificial intelligence projects. Project Leadership and Society.

2022;3: 100068. doi:10.1016/j.plas.2022.100068

15. Wieringa M. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–18. doi:10.1145/3351095.3372833

16. Touretzky DS, Cune CG-M, Martin F, Seehorn D. K-12 guidelines for artificial intelligence: What students should know. [citado 25 de agosto de 2023]. Recuperado: https://upload01.uocslive.com/ISTE/ISTE2019/PROGRAM_SESSION_MODEL/HANDOUTS/112142285/ISTE2019Presentation_final.pdf

17. Manders-Huits N. Moral responsibility and IT for human enhancement. Proceedings of the 2006 ACM symposium on Applied computing. New York, NY, USA: Association for Computing Machinery; 2006. pp. 267–271. doi:10.1145/1141277.1141340

18. Martin KE. Ethical Implications and Accountability of Algorithms. SSRN; 2018. Recuperado:https://play.google.com/store/books/details?id=6kb8zgEACAAJ

19. Derry R. Reclaiming Marginalized Stakeholders. J Bus Ethics. 2012;111: 253–264.doi:10.1007/s10551-012-1205-x

20. Mancilla-Caceres JF, Estrada-Villalta S. The Ethical Considerations of AI in Latin America. Digital Society. 2022;1: 16. doi:10.1007/s44206-022-00018-y

21. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. En: Friedler SA, Wilson C, editores. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR; 23--24 Feb 2018. pp. 77–91. Recuperado: https://proceedings.mlr.press/v81/buolamwini18a.html

22. Stray V, Hoda R, Paasivaara M, Kruchten P, van der Aalst W, Mylopoulos J, et al. Agile processes in software engineering and extreme programming: 21st international conference on agile software development, XP 2020, Copenhagen, Denmark, June 8-12, 2020, proceedings. 1 a ed. Stray V, Hoda R, Paasivaara M, Kruchten P, editores. Cham, Switzerland: Springer Nature; 2020. doi:10.1007/978-3-030-49392-9

23. Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: Organizing data for trustworthy Artificial Intelligence. Gov Inf Q. 2020;37: 101493. doi:10.1016/j.giq.2020.101493

24. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25: 1337–1340. doi:10.1038/s41591019-0548-6

25. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care.J Am Med Inform Assoc. 2020;27: 491–497. doi:10.1093/jamia/ocz192

26. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Jt Summits Transl Sci Proc. 2020;2020: 191–200. doi:10.1001/jama.2017.18391

27. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366: 447–453. doi:10.1126/science.aax2342

28. Buolamwini JA. Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Massachusetts Institute of Technology. 2017. Recuperado: https://dspace.mit.edu/handle/1721.1/114068?show=full?show=full

29. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. JAMA Dermatol. 2021;157: 1362–1369. doi:10.1001/jamadermatol.2021.3129

30. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018;154: 1247–1248. doi:10.1001/jamadermatol.2018.2348

31. Wiens J, Price WN 2nd, Sjoding MW. Diagnosing bias in data-driven algorithms for healthcare. Nat Med. 2020;26: 25–26. doi:10.1038/s41591-019-0726-6

32. Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Commun Med. 2021;1: 25. doi:10.1038/s43856-021-00028-w

33. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care:A call for open science. Patterns (N Y). 2021;2: 100347. doi:10.1016/j.patter.2021.100347

34. Red Iberoamericana de Protección de Datos. [citado 11 de agosto de 2023]. Recuperado:https://www.redipd.org/es

35. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018;169: 866–872. doi:10.7326/M18-1990

36. Tae KH, Roh Y, Oh YH, Kim H, Whang SE. Data Cleaning for Accurate, Fair, and Robust Models: A Big Data - AI Integration Approach. Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning. New York, NY, USA: Association for Computing Machinery; 2019. pp. 1–4. doi:10.1145/3329486.3329493

37. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical

imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A. 2020;117: 12592–12594. doi:10.1073/pnas.1919012117

38. Sramka M, Safavi-Naini R, Denzinger J, Askari M. A practice-oriented framework for measuring privacy and utility in data sanitization systems. Proceedings of the 2010 EDBT/ICDT Workshops. New York, NY, USA: Association for Computing Machinery; 2010. pp. 1–10. doi:10.1145/1754239.1754270

39. Gehrke J, Hay M, Lui E, Pass R. Crowd-Blending Privacy. Advances in Cryptology – CRYPTO 2012.Springer Berlin Heidelberg; 2012. pp. 479–496. doi:10.1007/978-3-642-32009-5_28

40. Pessach D, Shmueli E. A Review on Fairness in Machine Learning. ACM Comput Surv. 2022;55: 1–44. doi:10.1145/3494672

41. Chung Y, Kraska T, Polyzotis N, Tae KH, Whang SE. Slice Finder: Automated Data Slicing for Model Validation. 2019 IEEE 35th International Conference on Data Engineering (ICDE). 2019. pp. 15501553. doi:10.1109/ICDE.2019.00139

42. Stiennon N, Ouyang L, Wu J, Ziegler D, Lowe R, Voss C, et al. Learning to summarize with human feedback. Adv Neural Inf Process Syst. 2020;33: 3008–3021. Recuperado: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

43. Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion. 2021;76: 89–106. doi:10.1016/j.inffus.2021.05.009

44. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med. 2022;133: 102423. doi:10.1016/j.artmed.2022.102423

45. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58: 82–115. doi:10.1016/j.inffus.2019.12.012

46. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc. 2020;27: 1173–1185. doi:10.1093/jamia/ocaa053

47. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. pp. 1135–1144. doi:10.1145/2939672.2939778

48. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30. Recuperado: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

49. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15: e1002683. doi:10.1371/journal.pmed.1002683

50. Liao QV, Gruen D, Miller S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–15. doi:10.1145/3313831.3376590

51. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. NPJ Digit Med. 2022;5: 66. doi:10.1038/s41746-022-00611-y

CLIAS

CENTRO DE INTELIGENCIA ARTIFICIAL Y SALUD PARA AMÉRICA LATINA Y EL CARIBE

CIPS IMPLEMENTACIÓN E INNOVACIÓN EN POLÍTICAS DE SALUD
IECS INSTITUTO DE EFECTIVIDAD CLÍNICA Y SANITARIA