NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

ARTIFICIAL INTELLIGENCE FOR GOOD HEALTH

ETHICAL CONSIDERATIONS FOR THE IMPLEMENTATION OF AI SYSTEMS IN THE HEALTHCARE SECTOR

Maria João Morgado Marques

Dissertation

presented as partial requirement for obtaining the Master's Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ARTIFICIAL INTELLIGENCE FOR GOOD HEALTH – Ethical Considerations For The Implementation Of AI Systems In The Healthcare Sector

by

Maria João Morgado Marques

Dissertation presented as partial requirement for obtaining the Master's Degree in Advanced Analytics, with a specialization in Data Science

Supervisor: Prof. Roberto André Pereira Henriques

October 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 28 October 2023

To my parents and sister,
For the constant interest, endless support and for all the selfless sacrifices.
Thank you for believing in me, even when I couldn’t.

ACKNOWLEDGEMENTS

I want to express my gratitude to my supervisor, Prof. Roberto Henriques, for his guidance throughout the journey of completing this thesis. Your insightful feedback has shaped my research and strengthened my academic growth.

I am also immensely grateful to all the professionals who have crossed my path this year, particularly the ones devoted to the healthcare sector, whose encouragement, assistance, and constructive discussions have been fundamental in shaping my ideas and refining my work. Their passion and endurance have endlessly inspired me.

Furthermore, I thank my university colleagues from my study cycles for all the shared experiences, moral support, and friendship. Our shared moments gave me strength and have brought me here today.

Lastly, I couldn't cross this stage without the support of my family, closest friends, and boyfriend. Your constant belief in me and continuous encouragement were the main factors that have allowed me to achieve this important step.

ABSTRACT

The role and presence of artificial intelligence (AI) are increasingly growing in everyone's lives due to its employment in several industries, and healthcare (HC) is not an exception. With its particularities, many believe the quality of care can benefit from the contribution of such technology to an industry that guarantees one of human's fundamental rights. Many, however, are reluctant towards its disruptive power. Research showed how much of it lies with Ethical concerns that are seen as crucial to guarantee an application of AI that is trustworthy and value-upholding, which justifies the significant difference between published AI studies for HC and the implemented ones on the "bedside" of patients who could benefit from it. To bridge this gap, the present work explores how ethical considerations can be incorporated into AI development and implementation for HC. To answer this, a literature review was carried out on AI, AI presence in the HC industry, and ethics. Based on a quantitative research methodology, three case studies were analysed to propose a framework that encapsulates ethical considerations research showed to be more significant for the industry. The produced framework also reflects what the AI lifecycle should be when producing a trustworthy artefact moved by engaged multidisciplinary stakeholders. With the present work, Portugal's findings were analysed, essential practices and guidelines were highlighted throughout the framework, and a step further was given to leverage AI to a responsible end while avoiding an "AI winter".

KEYWORDS

Artificial Intelligence; Healthcare; Ethics; Framework

Sustainable Development Goals (SGD):

"Just as a microscope proved an invaluable tool in medicine and biology when it was developed and people learned how to use it, so AI will prove an invaluable tool as it is developed and people learn how to use it. It is a multipurpose tool that can transform healthcare."

Professor Mihaela van der Schaar
(John Humphrey Plummer Professor of Machine Learning, Artificial Intelligence and Medicine at the University of Cambridge, for Topol Review 2019)

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
ASI	Artificial Super Intelligence
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CP	Constraint Programming
CT	Computerized Tomography
CV	Computer Vision
DL	Deep Learning
DT	Decision Tree
EA	Evolutionary Algorithm
EC	European Commission
EMR	Electronic Medical Record
EMS	Emergency Medical Services
EU	European Union
FM	Feature Map
HC	Healthcare
LLM	Large Language Model
ML	Machine Learning
MRI	Medical Resonance Imaging
MSS	Master Surgical Schedule
NB	Naïve-Bayes
NLP	Natural Language Processing
NN	Neural Network
RNN	Recurrent Neural Network
RQ	Research Question
SSS	Surgery Scheduling Solution
SVM	Support Vector Machine
USD	United States Dollar

1. INTRODUCTION

1.1. CONTEXT

Artificial Intelligence (AI) was first defined by Alan Turing (1950) – considered its founding father - as "the science and engineering of making intelligent machines, especially intelligent computer programs". More recently, there are numerous ways of defining AI since it has become a prominent topic in today's society (Joint Research Centre European Commission *et al.*, 2020), and its potential to transform industries and revolutionize how we live, work and communicate has been widely recognized (Gates, 2023; Tai, 2020; Bossmann, 2016).

AI technology is already being used in various applications, from virtual assistants to self-driving cars (Ouchchy *et al.*, 2020), and the availability of large amounts of data, advancements in computing power and the development of sophisticated algorithms have fuelled its growth. However, as AI becomes more pervasive, concerns have arisen about its impact on society (Bossmann, 2016), including issues related to privacy, algorithm fairness and bias, liability, transparency, and job displacement (Gerke *et al.*, 2020; Stanfill *et al.*, 2019; Murphy *et al.*, 2021; Vogel, 2017).

Healthcare (HC) is not an exception to the increasing interest in AI, mainly since the outburst of the COVID-19 pandemic, which accelerated digital transformation. At the same time, HC professionals, institutions, authorities, and health technology market suppliers responded to the crisis while maintaining services under extreme pressure, according to the 2021 Healthcare Information and Management Systems Society (HIMSS) Annual European Digital Health Survey. Under the same study, diverse HC stakeholders reported that the pandemic exposed significant connectivity and integration gaps in many countries, which only shows that although hospitals and HC providers have a crucial and confirmed role in the economic and social dimensions of development, these are often the institutions that present an urgent need to improve their processes.

The AI breakout reached the HC industry worldwide but at different rates. A study conducted by Infosys in 2018 showed that although HC was one of the first practical applications for early AI systems like DENDRAL,[1] the average investment, at 4.7 million USD, was lower than the overall average across all industries. Worldwide, HC organizations still focus on investing in IT infrastructure (62%), while building AI into the company ethos stays at 46% (Infosys, 2018).

Focusing on the European Union (EU) context, and according to the 2019 HIMSS Analytics report on Health-IT predictions for Europe, HC systems haven't become much brighter over the analysed time as little progress was made, although the vast majority of patient records in Europe are digitised. Another key finding is that while the average European HC provider organisation spends between 2.9% and 3.9% of its total annual expenditure on digital products and services, the majority (63%) of HC employees think their institutions' IT budget is too low. According to the same source, this can be a sign of frustration *i.e.,* digital solutions are not delivering the expected benefits from an end-user perspective. A report held by the European Commission (EC) in 2021 also shows that the state of AI adoption in HC organizations across the EU suggests a slow and overall low level of implementation,

[1] DENDRAL, a chemical analysis expert system, was developed by AI researcher Edward Feigenbaum and geneticist Joshua Lederberg starting in 1965. Originally called Heuristic DENDRAL, the system would use spectrographic data to propose the molecular structure of a substance (Copeland, 2019).

with varying levels in each country (EC, 2021b). Converging to Portugal, the EC (2021a) identified the lack of regulations and policies as a general barrier to the further adoption of AI technologies in HC. This can be paradoxical when the scientific contribution of Portugal is relatively significant concerning its population (Portugal contributes approximately 8% of scientific output in the area of AI in HC, ranking 5th amongst EU countries). The leading publications refer to the domains of patient monitoring, disease diagnostics and the predictive power of AI in health, *e.g.,* triage waiting time in maternity emergency care or the use of data mining in cases of postoperative complications with gastric cancer patients (EC, 2021b). Concerning Portuguese hospitals, a report provided by Glintt and APAH[2] (2022) has shown that although image interpretation is the area with the highest expectation (36%), it is demonstrated that such topic has a lower implementation potential, mainly due to the lack of knowledge or absence of implementation plans in the near future and not to the fact that they are already being implemented. The most significant facilitator of AI adoption for the respondents would be the recognition of the benefits of integrating AI into their daily lives by health professionals (55%). One can easily perceive a significant gap in the HC academic and professional context since the second one cannot keep up with the first. A reason for this is pointed out by the EC (2021): Since AI systems will make critical decisions autonomously, transparency and auditability will be demanded by society to foster safety and ethical principles. This comes to show that definition and implementation of ethical guidelines are crucial for all involved – HC professionals, HC solution developers, and HC patients.

Amidst the birth of AI, Ethics has been identified as a priority concern in the development and implementation of AI across multiple sectors (Bossmann J., 2016; Gibney E., 2020; Ouchchy *et al.*, 2020). Several recent interventions from relevant personalities in AI development, such as Elon Musk[3] or Steve Wozniak[4], call for a pause on AI development, mentioning "profound risks to society and humanity" (Metz *et al.*, 2023), or defining some of the applications that AI has potentialized as "quite scary" (Geoffrey Hinton[5]). Endeavours as these contributed to a consensus regarding the appliance of ethical considerations when developing AI tools. Several institutions, from the EC (and some countries themselves), corporate enterprises and governmental entities have issued articles, reports (*e.g*. Infosys, 2018; Portugal INCoDe 2030 *et al.*, 2019; EC *et al.*, 2022; NHS, 2019) and even conferences (Kidd, 2020) regarding the communion of these two topics – AI and Ethics. In what concerns the HC system, these concerns are frequently intensified due to the characteristics of the HC industry. To some extent, ethical topics are commonly pointed as reason for some delay in the application of AI solutions in this environment when compared to other industries since it is perceived that only once these topics are addressed will AI be able to revolutionize HC (Whitby, 2014; Gerke *et al.,* 2020; Cordeiro, 2021).

1.2. STUDY OBJECTIVE

Health is a fundamental human right. The potential benefits of AI in improving health conditions imply that the promise of AI to advance human rights. However, it also risks challenging human rights by perpetuating existing societal biases through biases in data and algorithms and the opacity of increasingly complex AI processes. As such, an ethical debate regarding AI in HC becomes critical. Being so, the main ambition of this document is to provide an answer to the Research Question (RQ): "How can ethical considerations be incorporated into the development and implementation of AI for HC?".

[2] Associação Portuguesa de Administradores Hospitalares.

[3] CEO and chief engineer of SpaceX; CEO and product architect of Tesla, Inc.; CEO ad owner of Twitter.

[4] Apple, Inc. Co-Founder.

[5] Ex-Google; "Godfather of AI" due to being one of Neural Networks' pioneer.

While aiming to do so, an AI technical background will be provided to understand the execution of AI on three distinct applications for the HC industry. Ethical principles will be studied, and leading ethical factors will be highlighted to grasp how they have been addressed alongside technology, particularly in three case studies, whose analysis aims to set best practices for an ethically aware AI project. Based on these findings, a suitable answer for the RQ will be obtained while producing a relevant analysis for society.

As a secondary objective, data from Portugal will be analysed in multiple sections since one can point out that literature is scarce when it comes to it. Additionally, it is expected that literature mainly contains the HC user perspective rather than including the remaining stakeholders when approaching ethical concerns. Physicians and other practitioners, researchers, and AI solutions developers are examples of other roles that will be extensively addressed in this work besides the patient.

Ultimately, there is the drive to provide a framework to assist AI development and implementation in HC that contains how, in what phase and with whom the most relevant ethical topics should be addressed, according to the findings of the fundamental guidelines and regulatory documents in the field.

1.3. Study Relevance and Importance

According to Glintt and APAH (2019), AI supports clinical decision-making, resources and installed capacity optimization, improvement of citizens' experience and their contact with institutions, as well as anticipation of health states, whether of individual or public health. However, according to van de Sande *et al.* (2022), 90%–94% of the published AI studies remain within the testing and prototyping environment and have poor study quality. Clinical benefits fall short to the high set expectations.

This absence of clinical AI adoption is alarming and raises the possibility of an “AI winter”, in which the enthusiasm surrounding AI will wane and expectations will become unrealistic. Therefore, a suitable answer to the RQ can provide a practical approach to prevent such tendency. Due to the increasing voices of concern regarding the presence of this technology in the HC industry, the present study has the potential to be high relevance for the field and for society.

1.4. Methodological Outline

As proposed by Natrup, S. (2022), the present work will be composed of four distinct phases – Exploration Phase, Analytical Phase, Execution Phase and Conclusive Phase.

The first part is characterized by elaborating of the RQ and the study objectives. With these, it will be possible to identify publications and assess their relevance for the defined objectives. Moving to the Analytical Phase, the literature will be reviewed, making it possible to determine challenges that should be deconstructed throughout the developed work. Qualitative Research with content, and themati analysis will then be conducted using a case study approach, inserted in the Execution Phase. Lastly, to provide a framework for systematic ethical appraisals throughout the development of AI solutions for the HC field, all the collected information will be discussed, making it possible to fully answer the RQ and look into future developments and work limitations.

The following figure organises the described procedures and intermediate steps to achieve the proposed objectives.

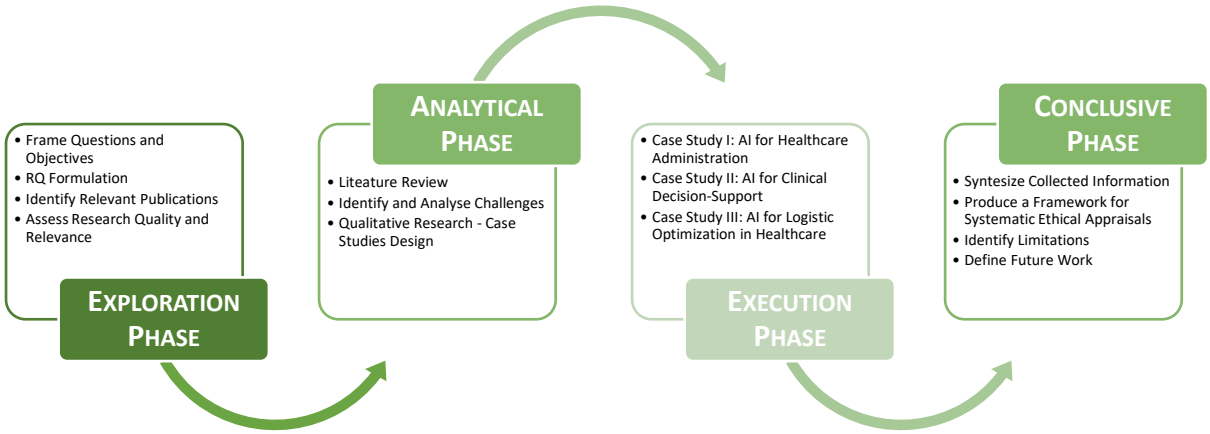

Figure 1.1 - Methodology Outline

1.5. Report's Structure

In the present chapter, a look closer at this work's aims and why it is relevant will be held.

The next chapter presents a summary of the literature, giving an overview of the main topics to grasp to produce results and discussing AI and ethics in the light of HC. This chapter summarizes an AI technical background, focused on its presence in three main areas of the HC sector, while exploring the state-of-the-art solutions for each of the case studies covered in the execution phase. Furthermore, an overview of the ethical theory emphasises the main topics impacting the implementation of AI solutions in the HC sector. Lastly in this next chapter, an outline of the most critical work developed to deliver a framework for building and supplying AI ethically will be carried out.

Chapter three will be used to describe the methodology used – Qualitative Research -, applying the case study approach. In the fourth chapter - Results and Discussion - the results of the developed work are observed, and discussed, sharing a reflection that compares the problem and objectives with the final product, while providing a critical outlook onto the case studies seen. In contrast, in chapter five, conclusions are drawn from the work developed and the comparison between what was expected, and the achieved results will be held. The last chapter describes the limitations found during this process and possible future work paths.

2. LITERATURE REVIEW

2.1. LITERATURE REVIEW METHODOLOGY

To answer how ethical considerations can be incorporated to deploy AI solutions for the HC sector, it is crucial to understand the developments related to this topic, as the present work also aims to produce a helpful synthesis of the literature regarding this.

The current document follows a systematic literature review approach, following the SALSA steps (Search; Appraisal; Synthesis; Analysis), which according to Grant & Booth (2009) are respectively:

- an exhaustive, comprehensive search, that time constraints may limit.
- a quality appraisal that determines the inclusion/exclusion of a particular paper.
- a narrative synthesis.
- an analysis that contains what is known, and recommendations for future research.

From the search step emerged literature selection criteria. The considered documents result from a catalogue searching for holdings and exploration of the computerized databases that are typically reviewed in academia and enterprises and are connected to the topics of AI, HC, Machine Learning (ML), Ethics for AI, Social Sciences, Medical Informatics, amongst others. These include articles, video resources, and reports by field authorities and popular media outlets, *e.g.*, Medium, World Economic Forum, HIMMS, EC, IBM, McKinsey, Deloitte. The mentioned documents were reached by using keywords identified during preliminary readings for the RQ formalization. Some examples of such are "Artificial Intelligence in Healthcare", "Ethical Artificial Intelligence", "Deep Learning", "Machine Learning Healthcare Applications", "Ethical Considerations", "Ethical Challenges", "Clinical Decision Support", "Administrative Support", "Logistics Support", and its combinations.

Regarding the appraisal part, there was an effort to present the most recent sources so the utmost recent information would be considered due to the volatility of these topics that are fast changing. It's relevant to further notice that the inclusion of certain research documents was determined by its relevance, established mainly by the number of times it has been cited, and by the characterization of a high number of the chosen keywords for the search development. The authors at stake are identified and quoted when the purpose of the present work justifies it. Additionally, grey literature was also considered, keeping the expert-driven criterion, consulting mainly organization/institution reports. Some additional (exclusion) criteria were also applied, namely if the article was neither in English nor Portuguese languages.

Since this is a theme that has shown to create a society-level interest, the synthesis and analysis part resulted into a literature review that contains a technical background regarding AI - its context and its most relevant techniques (that are related to the current work). This way, the full document will be accessible for a broader audience. After this, state-of-the-art research is addressed, namely concerning the bordering artefacts that concern the goal of this study – to produce a framework designed for an ethical-driven AI solution in an HC environment.

In the following graph, one can have a global perspective on the unprecedented rate of research publications covering some of the used topics to fuel the present work, demonstrating a growing interest in the field, backing its relevance.

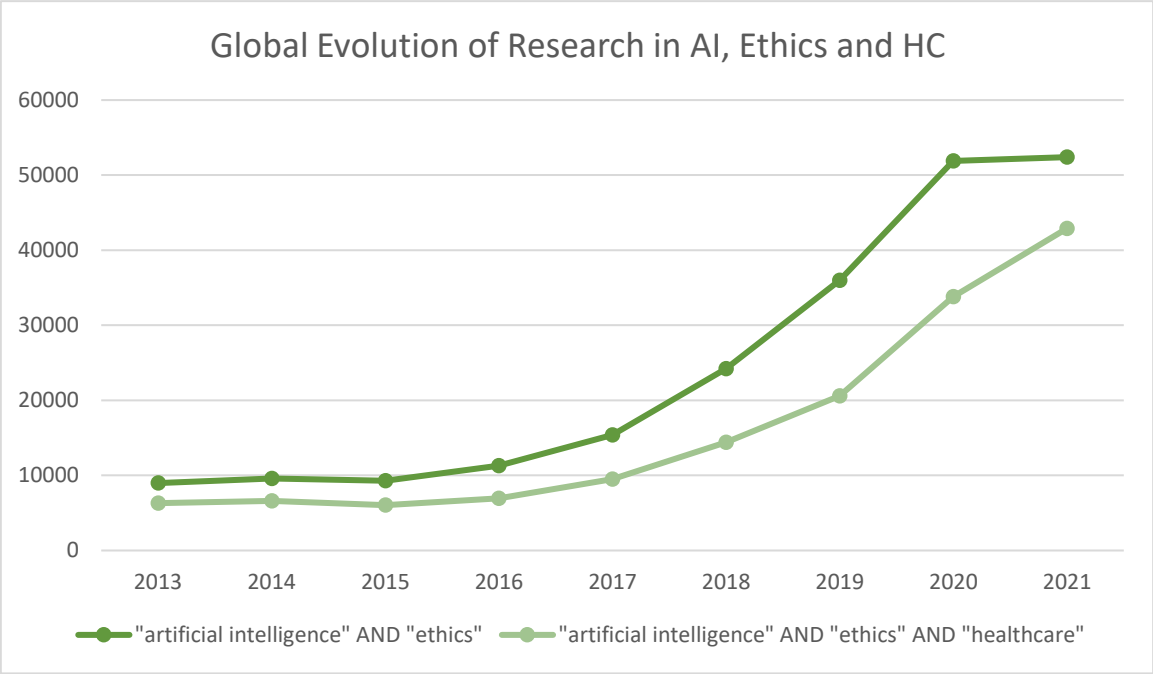

Figure 2.1 - Global evolution of research in AI, Ethics and HC during the last decade. The number of AI papers presented by Google Scholar was arranged by year, 2013–2021. The points represent the number of studies. The performed search included the terms exposed in the legend.

2.2. TECHNICAL BACKGROUND

2.2.1. Artificial Intelligence Essentials

As proven before, there is a visible interest in AI and all its possible applications. However, a standard definition for what AI means and concerns is not available (House of Lords' Select AI Committee, 2018). Specific approaches to human intelligence, or intelligence in general, have been used to characterize AI. Machines that act like people or can do tasks that call for intelligence are frequently mentioned in definitions (*e.g.,* McCarthy, 2007). The objective characterization of something as subjective and abstract as human intelligence offers a misleading sense of precision that cannot be achieved (Kaplan, 2016). As a result, most definitions in research, policy, and market studies are ambiguous and suggest an ideal aim rather than a quantifiable research concept.

Regardless of this scenario, there are commonalities when exploring definitions despite AI's multiple environments, that can provide a possible formulation of what might be regarded as the main characteristics of AI, according to Joint Research Centre (EC) *et al.* (2020):

- Environmental perception, considering the complexity of the real world.
- Information processing (gathering and analysing inputs in the form of data).
- Making decisions (including reasoning and learning) involves taking actions and performing tasks with some autonomy.
- Achieving particular goals is the primary motivation behind AI systems.

Being so, the High-Level Expert Group on AI, or HLEG, appointed by the EC and composed of 52 elements from the academia, civil society and industry, considers AI systems to be "software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions."

The term *artificial intelligence* was first used by John McCarthy, an assistant professor of mathematics at Dartmouth College in Hanover, New Hampshire in 1956 (Kaplan, 2016). However, it is believed that Alan Turing, an English mathematician, was the first to initiate field research with a lecture given in 1947 and a 1950s *Computing Machinery and Intelligence* article, where the Turing's Test was published to discuss criteria for determining whether a machine is intelligent. He stated that you should consider a machine intelligent if it could convincingly pass for a human to an informed observer (Kok, n.d.; Turing, 1950). Although it didn't satisfy all, it undoubtedly triggered several events in the field that brought us to the present. Michael I. Jordan[6] claims that the History of AI - and practically any other engineering branch - is basically attempting to construct things. In *Towards a Blend of Machine Learning and Microeconomics* (Open Data Science, 2020), Jordan explores the history and potential of ML while emphasizing that the algorithm is simply one component of a more extensive system of intelligence.

As seen so far, when looking at today's scope of AI, one should go further, given that it encompasses the development of intelligent systems, ML, natural language processing (NLP), decision support, and their applications across various domains - which naturally includes HC and its implied ethical consideration and social implications, as the following chapters will prove.

2.2.2. Artificial Intelligence Algorithms and Techniques

According to Kavlakoglu (2022), AI is "the broadest term used to classify machines that mimic human intelligence". Since it is used to predict, automate and optimize tasks once done by humans, it can have practical uses such as decision-making or speech and facial recognition. As such, AI can be categorized into three distinct categories – Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), or Artificial Super Intelligence (ASI). ANI or *weak* AI, according to Kaplan (2016), refers to systems that are restricted to a narrow domain, functioning, made, and developed for that specific task only, while *strong* AI, or AGI, proposes that machines do or eventually will have minds, which means, in a less "apocalyptic" manner, that these represent systems that exhibit general intelligent behaviour, compared to humans. Besides these, ASI, or *superintelligence*, would still exceed a human's intelligence and ability. Although AGI is in a premature phase, and ASI is still hypothetical, AI fields, namely ML, neural networks (NNs) and deep learning (DL) are quickly fuelling these trends.

The three first fields – ML, NNs and DL – are deeply connected since each depends on the prior. ML, first mentioned by Arthur L. Samuel (1959), enables computers to gain knowledge from experience

[6] Professor at the University of California and 2022 WLA Prize in Computer Science or Mathematics winner.

and improve over time without explicit programming. An ML algorithm can learn to make predictions or solve problems. Additional human involvement is needed to categorize data using traditional (non-deep) ML algorithms, such as feature learning. Looking deeper into the algorithms functioning for ML, UC Berkeley (2020) breaks it into three parts. Firstly, during the decision process, the input data will lead to a pattern that the model will aim to predict. Next, for the error function, an evaluation regarding the performance of the model when predicting will be performed. Here, one can compare it to known examples and assess its quality. With this assessment, it is now possible to reach the third part, where the model process is updated and optimized. Here, one aims to reduce the difference between the previous results and the known example until an accuracy threshold is met. With this example, the supervised learning method was introduced. However, other approaches, namely unsupervised learning, or reinforcement learning are still possible. For the first option, one intends to cluster unlabelled data by discovering patterns. In contrast, for the second one, the learning process takes place by trial and error, meaning that the algorithm isn't trained with sample data. In a simpler way, the main difference is whether there is labelled data available.

Before getting into the specifics of NNs, numerous other statistical, probabilistic, and optimization methods, including Support Vector Machines (SVMs), Decision Trees (DT), and Naïve-Bayes (NB), can be used as learning techniques. SVMs effectively find subtle patterns in large datasets (Aruna, 2011), and it seeks to establish a decision boundary between two classes to predict labels from one or more feature vectors. The hyperplane, a decision boundary, is oriented to be as far away from each class's nearest data point as is physically possible, and support vectors are used to describe these nearest places. The kernel approach, an alternate application of SVMs, enables us to model larger dimensional, non-linear models. Trial and error is the only approach to selecting the best kernel. For example, cross-validation can choose an optimal kernel function from a fixed set of kernels (Huang *et al.*, 2018).

Looking now at DTs, a non-parametric supervised learning approach that can be used for regression and classification, they aim to learn straightforward decision rules derived from the data features to build a model that predicts the value of a target variable. In the words of Rokach *et al.* (2005), a DT is made up of nodes that make up a rooted tree, which is a directed tree with a node called "root" that has no incoming edges, and all other nodes have exactly one incoming edge. Each internal node in a DT divides the instance space into two or more sub-spaces in accordance with a specific discrete function of the values of the input attributes.

NB classification consists of a probabilistic classifier that applies the Bayes theorem while making strong (naive) assumptions about the independence of the characteristics (Vembandasamy *et al.*, 2015). In other words, NB classifiers use the probability theory to find the most likely classification of an unseen instance. Despite these characteristics, NB has demonstrated success in various real-world settings, such as text classification and medical diagnosis (Rish, 2001).

Regarding NNs, a commonly used ML algorithm is being addressed. NNs aim to mimic how the human brain interprets and intelligently categorizes data. "Artificial neurons", or primary processing nodes, make up a NN. These nodes are interconnected in layers. In this process, data is "fed-forward" - each node receives data from several "previous" nodes and transmits it to a certain number of nodes placed "afterwards". The data that nodes receive is given a "weight" and an assigned value. When the algorithm is trained, the nodes' weights and thresholds are modified (via backpropagation), until similar data inputs yield consistent outputs. Importantly, NN algorithms are built to learn quickly from

input training data, enhancing the accuracy and effectiveness of the network's algorithms. As a result, NNs serve as an important illustration of the strength and capability of ML models. When speaking of DL, a more modern adaptation of NNs that employs multiple ("deeper") layers of synthetic neurons to tackle more challenging tasks is being mentioned. A higher volume of training data is being used, demanding more computer power. Since the middle of the 2000s, it has become increasingly popular as a technique. It is mainly responsible for the current surge in public interest in AI due to its strong association with human-level AI. Computer vision (CV) is fed primarily by CNNs (Convolutional NNs), and speech synthesis and recognition are possible via RNNs (Recurrent NNs), where sequential or time series data is leveraged.

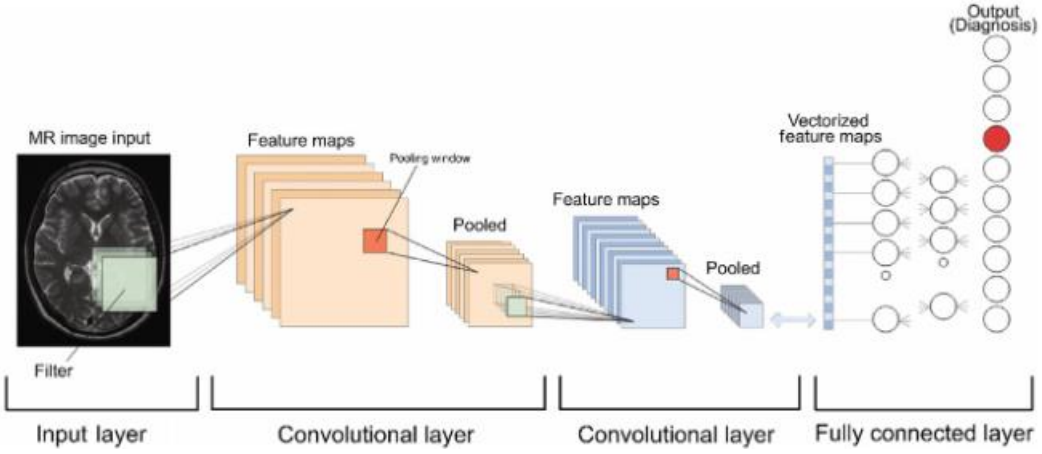

Figure 2.2 - The various stages of CNNs (Lundervold *et al.*, 2019).

Besides the topics already exposed, which are typically the most associated ones with AI and that have contributed the most to the evolution and public enthusiasm for the subject, there are also other topics worthy of mention when talking about AI, namely NLP, and AI for optimization problems.

The study of how computers can comprehend and use natural language text or speech is known as NLP. NLP researchers work to learn more about how people interpret and utilize language so that the right tools and methods may be created to help computers comprehend and manipulate natural languages to carry out the necessary tasks. Some applications of this technology have brought us sentiment analysis, chatbots and virtual assistants.

On the other hand, when using AI to solve optimization problems, a diverse landscape of problems, *e.g.,* inventory optimization, where one intends to minimize stock in excess, while minimizing the risk of running out of stock and losing sales, can be tackled. To solve them, Khamis (2023) mentions several algorithm "families" that can be used – deterministic search algorithms, trajectory-based algorithms, evolutionary computing algorithms, swarm intelligence algorithms and ML-based algorithms. The These problems are frequently combinatorial, so the solution aims to find the most beneficial choice for a set of mandatory goals or desired constraints. Some other techniques used for optimization problems are fuzzy logic, where AI mimics human reasoning in terms of linguistic variables and is built on straightforward IF-THEN rules (Cuevas *et al.*, 2020), constraint programming (CP), where constraints are used to reduce the set of values that each variable can take, while a search space keeps getting pruned, and anomaly detection approaches, that are commonly used to expose inefficiencies,

disclosing opportunities for optimization *e.g.*, traffic management systems leverage these to identify accidents or congestions, thus managing traffic flow.

2.2.3. Artificial Intelligence in Healthcare

Several authors have stated how AI can bring advantages to HC, by helping to solve indicators related to human error, and how there is motivation to enhance AI technologies as part of the solution. For instance, Graber *et al.* stated that "cognitive factors" were thought to be responsible for about 75% of diagnostic errors. These "factors", according to the author's investigation, included anchoring bias (staying with an initial impression), framing bias or faulty context generation (over-reliance on the specific way in which a question is posed), availability bias (tendency to draw conclusions based on recent events), satisfaction of search (not considering other options once a likely answer is found), and premature closure (accepting an answer before it has been verified). Not only have these errors contributed to rising HC costs, with misdiagnosed claims for malpractice estimated to be 300 000 USD per claim (Vinod Khosla, 2013), but they are also responsible for the perish of about 40 500 patients in intensive care units each year in the United States (Winters *et al.*, 2012).

Despite these (and other) challenges, the same authors point out various opportunities for the success of AI in the HC sector, namely the fact that AI solutions and algorithms cannot make the same mistakes as humans do, so they can act in a complementary way to diagnoses provided by physicians. Additionally, there has been a confidence increase from society when it comes to the "availability, sophistication and trust in computer expert systems". Moreover, there is great potential for increasing accuracy, productivity and efficiency in the sector, as there has been a reported discouragement from clinical staff when it comes to the amount of time spent with patients (Dugdale *et al*., 1999). Finally, AI also comes as a way for physicians to spend less time in administrative chores and use these cutting-edge computing technologies.

In the next three sections, the main cores where AI has been applied to the HC industry will be exposed. The intention is to understand if AI outcomes and techniques have improved processes in practice, and not only theoretically, as seen until this point. Later, with this knowledge, it will be possible to address more realistically how an AI solution is held when producing the framework. Throughout these three processes, a case study for each HC thematic will be chosen amongst the examples given so the ethical concerns and practices in each of these case studies can be dissected.

2.2.3.1. AI in Healthcare Administration

Enhancing service delivery and operational effectiveness has always been a priority since these are identified as "weak spots" for many HC institutions. AI applications include processing and analysing enormous quantities of information in the form of clinical notes, managing the supply chain and providing diagnostic assistance, among other things. There are several examples of what concerns the application of AI solutions to the improvement of the usual time-consuming tasks that characterize the administrative duties in the HC sector.

So far, researchers have applied AI to enhance HC administration via data warehousing and cloud computing, quality improvement, cost reduction, resource utilization, or patient management (Islam

et al., 2018). Focusing on the first topic, examples of research include the development of a clinical data warehouse and analytical tools for traditional Chinese medicine (Zhou *et al.,* 2010) or the creation of a large data repository and knowledge discovery using unsupervised learning (Mullins *et al.*, 2006). HC cost, quality and resource utilization have been addressed using classification and clustering algorithms to find that, having used medical claim data from 800 000 people from an insurance company, the absolute prediction error improved by over 16%. This was achieved by Bertsimas *et al.* (2008), with two prediction models, where both used medical information, and only one also used cost information. While inferring HC costs, both achieved similar accuracy scores, and both models surpassed traditional regression models while proving, in this case, that including medical information did not improve the cost prediction accuracy score.

Specifically to optimise resources in two Portuguese hospitals, an automatic scheduling solution developed by Glintt[7] (2022a; 2022b) was implemented with the aim of transforming the surgical scheduling paradigm by introducing AI into the decision-making process, valuing it as a critical element in the value chain and as an enabler of better HC. Powered by CP, a mathematical model where decision variables and constraints represent the problem to be modelled, the solution aims to achieve three main goals. Firstly, an automatic scheduling, given that an optimal solution for the intended timeframe is found among all possible combinations of surgeries on the provided waiting list. Secondly, resource optimisation, namely block rooms. Finally, a "blind" scheduling, which is crucial so ensure the suggested schedule follows the defined restrictions in a transparent way. This particular example will be further explored in case study I.

Another example is the efforts to predict the length of stay in a hospital. To do this, Hachesu *et al.* (2013), predicted this indicator for patients with coronary artery disease using SVMs, NNs, DTs and an Ensemble algorithm that combined SVM, NN and a DT algorithm – C5.0. The results showed that SVM led to the highest accuracy score – 96.4% - while showing that anticoagulant drugs, nitrate drugs, and diagnosis were the top three predictors for this context. When it comes to patient management, Koskela *et al.* (2010) identified risk factors such as "high body mass index, alcohol abstinence, irritable bowel syndrome, low patient satisfaction, and fear of death" for medical care "frequent attenders" using Bayesian classification techniques, which is particularly important since costs and resource utilization can be reduced as these are the patients that represent a large percentage of clinical workload.

EMRs or electronic medical records have been widely used to employ AI in HC, mainly using NLP. Some of its applications include extracting information, converting unstructured data into structured data, or categorizing data and documents. More specifically, NLP has been used to classify diseases based on medical notes and standardized codes, namely the ICD[8]. This system, managed by the World Health Organization, includes codes for ailments and their signs, different observations, situations, and disease causes. Figure 2.3 exemplifies extracting and identifying the ICD code from a description of clinical guidelines using NLP techniques. By searching for pertinent clauses in unstructured language and classifying ICD-10 codes according to how frequently they occur, unstructured text is turned into structured data. The NLP algorithm is performed at multiple thresholds to increase classification

[7] Portuguese technological company focused on providing solutions for HC institutions.

[8] International Statistical Classification of Diseases and Related Health Problems.

accuracy, and the data is aggregated for the output, according to Rangasamy *et al.* (2018). ICDS4IM[9] is another example of a similar system, developed in Portugal by Universidade do Minho, that translates clinical notes (natural language/narratives) into valuable data for analytics and aims to "extend the actual state of the art to support clinical decision-making" (Peixoto *et al.*, 2020).

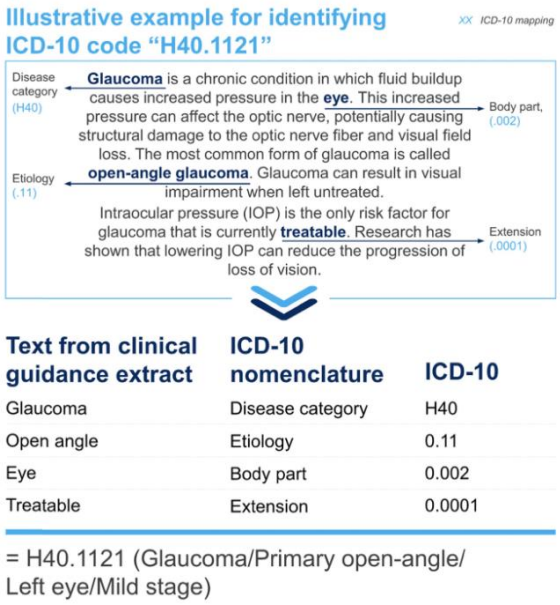

Figure 2.3 - ICD-10 mapping using NLP (Rangasamy *et al*., 2018)

2.2.3.2. AI in Clinical Decision-Support

In the event of rare or challenging-to-diagnose illnesses, timely, correct discovery significantly impacts on treatment outcomes. Medical professionals are unfortunately limited in the quantity of images and materials they can study, and their diagnoses are susceptible to human error. AI, on the other hand, can quickly and reliably process millions of samples. The search for support in this field arose in the 1970s, with a large wave of the so-called first generation of AI in medicine (Kulikowski, 2015), including Shortliffe's work with MYCIN (Shortliffe *et al*., 1975), Kulikowski's individualized clinical decision models (Sonnenberg *et al.*, 1994), and de Dombal's computer-aided diagnosis of acute abdominal pain (de Dombal *et al*., 1972). These systems included newer statistical reasoning techniques like probabilistic reasoning and NNs after CV's basic foundation was statistical signal processing.

When mentioning CV, there is enough research around leveraging it to support decision-making in the HC sector through imaging or video analysis. Strokes are an example of a condition that can benefit significantly from coopering with AI, as researchers like Rehme *et al*. (2015), Griffis *et al*. (2016), and Kamnitsas *et al*. (2016) proved. Strokes affect more than 500 million people worldwide and is a common and regularly occurring disease that is 85% of the time caused by a thrombus in the vessel called cerebral infarction (Jiang *et al*., 2017). In North America, it has ranked fifth and is recorded as the biggest cause of death in China (Heeley *et al.,* 2009; Saenger *et al.*, 2010). Stroke-related medical costs were around 689 billion USD globally, significantly straining on nations and families (Jiang *et al.*,

[9] ICDS4IM's Project Sheet can be found at https://algoritmi.uminho.pt/projects/icds4im-intelligent-clinical-decision-support-for-intensive-medicine/.

2017). Neuroimaging methods, such as medical resonance imaging (MRI) and computerized tomography (CT) scans, are crucial for this disease assessment. To help in stroke diagnosis, some studies, such as the following ones, have attempted to use ML techniques to neuroimaging data. Endophenotypes of motor impairment after stroke were recognized and categorized by Rehme *et al.* (2015) using resting-state functional MRI data and SVMs, which classified stroke patients with an accuracy of 87.6%. To identify stroke lesions in T1-weighted MRI, Griffis *et al.* (2016) utilized the Gaussian NB Classification, whose outcome is comparable to manual lesion delineation performed by human experts. In a multimodel brain MRI, Kamnitsas *et al.* (2016) attempted to segment lesions using three-dimensional CNNs. This particular example will be further explored in case study II.

Although NNs, and particularly DL, are currently the first choice to create CV algorithms for categorizing images of certain disorders like lesions in the skin or other tissues, Bohr and Memarzadeh (2020) demonstrated that video data is projected to include 25 times as much data as high-resolution diagnostic images like CT scans and may therefore offer a higher data value based on resolution over time. Although it is still early, video analysis provides a lot of potential for clinical decision assistance. As an illustration, real-time video analysis of a laparoscopic surgery yielded 92.8% accuracy in identifying each step and, interestingly, the ability to spot a missed or unexpected step, as Volkov *et al.* demonstrated in 2017, and the following figure illustrates.

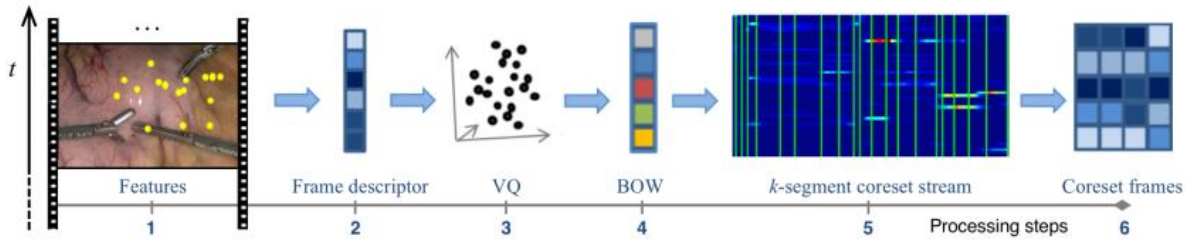

Figure 2.4 - Volkov *et al.* (2017) system overview. (Left to Right) Video Stream; Feature Extraction from video frames; Descriptor Representation; Vector Quantization; Bag-of-Words Representation; k-segment coreset reduction; Coreset Frames, as input to the system.

2.2.3.3. AI in Logistics Optimization

The allocation and use of logistics must be optimized for efficient HC management. One of the main reasons lies in swiftly increasing HC costs in developed countries, and logistics has been identified as one of the significant cost factors (Rais *et al.,* 2018). As already stated in this document, efficiency has been a frequent problem in several countries and health institutions. As such, the use of AI in this subject has been taken on the agenda of multiple researchers, to enhance results, applying mostly operational research techniques for optimizing processes.

AI as a means of HC logistics optimization was also used by Umoren *et al.*, 2021. Resorting to a hospital in a developing country – and facing the adjacent scarce resources and monetary means - the goal was to improve two leading indicators: quality of experience and quality of care. A framework capable of assessing and optimizing HC logistics was utilized to demonstrate how to provide adequate supply chain management. Type-1 fuzzy logic model was employed to optimize multiple objectives, with coefficients given statistically proven weights. To achieve this, in the first phase, an HC resource allocation plan was explored; a resource use schedule by patient class for a daily operational level was

established in the second phase; and in the third phase, the framework was created. In the end, additional research was conducted to comprehend the consequences of various tactics and how they interacted to determine the ideal resource supply chain.

In Portugal, KnowLogis[10], a solution by Glintt in collaboration with INESCTEC - an internationally-oriented multidisciplinary associate laboratory for R&D and technology transfer - was created to ease decision-making in hospital logistics, through the development of an intelligent reporting system comprising advanced AI software and a user-friendly and handheld dashboard (Figure 2.5) that, in integration with the databases of the current systems, actively and dynamically monitors and tracks the costs of hospital logistics products, automatically analyses the evolution of its stocks, incorporates historical data and suggests corrective and improvement measures to warehouse and inventory management. This way, it provides value in the logistical and financial sector by predicting needs, monitoring, and coordinating expenses with medications, medical equipment, and supplies through budget monitoring, modification, purchase coordination, and by suggesting supply policies (such as cycle and safety stock, designated space, replenishment cycle, etc.). With the challenge of taking logistics from a deterministic to a dynamic perspective, an intelligence and advanced analytics engine is based, among other indicators, on the classification of the various hospital items according to their consumption profiles – erratic, lumpy, smooth, and slow/intermittent, according to the Syntetos *et al.* (2005) classification scheme, based on the coefficient of variation of demand and the average interval between consumptions -, and analyses typically used in the logistics industry (ABC[11], FSN[12]). Knowing these insights in a user-friendly way has shown to improve operational efficiency while increasing the level of service. The results of the implementation of this pilot project in a D group hospital, according to the practised classification by ACSS[13] (2017), since February 2020 showed a reduction of the average inventory value by 10%, an improvement of 2 percentage points in the service level, and a reduction of the time spent in the whole process, from the detection of a need to the placement of an order by at least 20% (Oliveira, 2020).

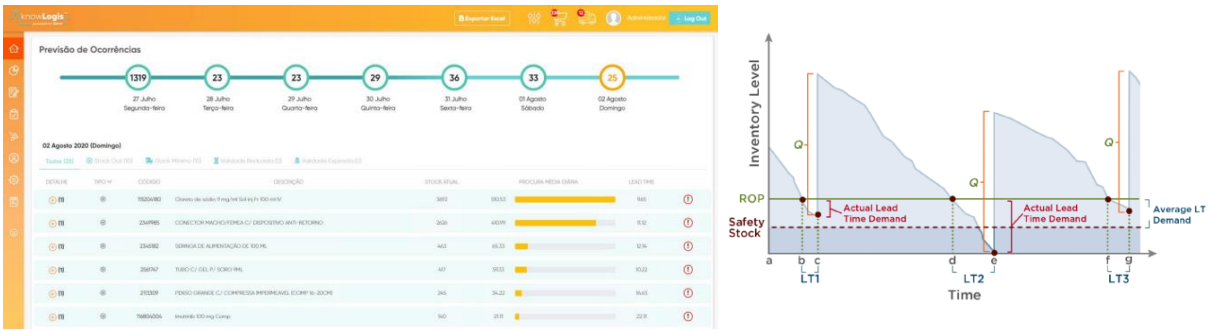

Figure 2.5 – Knowlogis' user-friendly interface, on the left. Reorder point (ROP) calculation, based on inventory level, safety stock, average demand, and time, on the right (Glintt, 2020).

[10] KnowLogis—GLINTT | INOV, available at https://inovglintt.com/projetos/knowlogis/.

[11] ABC analysis is a method of inventory classification that divides the products into three groups – A, B and C -, according to their revenue. In this analysis, "A" stands for the most important inventory, "B" for moderately necessary inventory, and "C" for the least important inventory.

[12] FSN analysis is used to categorize the products into fast-moving, slow-moving, and non-moving categories in the inventory control approach, based on their amount, rate of consumption, and rate of inventory use.

[13] ACSS, Administração Central do Sistema de Saúde, IP, available at https://www.acss.min-saude.pt/.

One last example that will be further explored in case study III is the Data2Help[14] project. With the ultimate goal of improving operational results by optimizing the allocation of resources, the Data2Help project—created jointly by INEM[15], IST[16], and INESC-ID[17] aims to give INEM new tools to optimize the allocation of resources, resulting in a better and quicker response to medical emergencies in mainland Portugal, according to INESC-ID (2019). To this goal, the project developed by Manquinho *et al.* (2022) is supported by three key pillars: Develop predictive models that use several anomaly detection approaches (*e.g*. SARIMA; LSTM) for emergency vehicle requests in each geographic area through the analysis of a large repository of integrated and historical data; integrate INEM's information systems with other pertinent external data (such as meteorology, epidemics, demography, and forest fires); and optimize the allocation of INEM's resources based on the predictive models to improve the response times to medical emergencies.

2.2.4. Artificial Intelligence Lifecycle: Components Overview and Stakeholders in HC

At this point, it becomes significant to explore further how the AI lifecycle has been described in the literature and who the stakeholders are when it comes to the HC industry. Looking at the schematization developed by Wirtz *et al.* (2020) and adapted by Reis *et al.* (2020) to provide a framework for AI's implementation (Figure 2.6), it is possible to locate AI Ethics as part of the process and composed by the cycles of AI Applications and Technology (1), AI Challenges (2) and Collaborative AI Governance (3).

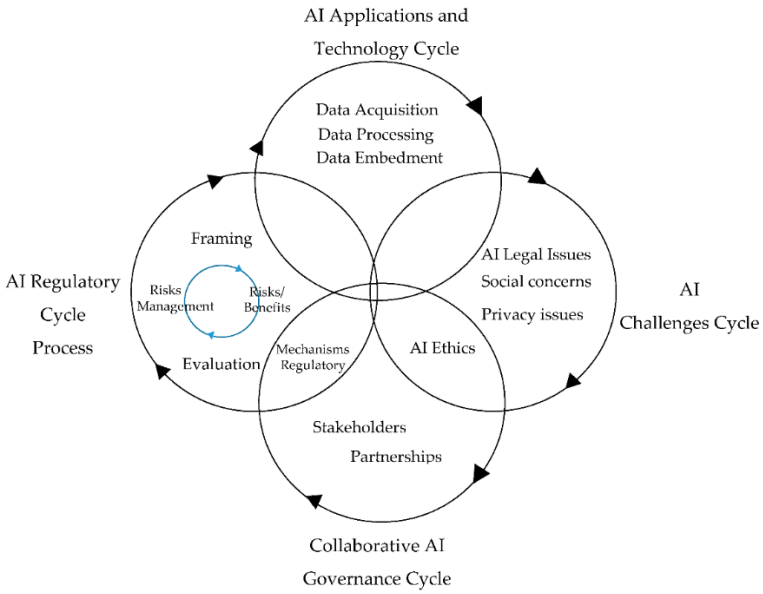

Figure 2.6 - AI's Implementation Framework, developed by Wirtz *et al.* (2020) and adapted by Reis *et al.* (2020).

Starting with cycle (1), the author implies almost every process related to data capture, process and treatment. With the same opinion as Reis *et al.* (2020) that this cycle should be broad and

[14] Data2Help, available at https://cegist.tecnico.ulisboa.pt/~cegist.daemon/projects/data2help-data-science-optimization-emergency-medical-services.
[15] Instituto Nacional de Emergência Médica
[16] Instituto Superior Técnico
[17] Instituto de Engenharia de Sistemas e Computadores – Investigação e Desenvolvimento

comprehensive, De Silva *et al.* (2022) divide nineteen stages into design, development and implementation phases, all concerning different profiles. On the other hand, Char *et al*. (2020), focusing on HC, mentions phases related to developing and implementing a solution as part of this same cycle. These comprehend stages as the conception, development and calibration of ML for HC applications. With the same focus on industry and application type, Conjeti (2023) divides the development and implementation of AI devices in HC with resources into three cycles – Software Development (Dev cycle from now on), ML and Operations. While the Dev cycle is mainly about identifying stakeholder requirements, developing a clinical concept, and translating stakeholder requirements into software requirements to create AI systems tailored for HC challenges, the ML cycle involves deriving algorithm performance requirements, establishing guidelines for accurate annotations, curating datasets, developing tailored AI algorithms, and verifying their performance against acceptance criteria before integrating them into HC systems. These two cycles communicate simultaneously, leveraging the findings of the ML cycle to return to the Dev cycle to perform quality validations, namely completing clinical performance evaluations and standalone testing on independent data and obtaining regulatory approval. Finally, the third and last cycle is about deploying cleared medical devices, developing marketing strategies, providing training and support, ensuring interoperability, offering customer support, monitoring performance, conducting post-market surveillance, and gathering clinical evidence to ensure the successful integration, adoption, and ongoing improvement of AI medical devices in HC settings. The Dev cycle is used again at this stage, where feedback is received for continual improvement, and product strategy and roadmaps are aligned.

Setting now attention on the common space between cycles (2) and (3), AI Ethics has its space in the AI lifecycle, according to Wirtz *et al.* (2020) and Reis *et al.* (2020). Ethics, and in particular AI Ethics, will be further discussed during this chapter. However, to enhance its understanding, it's crucial to recognize who are the stakeholders particularly in the HC industry. Paul *et al.* (2018) propose the division of the various stakeholders that "make up the HC ecosystem and work together towards the successful adoption and implementation of AI in HC" into five groups – practitioners, developers, research and industry bodies, government, and funders and investors. Although the described ecosystem is for the HC industry in India, the categories indicated seem to be interesting for more regions of the world, since their compositions are compatible with what occurs in Europe, for instance. In what concerns "practitioners", hospital, and other clinical staff are englobed in this category. For "developers", entrepreneurial developers, commonly part of companies ranging from startups to bigger corporations with international or domestic expansion, are included, and just like "research and industry bodies", these have already been mentioned in this work, and both can include computer scientists, and engineers, for instance. Professors, researchers, scientists, and firms that, for example, publish studies regarding AI and HC are examples of the last-mentioned category. When it comes to "government", every state company whose mission is to bring evolution to this field or governments who approve initiatives, laws or guidelines regarding the use or development of AI, as exposed already in this document, are part of this category. Finally, for "funders and investors", the most common examples are investors or venture capital firms that fund startups for the development of AI in the HC industry. The government and private companies can also belong to this category when acting as investors for endeavours with these thematises *e.g.*, NeuroPsycad is a startup whose mission is to be the provider of precision diagnostics for neurodegenerative diseases and has Philips HealthWorks as one of its investors (Tracxn, 2023). Char *et al*. (2020) add to the already mentioned stakeholders,

patients (or receivers, users) of HC involving AI solutions in any phase of the care provided, and oversight bodies with regulating medical practice.

2.2.5. Artificial Intelligence and Future Trends

UC Berkley believes that an increasing amount of data will continue to be produced as the majority of societal and industrial sectors continue to go digital. In their blog (2020), they state that one key to solving a wide range of problems is the ability to draw insights from these enormous datasets, and ML and AI are being used for this in almost every major industry, including business, government, agriculture, transportation, cybersecurity, marketing, and HC is no exception. For illustration purposes, IBM used the Watson system to browse through medical research material after winning Jeopardy in 2011, effectively "sending Watson to medical school" with the goal of Watson becoming a “very smart assistant” that will be able to aid physicians with the “rapid pace of incoming new research” (Lohr, 2012).

Yet, in the last year, an “AI boom” without precedents has come to change every person’s and business’s perspective on the usage of AI. With foundational models, more specifically Large Language Models (LLMs), a model called ChatGPT, that interacts conversationally, was trained and made available freely and on a worldwide scale by OpenAI[18]. OpenAI is a “research and implementation company” whose mission is “to ensure that artificial general intelligence benefits all of humanity” (OpenAI, 2023). Knowing this, making an AI model available for all was a no-brainer for the company. However, this new tool’s impact was perhaps unmeasurable, even for OpenAI, whose pillar is safety and responsibility when it comes to creating AI.

Generative AI has unique risks and enormous potential, namely hallucinations, bias, consent and security (IBM Technology, 2023). Due to this, now and in the future, responsible, explainable and accountable AI is becoming a major topic of both interest and investment, as executives expect to invest at least 40% more in AI ethics over the next three years and 80% of business leaders see at least one ethical issue amongst explainability, ethics, bias and trust as a significant concern, according to the same author’s Institute for Business Value (2023). Being so, despite some use-cases that are already applied in the “real world” - *e.g.*, Moderna[19]’s usage of proprietary LLMs for mRNA drug discovery (Pearson, 2023) -, the efforts on regulating models such as ChatGPT to guarantee that AI research (in Europe) is ethical and focused on people, has also become visible (Browne, 2023), *e.g.*, the EU AI Act[20], once approved, will be the first AI-related regulation in the world (Guillot, 2023). Due to all this, it is impossible to mention any topic of AI without addressing ethical considerations, as this will be key to ensuring an “AI winter” isn’t in the future for this technology.

Having at this point provided a broad but relevant technical background for the work held, it becomes essential to deep dive into what ethics is, and its presence when it comes to the communion with AI for developers and users. Being so, the most mentioned topics in the literature will be extracted and analysed. Afterwards, the state-of-the-art approaches for implementing AI in an ethical way will be scrutinised.

[18] OpenAI, available at https://openai.com/.
[19] Moderna is a pharmaceutical and biotechnology company focused on RNA therapeutics, primarily mRNA vaccines, available at https://www.modernatx.com/.
[20] EU AI Act Proposal, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

2.3. ETHICS

The Cambridge Dictionary (2023) defines Ethics as “a system of accepted beliefs that control behaviour, especially such a system based on morals” and as “the study of what is morally right and what is not”. The fundamental questions of practical decision-making make up the study of ethics, and its main concerns are the nature of ultimate value and the criteria by which human activities can be classified as right or bad. Ethics is, therefore, the same as moral philosophy (Singer, 2023).

Normative ethics – understood as the core of general ethics, elaborates and examines universally valid norms and values, as well as their justification (Natrup, 2022) -, metaethics - focused on exploring what morality itself is (DeLapp, n.d.) -, and applied ethics are the subfields of philosophical ethics (Marturano, 2002), with the latter one being the most pertinent to this dissertation. To better grasp certain situations that require ethical consideration and frequently take prescriptive stances, applied ethics leverages the concepts and discourses from moral philosophy, although it often resources to the other two subfields for that purpose.

2.3.1. Ethics in Artificial Intelligence

AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in developing and using AI technologies (Leslie, 2019). The need to dispute indicators like those found in the 2023 AI Index Report[21] that claim the number of incidents involving the misuse of AI is rapidly rising led to the quick emergence of this topic. According to the AIAAIC[22] database, the number of AI incidents and controversies has increased 26 times since 2012, which validates both an increased use of AI technologies and knowledge of potential abuse. Directly related to this, and according to an analysis of the legislative histories of 127 countries, the number of bills containing "artificial intelligence" that were passed into law increased from 1 in 2016 to 37 in 2022, openly indicating a rise in policymakers' interest in the technology (Perrault *et al.*, 2023).

Alongside AI ethics, the expression "ethical use of computing" refers to a computing practice that is morally upright, respectable, or socially acceptable. However, as stated and supported by Stahl *et al*. (2016), such intuition is typically based on more or less explicit norms and values that are recognized within a social group or culture. Quoting the same author, "not all computing professionals have a deep intrinsic interest in understanding the details of ethics". Even though it may not be immediately clear how specific technological judgments in this setting might be seen from an ethical perspective, many practical debates and conclusions are nonetheless guided by ethical notions and values. An example of such occurs often when handling Big Data. Large datasets can be employed for a vast array of applications that promise significant advantages, namely for training models for decision-making in CV for HC, as an earlier chapter exposes. Simultaneously, they may raise serious concerns regarding ideas like privacy and ownership. Finding the ethical difficulties in this situation is a demanding task in itself, and connecting them to technical decisions that strike the optimal balance between competing interests and values is equally challenging. Such conundrums might impose the use of ethical principles

[21] 2023 AI Index Report is available at https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

[22] AIAAIC database is an independent, non-partisan, public interest initiative that tracks incidents related to the ethical misuse of AI for algorithmic, automation transparency and openness.

like accountability, virtue, or utility when they are made explicit. For these reasons, Stahl *et al*. (2016) believe that a certain amount of philosophical ethics knowledge is necessary to understand the complexity of issues such as these. As such, the Stanford University HAI[23] uncovers how a rise in interest in AI ethics and associated research occurred from 2021 to the past year since submissions to FAccT[24] increased by a factor of two and by a factor of ten from 2018, as the following figure illustrates. Interestingly enough, although academic institutions still dominate FAccT, industry actors have recently produced more work than ever before in this field, displaying the interest of several stakeholders in grasping the best ways of deploying AI and showcasing once again the relevance of the RQ in hands.

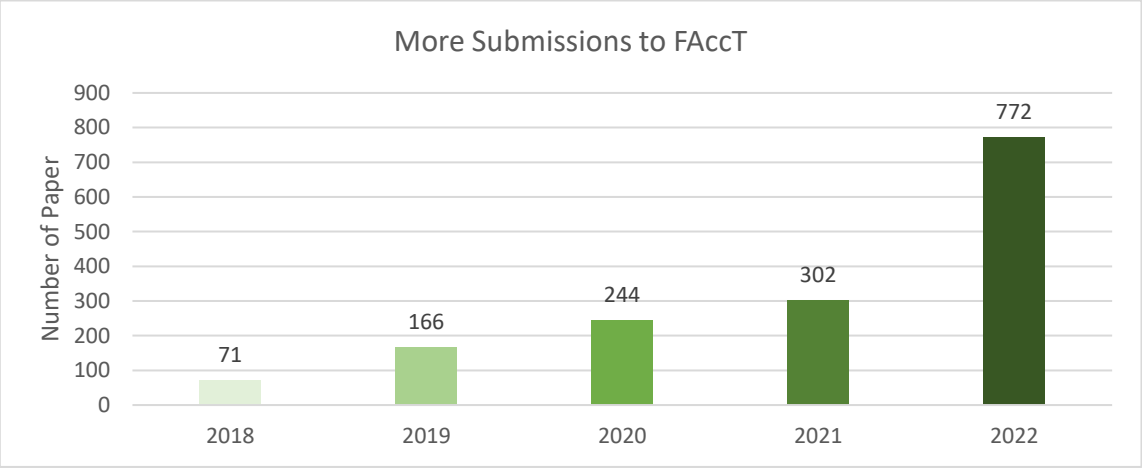

Figure 2.7 - There has been an increasing interest in incorporating ethical topics when working with AI, according to Stanford's HAI report on the 2023 State of AI (Lynch, 2023).

When addressing the sector of HC in the light of ethics topics, many authors reflect on their concerns. Examples are related with matters as inclusivity, discrimination, and bias (*e.g.,* Pfohl *et al.*, 2021; Naik *et al.,* 2022; Chen *et al.,* 2019; Obermeyer *et al.,* 2019; WHO, 2021), explainability of the models that produce what may influence a medical decision (*e.g*., van de Sande *et al*., 2022; Vogel, 2017; Diakopoulos, n.d.; IBM, 2023; UNESCO, 2021), accountability for the results (*e.g*., Vayena *et al*., 2018; AI Now Institute, 2023; Hagendorff, 2020; Deloitte, 2022; Char *et al.*, 2020), human's autonomy towards these models and a value up-holding use of these technologies (Tai, 2020; Paul *et al.*, 2018; Ouchchy *et al.*, 2020; Cordeiro, 2021; Topol, 2019; Rigby, 2019). Besides these, the particularities of the HC field lead authors also to show concern regarding multiple other topics, namely what the future of patient interaction will be (Sujan *et al.*, 2019; Whitby, n.d.; Cordeiro, 2021), and what steps need to be taken regarding the "education of an AI-literate workforce" (He *et al.*, 2019; Council of Europe, 2019; Böke, 2020; House of Lords' Select AI Committee, 2018). With this wide variety of concerns, a condensation of 43 sources regarding ethics in HC and in AI applications (see Appendix chapter) was executed to understand the most relevant topics within the available literature, which are represented in the following figure.

[23] Stanford University Human-Centered Artificial Intelligence, available at https://hai.stanford.edu/.
[24] Conference on Fairness, Accountability, and Transparency.

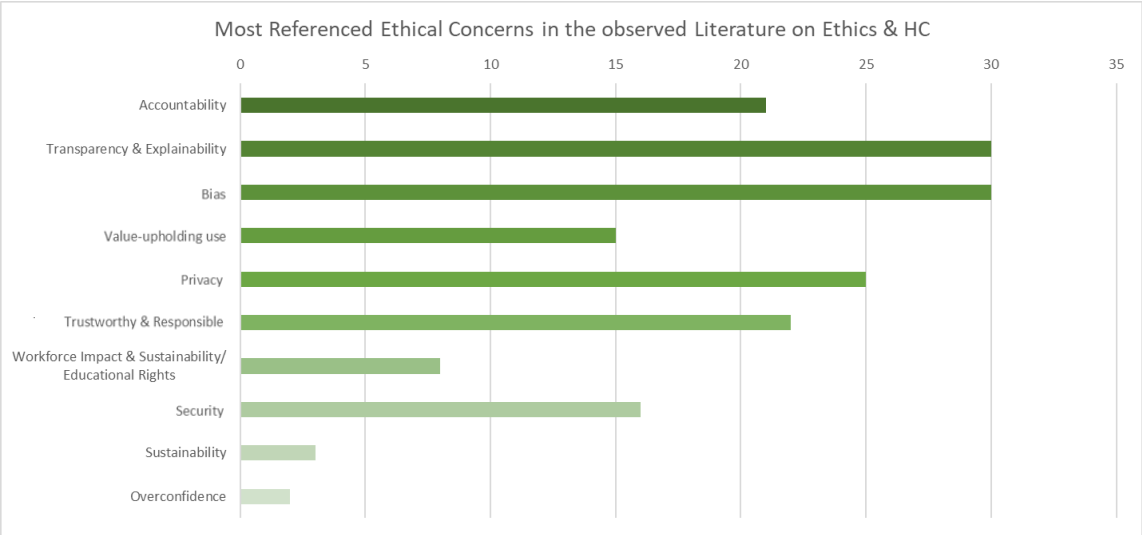

Figure 2.8 - Frequency of topics identified in the literature related to AI ethics in the HC sector.

Its examination shows that the five most referenced ethical worries from several stakeholders *e.g.*, HC providers, HC receivers, health organizations and institutions, are (1) the transparency, interpretability and explainability that any AI artefact employed must bring for its users, opposing to the "black box" vision of AI; (2) the bias in which it might occur; (3) the possibility of privacy and informed consent when utilizing AI solutions; (4) a responsible artefact that allows users to build a feeling of trust towards its outputs; and, finally, (5) the definition of who is the accountable party for the AI solution and its outputs.

Other referred topics are related to the value-upholding use of these technologies for the human, preserving one's autonomy, which can be associated with the centralization of the user, or the recipient, keeping the "human (…) "in the loop" for all targeting decisions" (Kaplan, 2016); the worries linked to what might be the impact in the HC workforce, and if these professionals have a sustainable future given the rise of AI, coexisting with it; the robustness of such solutions for patients and, linked to it, the security one can expect from AI implementations. Finally, in a residual number, there are also mentions of environmental sustainability and the overconfidence AI can bring. Concerning the first topic, Stanford's HAI (2023) elucidates how training a model comes with high environmental costs due to its considerable carbon emissions. As for overconfidence, the developmental psychologist from Berkeley, Kidd, exposed during the 2019 NeurIPS conference how algorithms can influence human beliefs, claiming that humans continuously form beliefs and that certainty diminishes interest. That said, feedback drives certainty, and less feedback may encourage overconfidence. All this culminates in humans creating beliefs quickly. Kidd advocates that these traits should be part of everyone's knowledge of ML.

Deep diving into each one of the five most referenced ethical concerns, bias is "as old as human civilization", citing Google's Chief Scientist for AI, F. Li (Vanian, 2018). It is "human nature for members of the dominant majority to be oblivious to the experiences of other groups", and it can easily be a way of exacerbating inequalities and discrimination overall. Bias can be present in distinct forms and origins: when the bias occurs unconsciously – implicit bias - or when it occurs consciously – explicit bias - it has become a motto for ethical but also technical and scientific research (Bender *et al.*, 2018). Five types of bias can be encountered concerning the setting of AI applications: association bias,

automation bias, confirmation bias, dataset bias and interaction bias. When training data for an AI solution shows a bias that is not based on causal effects, association bias has occurred. For instance, larger average male salaries may not necessarily indicate performance. On the other hand, automation bias happens when human control over semi-autonomous systems is limited, leading to inaccurate or unfavourable results (Skirpan *et al.*, 2017), or in the HC context when a state of complacency that sets in when an AI program takes over a task that was formerly performed by an HC professional (Anderson, 2019). When information that supports preexisting views or biases is selectively favoured, one is in the presence of confirmation bias. Search engines and recommendation systems that are based on user profiles frequently exhibit this type of bias (Chou *et al.*, 2017). On the other hand, when an AI system's dataset fails to represent a specific demographic accurately, it is said to have a bias against that population, and there is a dataset bias (*e.g.*, gender, sexuality, age, education misrepresentation). Internet data, for instance, is not gender-neutral since women are underrepresented or contribute to material in different ways than men, which can lead to an algorithm that may conclude that women are less capable or eager to contribute because of such misrepresentation. Finally, interaction bias reflects situations where an AI system learns from data on human conversation and deduces corresponding patterns. An example of this is ChatGPT that, citing OpenAI's CEO, "has shortcomings around bias" (Altman, 2023). Regarding bias presence in HC applications, AI-based models may amplify pre-existing human bias within datasets (Cho, 2021). Seyyed-Kalantari *et al.* (2021) and Cho (2021) exposed how underdiagnosis occurred in underserved patient populations, with a higher underdiagnosis rate in intersectional underserved subpopulations (*e.g.*, Hispanic female patients). Both studies highlight the issue of an AI algorithm's potential delay in providing care when it incorrectly labels a person with a condition as healthy.

When addressing the concepts of transparency and explainability, both are viewed as quality requirements for deploying AI systems (Balasubramaniam *et al.*, 2022). The EC (2019) states that "technological transparency implies that AI systems are auditable, comprehensible and intelligible by human beings at varying levels of comprehension and expertise". From the research of Felzmann *et al.* (2020), transparency can be understood from different perspectives – as a virtue, a relation, and a system. In the RQ's context, all three come together as transparency, defined as the persistent openness regarding one's operations, behaviour, intentions, or considerations, and is regarded as an innately important attribute of agents, systems, or organizations. Transparency, however, cannot be understood independently of the relationship between an agent and a recipient and is the central tenet of a systemic perspective, which takes into account the institutional context of the relationships of transparency and makes it possible to have a practical understanding of its effects and implement it successfully. In practical terms, building trust between businesses and the users of their AI products is facilitated by transparency. It is also important to notice how, based on the same authors, transparency can refer to explainability, interpretability or interpretable AI, openness, accessibility, and visibility. The absence of any of these concepts is commonly known as the "black box issue" (Anderson, 2019). Assuming the example of an AI program's higher reported success rate for spotting cancer cells compared with a human success rate, resorting to the said AI program would be supported to at least augment human pathologists' identification of cancer cells, according to the same author. It could even be argued that using the AI program could help train pathologists since it has a higher success rate than the human eye alone at this stage. However, the "black box issue" addresses how the professional doesn't know how the AI program spots cancer cells. Nevertheless, assuming that the program is used to augment rather than replace the work of a pathologist, a pathologist could become

more knowledgeable about which cells are cancerous with the presence of more transparency in the given program, given that the presence of this enhanced ethical feature would, according to this author, illuminate the "black box", and the achieved explainability could enable pathologists and patients to feel more comfortable relying on the AI program.

Privacy is a significant issue impacting how AI is developed and evaluated (Stanfill *et al.*, 2019). Its principal concerns lay in how the data used in AI systems is collected, stored and used in a way that respects individual privacy rights, which is deeply related to the legislation employed in the distinct geographies. The most significant change in data protection regulation is the EU General Data Protection Regulation (GDPR), which occurred on May 25th, 2018 (Wolford, 2018). In a nutshell, the GDPR comes as an essential player in assuring privacy by addressing issues of processing personal data in the light of globalization and technological development while guaranteeing the protection of fundamental rights and freedoms for EU citizens and harmonizing data protection laws across Europe. Parallel to this, one can't discuss the prospect of privacy without considering the security or robustness of any solution since organizations must ensure protection against unauthorized access, misuse, or data breaches. An example occurred in 2017, when the UK Information Commissioner's Office determined that the Royal Free NHS Foundation Trust had violated the UK Data Protection Act of 1998 by sharing the personal data of around 1.6 million patients with Google DeepMind (Gerke *et al.*, 2020). This data sharing occurred for testing the "Streams" app, designed to aid in diagnosing acute kidney injury. The breach occurred because patients were not adequately informed about how their data would be used during testing. This also brings to discussion the privacy-related topic of data ownership, which gains extra relevance in the HC field since health data can be worth billions of dollars (Gerke *et al.*, 2020). Once again, if patients and clinicians don't trust AI, its successful integration into clinical practice will ultimately fail.

When it comes to accountability, the unease relies on the possibility of error or inaccuracy that, particularly for the HC field, can have shattering effects on the patient who is the victim of the error because it is when one is most vulnerable that turns to the physician's guidance. In practical terms, this topic aims to clearly define which stakeholder is responsible for the AI solution's outputs, if necessary. Once again, the particularities of the HC field augment the worries and complexity of addressing topics as this, because technologists, unlike doctors, are not required by law to be held responsible for their conduct; instead, ethical standards of practice are used in this field, which sums the argument over whether technologists should be held liable if AI negatively impacts patients (Naik *et al.*, 2022). Therefore, if a clinician decides to use that data, they will not be able to properly defend their decisions if they cannot account for the output of the AI solution being used.

Finally, trustworthy or responsible AI means, according to the EC (2019), any AI solution that is lawful, ethical and robust, which can be achieved by including human agency and oversight, keeping the human-on-the-loop, and all the last four ethical topics already exposed. For Bracamonte (2019), trustworthiness comes above all from the transparency and interpretability a given solution with AI can bring, and for Ribeiro *et al.* (2016), if a model's outcomes can be understood, that knowledge can be used to assess the model's reliability. These authors also showed how explanations are helpful for a range of models in tasks involving trust in the text and image domains, with both expert and non-expert users: selecting amongst models, determining how trustworthy they are, enhancing untrustworthy models, and gaining insights into predictions. The AAAI 2021 Workshop: Trustworthy AI for Healthcare discussed how not many clinical AI solutions are currently implemented in hospitals

or actively used by doctors, although existing results are promising. According to the speakers, the reason lies with many current approaches making clinical decisions in a “black box” manner, making the decisions more opaque and difficult to understand. Security and privacy issues are raised because existing systems are not resistant to adversarial attacks or even minor disturbances. Additionally, current methodologies frequently favour particular ethnic groupings or subpopulations. For different ethnic groups or subpopulations, these biases could lead to unfair, less trustworthy forecasts. All these issues reduce the reliability of the current solutions. As a result, physicians are hesitant to employ these solutions because making clinical decisions requires high reliability and confidence. So, once again, trustworthiness is pointed out as a characteristic that comes from the combination of several ethical issues that have already been addressed.

2.3.2. Frameworks for Ethical Implementations of AI

Framework is defined as “a real or conceptual structure intended to serve as a support or guide for the building of something that expands the structure into something useful”, according to Lutkevich (2020). From the available literature, there are distinct ways of addressing the implementation of ethical concerns during the development of an AI solution. One can find four types of approaches. Each one is based on:

- questionnaires filled at different stages of a project development.
- scoring methods.
- guidelines designed to avoid ethical concerns.
- to the most appropriate phase of the project, ethical norms are told to be addressed by the developers.

Relative to the first approach, Diakopoulos *et al.* (n.d.) propose a Social Impact Statement for Algorithms that allows developers to use ethical principles as a guiding structure. It consists of several questions for each principle that should be answered and act together as a “public form of transparency”. The authors also suggest a revision of the said statement in distinct phases of the design and development phase. The developed work is based on the belief that "the algorithm did it" is not a suitable justification if people are building and creating algorithms and the data that powers them. However, there is no reported evidence related to the effectiveness of the work besides the theoretical indications.

Moving on to the second approach – scoring systems -, governmental organizations lead this. As seen before, governments have taken measures to bring some kind of legislation or guidance to AI developments, as it is the example of Canada and the UK. So, to promote the responsible use of AI by measuring the risks associated with automated decision systems, Canada developed an Algorithm Impact Assessment Tool (AIA) in collaboration with public institutions, academia and civil society (TBS OCIO[25], 2021). The AIA[26] should be completed twice: once at the beginning of the design phase of a project (to direct the mitigation and consultation requirements to be met during the implementation of the project), and once before the production of the system (to confirm that the results accurately reflect the system that was built). The tool is made up of questions in several formats to evaluate the

[25] The Office of the Chief Information Officer (OCIO) at the Treasury Board of Canada Secretariat (TBS).
[26] The AIA can be found at https://open.canada.ca/aia-eia-js/?lang=en.

risk in six different categories – Project; System; Algorithm; Decision; Impact; and Data - and to assess the mitigation measures to manage the identified hazards and evaluate effects in a variety of domains, such as individual or community rights; the physical or mental health of people or groups; the financial interests of people, organizations, or communities and the ecosystem's continued viability. The risk and mitigation scorings are obtained with one or more questions in each area where each question is given a value based on how much risk it either increases or decreases for the automation project. Depending on the obtained score, a level of impact is assigned. These are divided into four categories: *I* being the least impactful and *IV* being the most. To lessen the risks highlighted, a directive is also provided that specifies the mitigation actions needed for each of the four impact levels. Concerning the UK, the Data Ethics Framework is governed by three broad concepts, applicable throughout the entire process – Transparency; Accountability; Fairness -, and five specific activities, to provide practical reflections (UK CDDO[27], 2020). It attempts to aid in understanding ethical considerations, resolving these within projects, and encouraging responsible innovation with the intention of guiding proper and responsible data use in government and the larger public sector. Just like with AIA, there are several questions to guide through various ethical considerations concerning the project in hands. However, in this case, a low self-assessment score may mean the need for extra verification and possible adjustments to make the project more ethical. Each component of the framework is intended to be often reviewed during the project, particularly whenever any modifications are made to the procedures for gathering, storing, analysing, or sharing data. Another difference resides in the proof for evidence related to the effectiveness of these methods – while TBS OCIO makes available the implementation of the AIA for other players, the same wasn't found for the UK CDDO.

When it comes to an approach based on guidelines to avoid ethical harm, Prakash *et al.* (2022) propose a conceptual framework that includes a five-step approach for locating, evaluating, and putting into practice interventions to deal with the ethical and legal issues surrounding the use of AI in HC for the benefit of patients - The Pent'E Approach. Based on the string of thought that "AI presents ethical concerns that impede the progress of its usage in the HC field" and "(…) patients prefer empathetic humans to treat them rather than artificial systems", it was understood that "(…) an AI system, under the able supervision of HC professionals, has immense potential to bring about beneficial reforms in the HC system". In the study, five guiding principles – Evaluate; Enumerate; Engage; Enforce; Execute – made a framework ranging from identifying problems to implementing solutions, while addressing pressing AI ethics issues with the collaboration of stakeholders, policymakers, developers, and HC providers. On another work, Magrabi *et al.* (2019) defined indicators for monitoring AI-enabled Clinical Decision Support. Being so, four main guidelines were developed that aimed to underline the importance of understanding the context, goals and usefulness for the clinical practice and stakeholders of such applications, while mitigating possible ethical problems *e.g.*, it's crucial to guarantee enough coverage, specificity, and validity of the data for AI applications that are based on massive data sets, particularly genomic, biomarker, and phenotype data from across the health system. In parallel, assuring feedback from both experts and user groups takes an equal part in the monitoring efforts.

Concerning the last approach, there are also authors whose work aimed to encompass ethical concerns throughout the process of developing an AI solution. An example of such is the work developed by Char *et al.* (2020), which combined the first approach based on key questions along with ethical

[27] United Kingdom (UK) Central Digital and Data Office (CDDO).

considerations amongst the process. The model's objective is to help with systematic ethical evaluations of ML HC applications by outlining the conception, development, implementation, and concurrent evaluation and oversight tasks of these solutions on a map, and then layering key questions, value-based issues, and ethical considerations on top of this map. To comprehend and subsequently handle the ethical implications of ML HC applications, interdisciplinary engagement and dialogue of many stakeholders and their viewpoints and concerns is both possible and necessary. Another example was created by van de Sande *et al.* (2022). Unlike others, there is a focus on making it accessible to stakeholders without prior AI knowledge which is materialised by providing a step-by-step approach that includes all the necessary components and current best practices for implementation, assisting in bringing AI "from bytes to bedside". The end result is an organized overview of the development and implementation of clinical AI, with key steps within five phases and stakeholder groups (knowledge specialists, decision-makers, and users) provided. Alongside this model, significant actions are divided into the five phases to unify available regulations, obstacles, and best practices that are vital to the development of medical AI. It's worth mentioning that there is no reported evidence related to the effectiveness of the work, for both examples, besides the theoretical indications.

Having presented the state-of-the-art research concerning the goal of this study – to produce a framework designed for an ethical-driven AI solution in an HC environment -, it is now possible to move to the following chapters, where the most relevant ethical will be inserted in the AI life cycle. For this matter, the next section will present the methodology to achieve the intended result – quantitative research whose results will be based on the analysed content throughout the literature and case studies addresses, to extract best practices when implementing an ethical-based AI system.

3. METHODOLOGY

As stated before, this study aims to explore the ethical issues for HC providers and receivers and how AI tools can be developed accordingly. The proposed methodology is designed following the Qualitative Research Method Model. According to Oun *et al.* (2014), the qualitative researcher's goal is to collect the understanding of human behaviour and the result that led to such particular behaviour. In other words, the Qualitative Research Method examines and answers questions of how, where, what, when and why a person would act in a certain way toward a specific matter. Furthermore, Malagon-Maldonado (2014) states that Qualitative Research is a very robust and essential research method to be used when little information exists about a subject, there are few instruments to measure the phenomena of interest, or when the research aims to understand the participants' perspective. Qualitative Research can be defined as a form of social inquiry that aims to understand how people interpret the environment in which they live, work, or receive care (Malagon-Maldonado, 2014; Atkinson *et al.*, 2001).

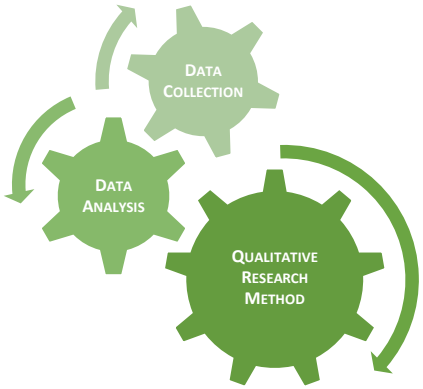

Figure 3.1 - Qualitative Research Method Model - Data Collection and Data Analysis processes can be thought of as the gears that move the Qualitative Research forward (Malagon-Maldonado, 2014).

Data Collection and Data Analysis happen together at the same time, according to the previous authors, who end up adding that "as data is analysed using the constant comparative method and conceptual formulations begin to emerge, the researcher continues to sample theoretically to develop and enrich the categories". Regarding the first of these two phases, data for this study will be collected through a hybrid approach, aggregating text analysis, and case studies. This approach is important due to the lack of sources and available literature regarding ethical issues from some stakeholders' point-of-view, *e.g.,* developers and data scientists from the HC sector. For case study I, data was collected following the participant observation research method (George, 2023). For the remaining case studies, the method used was the same as for the literature review (text analysis after content curation of computerized databases). Regarding the Data Analysis phase, the data collected from the previous activity will be analysed using content and thematic analysis. Data will then be reviewed to identify recurring practices and patterns. The identified themes will be categorized and compared to existing ethics literature in HC to validate the findings. This is a needed process to understand if the retrieved themes can be encapsulated in a framework for ethical appraisals in AI solutions for HC.

3.1. CASE STUDY METHODOLOGY

Now knowing a Qualitative Research will be performed, it is possible to specify the approach to the case study methodology. The case study approach was selected because it thoroughly examines a situation or phenomenon that is complex and specific to one or more real-life settings, according to Pope *et al.* (2020). HC settings are one of the many examples where this approach has been widely used. An example was when, according to the previous authors, sociologists Becker, H. S. and Goffman, E. used case studies to understand HC institutions and the experiences of those working in and receiving care. This kind of studies established the utility of qualitative methods for increasing understanding of what occurs in HC settings, and they offered powerful explanations for the behaviours and attitudes of the people found in these settings.

Qualitative case studies stem from qualitative research and adhere to the principles of the constructivist paradigm, which holds that truth is subjective and dependent on one's viewpoint (Baxter *et al.*, 2008; Simons, 2008). However, the distinctive feature of qualitative case studies is that the focus is on explaining and examining the context of a phenomenon and its influence (Yin, 2009; Cresswell *et al.*, 2016). This clarifies that case studies are fundamentally about enabling a holistic understanding of a phenomenon. Two crucial aspects determine whether a subject can be considered as a case for a case study. The first key factor is the subject itself, which can be an individual, location, small group, organization, family, social group or system. The second factor is the presence of specific or fundamental characteristics related to the research interest and question, forming the analytical framework and providing the theoretical and scientific foundation for the case study (Yin, 2009). Available literature also shows how case studies are defined, approached, and utilized varies significantly in different sources.

In a final analysis, the case study approach used in this study intends to examine projects where AI has reportedly been effectively used for improving HC. Therefore, the aim is to understand what distinguishes these projects in approaching ethics from the rest. Following what was stated above, the present work aims to use this knowledge to provide valuable insights into the ethical issues in HC, and how to address them in an AI system context, while contributing to the existing literature on the topic by producing a useful and practical artefact. The following table showcases a summary of the present chapter.

Table 3.1 - Methodology Summary

METHODOLOGY OUTLINE		
	Methodology	▪ Qualitative Research
	Approach	▪ Systematic Literature Review (SALSA method) ▪ Case Study Research
	Sources of Data Collection	▪ Catalogue & computerized databases searching (LR; case studies II and III)
		▪ Participant observation method (case study I)
	Data Analysis Techniques	▪ Thematic analysis ▪ Content analysis

4. RESULTS AND DISCUSSION

So far, AI has been characterized as a possible disruptor that will change how several activities in HC have been practised. The volume of data gathered and made available in this industry, along with improvements in computing power, have facilitated developments in AI and an explosive surge in publications. However, the creation of AI applications does not ensure that they will be used in everyday life. If AI adoption is not understood correctly, there is a risk that advantages for several stakeholders - patients, personnel, and society - may not be realized despite the resources invested. Additionally, it was understood that the HC environment is featured by diverse stakeholders, apart from the ones already pointed out, who are undertaking different, but impactful roles. This reality, paired with the main ethical problems uncovered in the last phase, will be explored in-depth with three case studies.

The intent of these case studies is to understand best practices for AI applications in HC, by leveraging the concepts exposed during the technical background. Relying on impactful and well succeeded AI solutions for the improvement of the HC industry in any way gives reason to uncover how ethical aspects were addressed and if that approach was vital to unlock the success of the implementation. This way, it will be possible to extract insights on developing AI sustainable solutions for HC purposes, and to bring more developments "from paper to bedside". Besides this, this is an excellent opportunity to uncover, whenever possible, who the main stakeholders involved are, and how these different realities work together.

Afterwards, with the gathered knowledge from the literature review and the case studies' best-practices whether in terms of defining an AI-lifecycle project or when approaching it on ethical terms, a framework will be presented and analysed.

4.1. CASE STUDIES EXECUTION

Following the HC AI topics reviewed in the literature review chapter, the first case study will focus on the reality that AI brought to the administration of two medium-sized hospitals. The collected knowledge derived from the participant observation of a Portuguese company project. As for the second case study, the intent is to unveil how clinical decision-support has been improved when leveraging AI techniques, namely NNs. Finally, with the third case study, one can understand how logistic processes were the target of an AI tool to make mainland Portugal emergency operations more efficient.

All case studies will be explored technically prior to the ethical analysis of each project, as it will contribute to a critical sense needed to effectively scrutinize any ethical approach. The roles of the involved stakeholders will be looked at throughout the case studies to try to understand if these contributions were important for the development and/or implementation's success.

4.1.1. Case Study I

As seen earlier, administration chores in an HC environment are the tasks that consume hospital staff the most, diminishing time dedicated to these professionals' core services, *e.g.*, surgeries, or consulting

patients. To contradict this tendency, and augment existing care, Glintt has developed a Surgery Scheduling Solution (SSS).

The SSS has emerged as a response to a surgical scheduling activity in a hospital context that is currently characterized by often being a cumbersome and time-consuming process for the staff involved. Surgical scheduling activity is challenging due to the complexity of several elements to take into consideration and that need to coexist at a specific time and place for surgery, namely human resources, hospital resources and the timetables set for the procedures.

Conditioned by the context above, and because it is commonly an *ad-hoc* process, there are recurrent instances of scheduling that aren't transparent and with inefficient use of hospital resources. Therefore, the SSS intents to transform the surgical scheduling paradigm by introducing AI into the decision-making process, valuing it as a critical element in the value chain and an enabler of a better HC service. Knowing all this, the primary goals of this tool are to leverage AI to get automated scheduling (an optimal solution is found among all possible combinations of surgeries on the waiting list), resource optimization (concerning surgery rooms and other hospital resources) and to allow "blind" scheduling (crucial to assure the suggested schedule follows the defined restrictions by the Portuguese Health Regulatory Authority[28] in a transparent way).

To fulfil the listed goals, the SSS has several components that constitute the final scheduling proposal. Regarding the mathematical model for optimization that fuels it, CP was the AI core technique chosen to solve a problem that can be defined as a hard combinatorial one, *i.e.*, "problems for which no low-degree polynomial-time algorithms are known" (Hromkovič, 2013). Deep diving into CP, it can be defined as the study of finding solutions to constraints, meaning that imposing restrictions on all the possible solutions is considered a way of "searching in a large space" (Mayoh, 1994) of candidates to find a solution that satisfies all the constraints, *i.e.*, feasible solutions. Usually used for scheduling or resource allocation problems, in the context of the SSS, CP was exploited due to its capacity to choose both a feasible and efficient solution since it can be used to minimize the search space while enforcing desirable properties for a desired solution.

In this case, to define the particularities of the constraints, some inputs need to be configured, particularly the MSS[29]. The MSS is a cyclical calendar that specifies the quantity and availability of surgery rooms, their hours of operation, and the surgeons who will be given priority. In a block system, the MSS allots a specific surgeon or speciality a set amount of time on a given day. The MSS typically has a one-week timeframe, though many organizations make exceptions (Costa, 2015).

Day & Shift \ Room		1		2		3		4	
		Surgical specialty's ID	Surgeon's ID	Surgical specialty's ID	Surgeon's ID	Surgical specialty's ID	Surgeon's ID	Surgical specialty's ID	Surgeon's ID
Tuesday	M	OTO OTO	SRG 2 SRG 208	OTO OTO OTO	SRG 69 SRG 145 SRG 215	GYN		GYN GYN	SRG 59 SRG 39
	A	OPH	SRG 120	OTO		GES GES	SRG 86 SRG 116	GYN GYN GYN	SRG 47 SRG 119 SRG 168
Thursday	M	OPH	SRG 104	OTO OTO	SRG 69 SRG 145			VAS	SRG 49
	A	OPH	SRG 120	OTO	SRG 139	GES GES GES	SRG 143 SRG 34 SRG 16	VAS	SRG 49

Figure 4.1 - MSS example (Marques et al., 2019).

[28] Entidade Reguladora da Saúde, or ERS, for the Portuguese context.

[29] Master Surgical Schedule.

Besides the MSS, the surgery enrolment list[30] is needed. This document contains information related to the patient's diagnosis, ICD code, surgery type, and physician responsible for its enrolment, among other relevant clinical information. Besides this information, this list contains the maximum guaranteed response time (TMRG)[31] for each entry. According to ERS, this date is intended to ensure the right of patients of the National Health Service (SNS)[32] or private providers contracted by the SNS, to access the various types of non-emergency HC at a time deemed clinically acceptable for their condition. For example, this response time also varies if the patient has neoplasms, which is also broken down in the mentioned list.

Having all this information, the SSS has the needed details to define hard and soft constraints. Hard constraints are rules that make a solution feasible, so these always happen. Examples of hard constraints are ensuring that all surgeries *must* be scheduled inside the defined MSS block. Soft constraints are rules that are desirable in a feasible solution, but not mandatory. One example of a soft constraint is to ensure that the suggested schedule *should* try to comply with all pre-scheduled surgeries. Therefore, it can be understood that soft constraints are expressions to be minimized or maximized, whereas hard constraints are described as equalities or inequalities. Having this information, this scenario will now be simplified to better express the problem solved by SSS. Further on, this will not be an opaque process when addressing SSS ethical implications.

Using the use-case produced by Champion (2022) as guidance, let's shorten it and start by defining the set of variables whose values get solved with the determination of each schedule. Let's assume R to be the set of affected rooms, S_r the set of slots in room r, and P the set of surgeries. This way, to link each slot in a room to a surgery, the intent is to identify a set of variables that satisfy hard constraints and tries to satisfy the soft constraints as much as possible. That said, the variables are described as follows. If in room r, the slot s is taken by the surgery p, then $x_{rsp} = 1$:

$$(x_{rsp})_{(rsp)\in R\times S_r\times P}, with\ x_{rsp}\ \in \{0,1\}.$$

Simultaneously, with Ω as the set of all solutions that satisfy all of the hard constraints, H_j, where $j = 1, 2, \dots, n$, we have that:

$$\Omega = \{(x_{rsp}) \mid H_j\}.$$

Once again, having in mind the simplification of the problem, let's reduce it, for notation effects, to one hard constraint ("There *must* be at most a single surgery in every slot") and to one soft constraint ("Surgeries with a lower TMRG *should* be given priority"). Regarding the hard constraint, it can be broken down into two possible scenarios for each slot s, in every room r: (1) There is a surgery p that is unique, or (2) the slot is without a surgery assigned, meaning it is empty. Regarding the soft constraint, it can be translated into the need to minimize the difference between the scheduling day and the TMRG. To illustrate this preference, let's assume we have two surgeries, a and b left to be assigned and only one slot available. Having $TMRG_a = 1$ (day) and $TMRG_b = 10$ (days), it is expected that, keeping all other conditions similar, the scheduled surgery is a, due to this soft constraint. Putting it all into notation, one has that H_1 and S_1 are expressed, respectively, as:

[30] Lista de Inscritos em Cirurgia or LIC, for the Portuguese hospital context.
[31] Tempos Máximos de Resposta Garantidos, or TMRG, for the Portuguese hospital context.
[32] Serviço Nacional de Saúde, or SNS, for the Portuguese context.

$$\forall r \in R, \forall s \in S_r, \sum_{p \in P} x_{rsp} \leq 1 \qquad (H_1);$$

$$\min_{\Omega} \sum_{r \in R} \sum_{s \in S_r} \sum_{p \in P} x_{rsp} TMRG(p) \qquad (S_1).$$

Keeping the simplistic approach, one can also expose the problem to the Pareto front, particularly when accessing which solutions best meet the objective functions in a graphical form. Plotting all feasible solutions, meaning all combinations of surgeries that together comply with the hard constraints in place, the ones that best meet the criteria imposed by the soft constraints, get unveiled. As an illustration of such, the following figure by Bre *et al.* (2017) schematizes this visual approach.

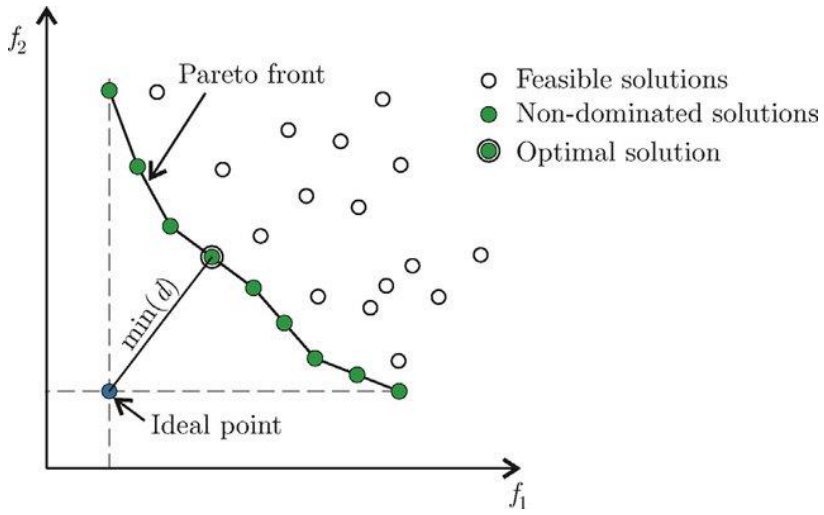

Figure 4.2 - Pareto Front schematization (Bre *et al.*, 2017). In this example, the objectives are defined by f_1 and f_2 to define that, amongst all the possible choices (feasible solutions), only the ones featuring on the Pareto front are suitable candidates to the optimal solution.

As stated previously, this case study reflects a problem that was summarized and simplified to understand the intricacies of CP for optimizing scheduling problems. In real circumstances, there are a lot more details that have not been included, such as the possibility of defining preferred physicians for certain surgeries or defining the maximum or minimum of surgeries to be placed in each block of slots of a specific room, amongst many others that were not presented in the expressions used. To answer this problem, an open-source C++ toolkit was leveraged – Gecode. Developed under Prof. Christian Schulte[33]'s direction, it is a "highly-efficient and award-winning constraint solver" (Dornbach, 2018).

Assuming that the situation at the time of the data collection remains the same in the present, the SSS is implemented and incorporated into the work process of the clinical-administrative staff of a B group and a C group Portuguese hospitals, according to the practised classification by ACSS. In the practical context, it was proven how some solutions produced by the SSS for the speciality of General Surgery could solve two months of delay and ease surgeries that weren't complying with the defined TMRG. This project used a waterfall methodology since new implementations and the approval of new ones highly depended on meetings with each hospital practitioners and management board to participate in point-of-situation meetings. All these stakeholders were essential due to their collaboration with

[33] Computer science professor at Stockholm's KTH Royal Institute of Technology.

the developers contributed to the various stages of the process. Besides this, it is also worth mentioning other elements from the hospitals' staff that impacted the solutions' success, as it is the case of the IT departments. Once the development phase was terminated, the implementation occurred with an on-site formation to the hospital staff that would use this tool. These sessions were relevant to clarify doubts and concerns. They contributed significantly to the acceptance of a somewhat controversial tool due to its AI component, which many fear can negatively impact on the workforce (Topol, 2019).

Crossing the development and implementation of the SSS with the five retrieved main ethical concerns, avoiding any bias seemed to be a concern from the conception of the solution. While using CP (instead of an NN, for instance), there is no need to train the model, which means not needing historical patient data, which is particularly relevant when discussing health data. Another aspect is having a "blind" scheduling, which means choosing surgeries that strictly comply with TMRGs while optimizing the use of the surgery rooms. It is worth inspecting how is this achieved, having a list of patients waiting for surgery as base, with medical and personal information. Although using CP does not require an exhaustive and thorough cleansing and other forms of treating data as other methods, only the strictly necessary information was retrieved to have all the details needed to schedule a surgery, which means that everything that could identify a patient was excluded from every process. With this, concerns regarding patient's privacy were being answered simultaneously.

Transparency and explainability were also a big concern for the SSS developers. The solution was incorporated into user interfaces designed to exhibit the proposed schedule for the MSS and LIC details, and other candidate surgeries that could be part of the suggestion. To allow this to happen, besides showing an agenda view on each room indicated in the MSS, the output also displayed a view to additional rooms – room 99 and room 100. The first one has the objective of showing pre-scheduled surgeries that the model couldn't account for, leading to them not being included in the scenario produced. This is due to, for instance, another concurrent surgery being pre-scheduled at the same time and this one leading to a greater occupation of the block, and thus, to a greater optimization of the block. Another example is if another concurrent surgery pre-scheduled at the same time as the one in room 99 has a TMRG closer to the dates featured in the MSS, having priority over the one in room 99. On the other hand, the purpose of room 100 is to give visibility to surgeries whose characteristics are not compatible with the MSS defined for the speciality in question and, therefore, cannot be scheduled. Although this may seem unlikely or even not useful, this room sheds light on, for example, human errors during the characterization of the MSS. This attribute can also contribute to the trustworthiness and responsibility of the solution.

Besides this being available on an agenda view, the SSS also allows the interpretation of the proposed schedule with the tabular view, using the LIC as guidance (treated as exposed earlier, to bring privacy and avoid human biases, for instance, when analysing the solution using this view). A colour scheme is used to facilitate the interpretability of the solution with surgeries featuring rooms 99 or 100 appearing with the colour orange, and red, respectively. Selected surgeries for the solution are shown in blue, while the remaining ones are without any colour. This view is ordered by the TMRG-defined rule and by neoplasms (that affect TMRG), contributing once more to the compliance with the defined rules by ERS (this aspect in particular contributes to the solution's responsibility). This view also invites the user to think critically towards the solution as it sheds light on the biggest differences, if any, between the surgeries that should be prioritized and the ones the SSS suggests, given the provided context, to

optimize the use of the room. Also contributing to the solution's transparency and trustworthiness, it is possible to check whether the model has found the optimal solution. Theoretically, given infinite time and processing power, the optimal solution would always be found. However, it is recurrent that these conditions aren't the ones in which the user finds themself. The best scenario found at a certain point in time is always possible to be analysed since the application is built to present that one (sub-optimal solution) until a better one is found[34], which also brings trustworthiness to the solution: the scenario seen by the user at any time will always comply with hard restrictions, and is always possible to be used, although it may not be the one that, in the end, brings the most optimization to the resources being employed by the hospital (optimal solution). Nevertheless, there is clarity about that since the scenario will not be signed as the optimal solution. As exposed before, several types of meetings occurred with different stakeholders for this project. Sessions whose aim was to give training to the administrative and clinical staff (intended SSS's type of end-user) and to show useful results produced by the model with real scenarios were also a key contributor to the transparency, explainability and trustworthiness of the solution.

Finally, connected to the topics of responsibility, trustworthiness, and accountability, the SSS team has always presented the solution as a complementary tool to the administrative and clinical staff's work to lessen the burden of tasks that are mainly manual until this point. Therefore, critical thinking seemed to be encouraged in every interaction with the SSS. Although there is not an official statement regarding accountability, both the final decision and main guidelines for the SSS's design originated from the practitioners' category of stakeholders (as defined earlier in chapter 2.2.4.), particularly each speciality clinical team, responsible for the final manually performed surgery schedules.

4.1.2. Case Study II

Exploring AI to bring clarification to assist medical staff, acting as a "second opinion", is not a recent employment (Shiraishi *et al.*, 2011). Leveraging NNs for dermatologic-level classification of skin cancer (Esteva *et al.,* 2017) is another example of such employments in the HC sector, besides the ones already pointed out during the literature review. DeepMedic emerged to deepen the knowledge in this particular use of AI in HC. DeepMedic is an AI software that resulted from the research of elements from the biomedical image analysis group (Kamnitsas *et al.*) of the Department of Computing from the Imperial College in London, United Kingdom.

A three-dimensional (3D) fully connected Conditional Random Field (CRF) and a multi-scale 3D deep CNN are the foundation of DeepMedic. This program for segmenting brain lesions has yield excellent results, winning the ISLES[35] 2015 competition for complex lesion segmentation tasks, such as lesions from ischemic strokes, brain tumours, and traumatic brain injuries.

[34] Whenever a feasible solution (or schedule) doesn't comply with a soft constraint, that said scenario gets "penalized" with a value that depends on how important for providing a quality solution that constraint is. Being so, the best solution possible is the one whose score is the lowest, and the model will be searching for the combination of surgeries that can present the lowest score possible, keeping in mind all remaining constraints (particularly the hard constraints).

[35] Ischemic Stroke Lesion Segmentation (ISLES) is a medical image segmentation competition hosted annually by the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). Many colleges in Germany, the United States of America, and Switzerland are organizing this event with Amazon's sponsorship.

According to the authors, the segmentation and the subsequent quantitative assessment of lesions in medical images are insightful tasks for retrieving information concerning the analyses of neuropathologies. Simultaneously, these allow the planning of treatment strategies and track illness characteristics, progression, and consequences. However, to perform the said quantitative assessment of lesions, multi-modal, 3D images are required and must be precise, which is demanding due to the various aspects or locations traumatic brain injuries can have. Per se, this is what poses the main challenge when defining rules for any discriminative model. Simultaneously, performing manual delineations of such lesions is a "tedious, expensive, time-consuming, (and) impractical in larger studies" task, giving this research a strong relevance for medical practice (Kamnitsas *et al.,* 2016).

From a high level, the architecture of this AI software consists of a CNN with eleven layers deep, 3D and multi-scale. While the 3D CNN allows highly accurate (soft) segmentation maps, a second element – a fully connected 3D CRF -, allows the introduction of constraints that will regulate the CNN output and thus produce the (hard) labels for the segmentation. To get here, the process started with the team defining a baseline. This baseline consists of four layers with 5^3 kernels used for feature extraction, culminating in a 17^3-sized receptive field. Still part of the baseline, there is a classification layer whose 1^3 kernel is placed as convolutional. The goal of this is to allow an "efficient *dense inference*".

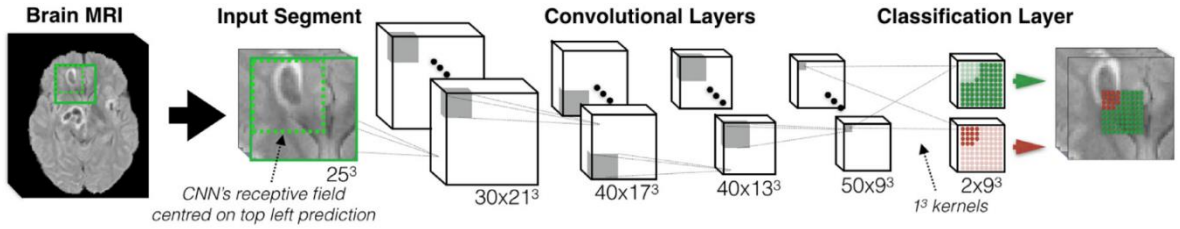

Figure 4.3 – Representation of the baseline model for the DeepMedic software. In the figure, the number and size of feature maps (FM) are represented as (*number* x *size*) (Kamnitsas *et al.,* 2016).

Regarding the FMs in the baseline, each one consists of a collection of neurons that "look" for a specific pattern or feature in the channels of the layer placed before. The said patterns are defined by the weights of the kernel associated with the FM in question.

To better understand the 3-dimensional component of this schema, if the m^{th} FM neurons in the l^{th} layer are arranged in a 3D grid, the image $y_l^m = f(\sum_{n=1}^{C_{l-1}} k_l^{m,n} \times y_{l-1}^n + b_l^m)$ is made up of the activations of those neurons. This is obtained via convolution of each of the channels of the preceding layer with a 3D kernel $k_l^{m,n}$, a learnt bias b_l^m, and a non-linearity f. C_l represents the FM of a layer l. In a final remark relative to the baseline in what concerns the last layer L, also known as the classification layer, the activations of its neurons correspond to specific segmentation class labels. As a result, the neurons are organized into C_L FMs, one for every segmentation class, c, and a position-wise softmax function is fed information about their activations. Having this model as a starting point, DeepMedic, after numerous improvements and tests, led to the already stated composition, whose representation can be seen in the figure that follows, which puts into perspective the road to go down from the initial model (Figure 4.3) to the final one (Figure 4.4).

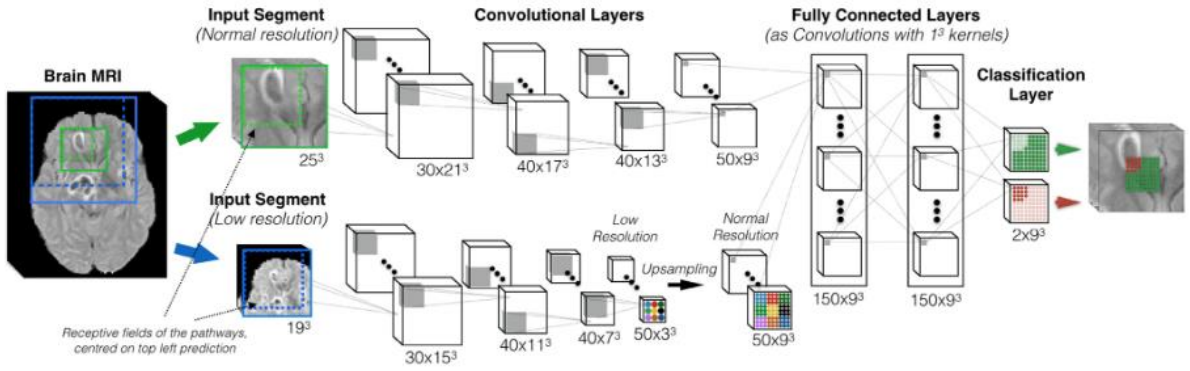

Figure 4.4 - Representation of the DeepMedic model, a multi-scale 3D CNN with two convolutional pathways (Kamnitsas *et al.,* 2016).

At this point, maybe the first difference one notices is the multi-scale element. Multi-scale is used in CNNs to modify the feature extraction part by resizing the inputted image to distinct resolutions, as is exposed in Figure 4.4. This creates a parallel convolutional pathway that centers in the exact location of the image as the first one. On the other hand, some aspects are maintained throughout the work developed *i.e.*, only $(1 \times 1 \times 1)$ strides are employed to enhance an accurate segmentation since it has been reported that bigger strides can downsample FMs (Sermanet *et al.,* 2013). Additionally, the choice of using a fully connected CRF is worth mentioning. CRF overcomes the studied limitations of previous models as it can handle arbitrarily large neighbourhoods while preserving fast inference times. Besides this, the authors believe that DeepMedic is the first use of a fully connected CRF on medical data.

Apart from the undisputable contributions in the technical and innovative areas, these developments were the product of the intervention and work of three stakeholder groups - developers, research bodies, and funder institutions. Supported by organizations such as the EC and the UK's Medical Research Council Program Grant, among others, DeepMedic is, as far as this research goes, not being employed in any "real-world" setting. However, the authors state their hope in the possibility of adopting DeepMedic for either a research or clinical environment, advocating its computational efficiency and improvement of results for state-of-the-art approaches. While this adoption hasn't taken place, multiple researchers employed this AI software for their investigations, showing the generalized acceptance of the research community *e.g.*, Battalapalli *et al.* (2022); Denisova (2023); Huang *et al.* (2022); Lippert *et al.* (2018). This is only possible due to the free and open-source access defined by the authors, with the goal of "facilitating further research and encouraging other researchers to build upon our results" (Kamnitsas *et al.,* 2016). The developed work is based on a widely used language (Python), and platform (TensorFlow) and the source code is available on GitHub[36]. There are ready-to-use packages that can be installed (compatible for Linux and Windows) via Anaconda[37], and images can also be accessed via a pull request on Docker[38]. In parallel, the team made several other resources available to further explain the software: a YouTube video[39] to visualize

[36] DeepMedic's source code and more resources are available at https://github.com/deepmedic/deepmedic.
[37] Package *bioconda/packages/deepmedic 0.6.1.* available at https://anaconda.org/bioconda/deepmedic.
[38] Docker instalation of DeepMedic available at https://hub.docker.com/r/medphys/deepmedic/tags.
[39] Brain Lesion Segmentation by DeepMedic available at https://www.youtube.com/watch?v=V68WRK-CYUw&ab_channel=BenGlocker.

the DeepMedic algorithm's output for the case of a patient scan with traumatic brain injuries (Ben Glocker, 2016) and a website[40] with further details and resources on DeepMedic's algorithm and team (Imperial College London, 2017).

According to the work developed during the literature chapter, all of these endeavours from DeepMedic's team have contributed to an undisputable transparency, one of the most relevant ethics topics. For this matter, and for the sake of explainability, the development of two papers (Kamnitsas *et al.*, 2015; Kamnitsas *et al.*, 2016) also contributed to accountability, particularly when results and technical choices are depicted as in these papers and when the developed work was assessed by juries in a competition (2015 ISLES). According to the authors, the "detailed analysis of the network" they provide sheds light on what they call "the powerful black box of DL with CNNs". As such, for Kamnitsas *et al.* (2016), the discriminative feature of their work is seen as one of its contributions due to its potential to be combined with the clinical knowledge acquired throughout the years by researchers and clinical staff. On the other hand, other author resorting to DeepMedic's algorithm to develop their work builds the trust gained in such technology.

Multiple scenarios emerge regarding the concern of privacy of the patients whose images were analysed. For the case of the traumatic brain injury images, the Local Research Ethics Committee granted ethical approval, and signed consent was obtained for the whole universe of patients (sixty-six) of a UK HC facility in which the images were collected within one week of a moderate-to-severe injury, via consultee agreement. Similarly, to perform ischemic stroke lesion segmentations, the Local Ethics Committee of the HC institutions supplied the scans (Maier *et al.*, 2017). In this case, as these images were provided for the participation in the ISLES 2015 challenge, the scans were also available on the SICAS Medical Image Repository platform[41]. According to Maier *et al.* (2017), by deleting all patient information from the files and the facial bone structure from the photos, complete data anonymization was made possible, thus ensuring both patients' privacy and obliterating any bias that could occur related to race or cultural values, for instance. Just as in the previous case, for assessing brain tumours, DeepMedic also relied on data from a challenge, namely the 2015 Brain Tumour Segmentation Challenge (BraTS). Once more, this data was available through the same repository platform[42], and had the same process of guaranteeing full anonymization, privacy and unbiased treatment to the patients – data was skull-stripped (Menze *et al.*, 2015).

4.1.3. Case Study III

Logistics applied to the HC sector has reportedly (from 2019 to 2022) been facing an increase in research publications (Božić *et al.,* 2022). Matters related to the procurement of medical acts and medicines or the need to make the activation of emergency medical services (EMS) more efficient are some examples of where the study of optimization methods for the HC industry can bring great improvements. This opportunity is starting to be recognized by the HC field, as the indicator given by the previous authors shows. To employ AI to optimize the EMS in mainland Portugal, Data2Help arose.

[40] DeepMedic's oficial website available at https://biomedia.doc.ic.ac.uk/software/deepmedic/.

[41] ISLES 2015 Challenge training and testing data can be accessible through the SMIR platform, available at https://www.smir.ch/ISLES/Start2015.

[42] BraTS 2015 Challenge training and testing data can be accessible through the SMIR platform, available at https://www.smir.ch/BRATS/Start2015.

Data2Help is a project based on the Portuguese EMS – INEM[43] - characteristics, which will be explored further. According to INESC-ID (2019), the aim of this project is to provide INEM with new tools to improve its operational results by optimizing the allocation of resources, resulting in a better and faster response to medical emergencies in Portugal. As seen in the literature review chapter, Data2Help focuses on achieving three main points: integrating INEM's information systems with other relevant external data (*e.g.,* meteorology, epidemics, demographics, forest fires); developing predictive models for emergency vehicle requests in each geographical area by analysing historical data; and optimizing the allocation of INEM's resources based on the predictive models to improve response times to medical emergencies (Manquinho *et al*., 2022). Part of the partnership between the three entities that made Data2Help a reality – INEM, INESC-ID, and the OpLog branch of CEG-IST[44] - another research work was developed with similar goals, but leveraging distinct AI tools, according to Grilo *et al.* (2023). This second work, developed by Abreu *et al.* (2023b), will also take part in this chapter to bring more recent insights and a second vision of what it means to apply AI to logistics optimization for the HC industry.

In Portugal, requesting EMS starts with situations being reported to INEM via a phone call to the number 112. Then, HC professionals in CODU[45] process these requests to be able to allocate resources to each reported situation, *e.g.*, specialized medical technicians or the most suited emergency vehicle. The CODU section of INEM serves as a dispatch centre in charge of responding to 112 calls. The remaining calls after excluding fake ones, have their source on 112, SNS24[46], or Inter-Hospital Emergency Transportation[47] (Abreu *et al.*, 2023b). Although CODU facilities are concentrated in four Portuguese cities, the emergency vehicles that need to be deployed must be sent out across the entire nation (FCT, 2020). Being so, one can conclude that there are two critical times in INEM's operations - the time it takes to respond to each emergency call and the time it takes for the emergency vehicle to travel from dispatch to the emergency location, which is determined by the INEM's rule of using the closest-idle ambulance for an occurrence. As reported by Santos (2023), while an E group hospital, according to the practised classification by ACSS (2017), received about 462 daily emergency patients during 2022, an average of 4100 calls a day were answered at INEM's CODU during the same year. As such, for the same period, CODU activated 1.4 million emergency medical resources during the same year. Shedding light on indicators such as these, it becomes clear how operational productivity is vital to guarantee minimum time responses, as these contribute to a higher probability of a patient's recovery or even survival.

To help achieve that, Data2Help is based on the premise of creating a tool to optimize resource allocation and, consequently, improve the quality and response time of EMS in mainland Portugal. In the first instance, Manquinho *et al.* (2022) worked on implementing best practices for multidimensional database modelling to combine emergency event data with available sources of situational information for context-aware data analysis. The desire to get awareness into such events

[43] INEM, or Instituto Nacional de Emergência Médica de Portugal, for the portuguese context.

[44] OpLog, or Operations, Logistics and Supply Chain Management, is a research group part of CEG-IST, or Centro de Estudos de Gestão do Instituto Superior Técnico.

[45] CODU, or Centros de Orientação de Doentes Urgentes, for the portuguese context.

[46] SNS24, sourced from SNS, consists of a line committed to offer advice and recommendations in medical and medication-related circumstances. According to Abreu *et al.* (2023), these calls are routed to dispatch centres in the event of medical emergencies, where they are handled like 112 calls.

[47] Inter-Hospital Emergency Transportation, as the name suggested, consists of a line specifically designed to facilitate transfers of hospitalized patients to other hospital facilities.

stems from the conviction that emergencies and responses are highly influenced by factors such as the weather, the setting of an epidemic, urban traffic, significant events, or demographic settings. As an example of the benefit that may come from the integration of emergency event data with publicly available sources of relevant situational context, the authors point out the effect that festivals can have on the number of intoxication emergencies or the impact that weather conditions can have on the frequency of cardiorespiratory emergencies. The authors reported that the AS-IS data storage processes for INEM consisted of data stored in a relational database associated with medical emergencies. Simultaneously, assigning medical emergency vehicles in advance was only occasionally done, particularly when required by law, such as for special events where the event's organizer needed to ensure nearby emergency resources. Knowing all of this, part of the developed work was to enable the prediction of emergencies as it is critical for resource allocation, reducing response inefficiencies, and assisting with vehicle allocation at large gatherings, which is something that INEM had not considered until the establishment of this study.

From a high-level view, the proposed solution by Manquinho *et al.* (2022) was an integrated data warehouse that facilitated a quick retrieval of pertinent information from multiple sources to help with various computing tasks. A multidimensional schema with four key attributes of interest was used to incorporate the sources of emergency and situational context data that were available on public sources (*e.g.*, INE, ANSR, IPMA[48]): (1) Multiple calendric and territorial hierarchies with spatial and temporal dimensions; (2) context-specific dimensions and facts introduced to capture one-to-many relationships between emergencies and situational context sources following the spatiotemporal footprint of the observed large-scale events and sensor measurements; (3) multiple fact tables instantiated to distinguish between complete and incomplete information for EMS; (4) complementary dimensions taken into account to hierarchically describe the diagnosis, severity, dispatched vehicles and assistance provided on emergency occurrences. In addition to a straightforward multidimensional data model and an expressive OLAP[49] querying, the developed work provided a service layer to support parametric queries for sophisticated context-enriched spatiotemporal analytics. An example is the possibility to analyse unusual increases in the frequency of emergencies and other constraints through the research and use of various anomaly detection methodologies, namely LSTM[50], or SARIMA[51]. Applying these anomaly detection methods for time-series was only possible due to the proposed multidimensional data model.

In a second instance, Abreu *et al.* (2023b) continued this work, giving response to another goal of the Data2Help project: “develop predictive models for the expected demand of emergency vehicles in different geographic areas” (FCT, 2020). The first steps were to collect, gather and prepare information regarding aspects such as the dispatch centres’ functioning, data produced and current demand for

[48] Portuguese Institutes for Statistics (INE, or Instituto Nacional de Estatística), of Sea and Atmosphere (IPMA, or Instituto Português da Atmosfera), and Authority for Road Safety (ANSR, or Autoridade Nacional para a Segurança Rodoviária).

[49] Online Analytical Processing.

[50] LSTM, or Long Short-Term Memory, is a RNN that has a "forget gate" that regulates whether information should be saved for the long term and which information should be saved for the short term. This is significant because in RNNs, connections between nodes can lead to a loop where output from one node might influence input to a different node later on. However, the LSTM network emerged as a solution to this issue because RNNs are unable to memorize data for an extended period of time and start to forget their prior inputs (Melichov, 2022).

[51] SARIMA model, or Seasonal Autoregressive Integrated Moving Average, is a variation of the ARIMA model since it adds a seasonal parameter to the last model. SARIMA method has into consideration seasonal autoregressive order (P), seasonal difference order (D), seasonal moving average order (Q), and the number of time steps for a single seasonal period (m) (Hanbanchong *et al.*, 2012; Brownlee, 2018).

these facilities. At this stage, analysis resorting to both INEM's and public contextual data allowed to infer multiple aspects from population growth to its ageing.

Still prior to the modelling phase of the forecasting models, there was the need to reduce the impact on neighbouring zones because ambulance services are not centralized, contrary to mainland Portugal's phone handling operations. The influence on surrounding zones is closely tied to the administrative divisions and the decentralized emergency vehicle operations. Knowing this, the authors decided to implement a DT to conduct a study in discrete zones without neglecting the demand shared by neighbouring zones. The algorithm used the coordinates of all the bases[52] spread over the country and the coordinates of all occurrences to aggregate the demand considering the nearest bases. Still in an exploratory phase, two datasets emerged: one focused on call volumes, where two priority levels were defined (P_1 and P_3), and another focused on dispatch volumes considering vehicle types (AEM, AMBSIV and VMER[53]).

Regarding which independent variables would be used in the forecasting models, correlations between an initial set of seventy variables were analysed. Twenty variables were kept to perform the predictions. For instance, keeping temperature, humidity and wind speed variables may help to explain incidences related to forest fires, a common event in Portugal. Additionally, the intention to also have a classification model to predict calls needed further work on the definition of the dependent variable, by turning it into a categorical variable instead of a continuous one. For that, k-Means algorithm was employed, ending in the definition of three clusters for both P_1 and P_3 calls, based on unsupervised learning (elbow method, which displays the SSE[54]).

In the modelling phase, resorting to the Keras API, NN models were built using a Feed-Forward NN architecture with dense layers. The FFNN architecture, whose connections amongst nodes don't form a cycle, having the information processed in a single direction, has been used before to produce spatial-temporal models, taking contextual information into account (Setzler *et al.*, 2009; Huang *et al.*, 2019), as well as to enhance decision-making in HC when it comes to prediction, classification, or diagnosis tasks (Shahid *et al.,* 2019). After performing a fine-tuning of hyperparameters and a (stratified, for classification models) 3-fold cross-validation[55] to avoid bias, the authors state with a level of confidence of 95% that the chosen independent variables aid in explaining the behaviour of EMS call demand in mainland Portugal, except for national holidays and one district in the regression model for P_1 calls, a few days of the week and a few months in the models for P_1 and P_3 calls.

At this point, it is understood that the combination of these two studies which leveraged AI under the Data2Help project intended to give a response to the already stated goals and to provide a response to forecasting CODU's anticipated workload, optimizing CODU staff schedules to meet anticipated demand, and developing software tools to maximize the number of active emergency vehicles and personnel across the nation in each work shift. From a practical point-of-view, so far, the delivery of this project has been the responsibility of the stakeholders involved in the scientific research team,

[52] Location where emergency vehicles are housed.

[53] For the portuguese context, AEM, or Ambulância de Emergência Médica; AMBSIV, or Ambulância de Suporte Imediato de Vida; VMER, or Veículo Médica de Emergência e Reanimação.

[54] Sum of Squared Error.

[55] Through the process of k-fold cross-validation, one can assess the model's performance on the test dataset—that is, its ability to generalize—by randomly dividing the entire dataset into subsets (train, validation, and test dataset) either once or more times (in this case, three times, meaning that k=3).

and the ones INEM has allowed to guide the said team, disclosing the processes' know-how. However, the authors state that closer collaboration with practitioners and decision-makers is needed when it comes to the models' improvements.

Now looking at the ethical preoccupations throughout and after the development of this project (that started in 2019, and is ongoing (CEG-IST, n.d.)), one can assume that accountability and trustworthiness are topics that go hand-in-hand and that are achieved through the collaboration of three renowned institutions in the fields of scientific research and teaching – IST, FCT, and INEM. This engagement gives credibility to the work developed, which also contributes to the trustworthiness that the various HC stakeholders can place on the project. This is something that, by consequence, can undoubtedly influence the implementation's success of the solution in a real-world setting. Along the same line of thought, the € 294.036 financing by FCT, under the INCoDe.2030[56] program (Marmé, 2018), contributes to similar feelings among HC stakeholders *e.g.*, INEM's HC professionals, INEM's decision-makers. Other aspects that contribute to said ethical topics are (1) the safeguards the researchers draw attention to, namely the planning context or the operational level in which the produced models should be used; and (2) the tested applicability of the developed work, in the first paper by promoting query efficiency, which validates the relevance of the data warehousing system, and on the second paper by applying the design forecasting models to the Portuguese case study.

Moving on to the topics of transparency and explainability, the presentation to society through various formats (*e.g.*, papers' publication, presence at conferences, namely the 2023 ICE (Cunha *et al.*, 2023)), and mentions in area-related magazines article (Grilo *et al.*, 2023), gives the project the needed transparency for HC endeavours, and gives opportunity also to build explainability. Just like these, making more documents available, annexed to the published papers (Abreu *et al.*, 2023a), that go deeper into the technical bases, considerations and reasons for the choices made throughout the project, as the independent variables or the chosen architecture, are undoubtedly factors that contribute to the said ethical considerations. It is also indisputable how endeavours such as those exposed here promote the deconstruction of concerns that are usually adjacent to AI, such as the "black box" one.

On the last two ethic topics – bias and privacy - few are the considerations made by the authors. However, some reflections can be made since there is a lot of information made available by the authors regarding the used data. The mentioned contextual data used in both studies is sourced from national institutions, *e.g.*, INE, that publish distinct and various sets of data to perform analysis as the ones analysed here. As such, these entities are required by law to comply with national, international and European rules and to have information security, privacy and personal data protection policies in place. Regarding the data prevenient from INEM, very little is known besides that it consists of historical data retrieved from INEM's system under study. However, when analysing the selected variables for the forecasting models, although some can be classified as able to handle SPI[57], no combination of variables that can identify the individual whose occurrence was intended (PII[58]). When it comes to any bias that the solution may present, no considerations are given by the authors in this regard. However, all available data seemed to be used in the various processes, from choosing the final

[56] For the portuguese context, INCoDe.2030, or Iniciativa Nacional Competências Digitais 2030, aims for the strenghten of digital competencies for the portuguese society.
[57] Sensitive Personal Information.
[58] Personally Identifiable Information.

independent variables to applying all of them, in distinct instances, to the same models. Therefore, one is led to conclude that since every record went through the same models, the resulting strategy is the same for all outputs prevenient from the same model, affecting all locations and citizens served by INEM the same way in a real-case scenario.

At this point we analysed three projects using AI in the HC sector for three distinct areas – Administration, Clinical Decision-Making, and Logistics. The first case study showed how the transparency of processes, and the approach of shedding light into a "black box" process via educational sessions could bring confidence to a fearful clinical staff. The second case study showed how an early-on adopted privacy strategy and the multiple adopted forms of spreading the work developed could contribute to build trust and avoid bias despite the AI technique employed (DL is often used to illustrate what the "black box" paradigm is). The last case study was essential to show how, despite the volumetry of data and process' complexity and responsibility, stakeholders' engagement take an essential part in defining the success and quality of a project. Having identified key practices as these, it is now possible to leverage them as an informed base for producing a framework that will allow the development of AI in the HC sector due to the merging of ethical considerations.

4.2. Discussion & Proposed Framework

So far, we have seen how research based on AI has been increasing in the past years (Figure 2.1) and have covered in-depth three opportunities where AI has proven to have the potential to improve HC quality for HC staff, patients, and society overall. Simultaneously, it was reviewed how ethical concerns and studies have increased, mainly when allied to new emerging technologies that have arisen to the central public at a pace like AI (Figure 2.7). Nevertheless, few are the solutions that enter HC facilities to assist those they were intended to. This may be due to some of the AI-related concerns already exposed throughout the present work. These apprehensions, from which the ethical ones can be fairly highlighted, may lead us to an "AI winter", bringing all involved stakeholders into a disenchantment state about the potential of this technology due to the high expectations set by an AI hype that has characterized the last years for many, whether from the academia, business or institutional side – all expect to improve processes, but most are afraid of change.

As stated previously, one of the intentions of the present work is to avoid the said period for AI by bringing tools to the table that allow a development of this technology and brings confidence to users and developers by connecting the retrieved ethical concerns to a holistic process based on the best practices studied so far for developing AI solutions. The now presented framework (Figure 4.5) is based on several aspects of what was studied and presented so far, namely the summarization of key phases during the development of an AI project, aligned with the insertion of the five most relevant ethical concerns – accountability, transparency and explainability, bias, privacy, and trustworthiness and responsibility. In this process, each phase (and correspondent steps) will be dissected. Then, the presence of each of the five ethical topics in the framework will be analysed along with essential guidelines, regulatory documents, and best practices to meet the goal of creating and utilizing AI in HC environments in a secure manner while keeping the "human-in-the-loop" as the indication of which stakeholders should be involved throughout the various phases of the project will showcase.

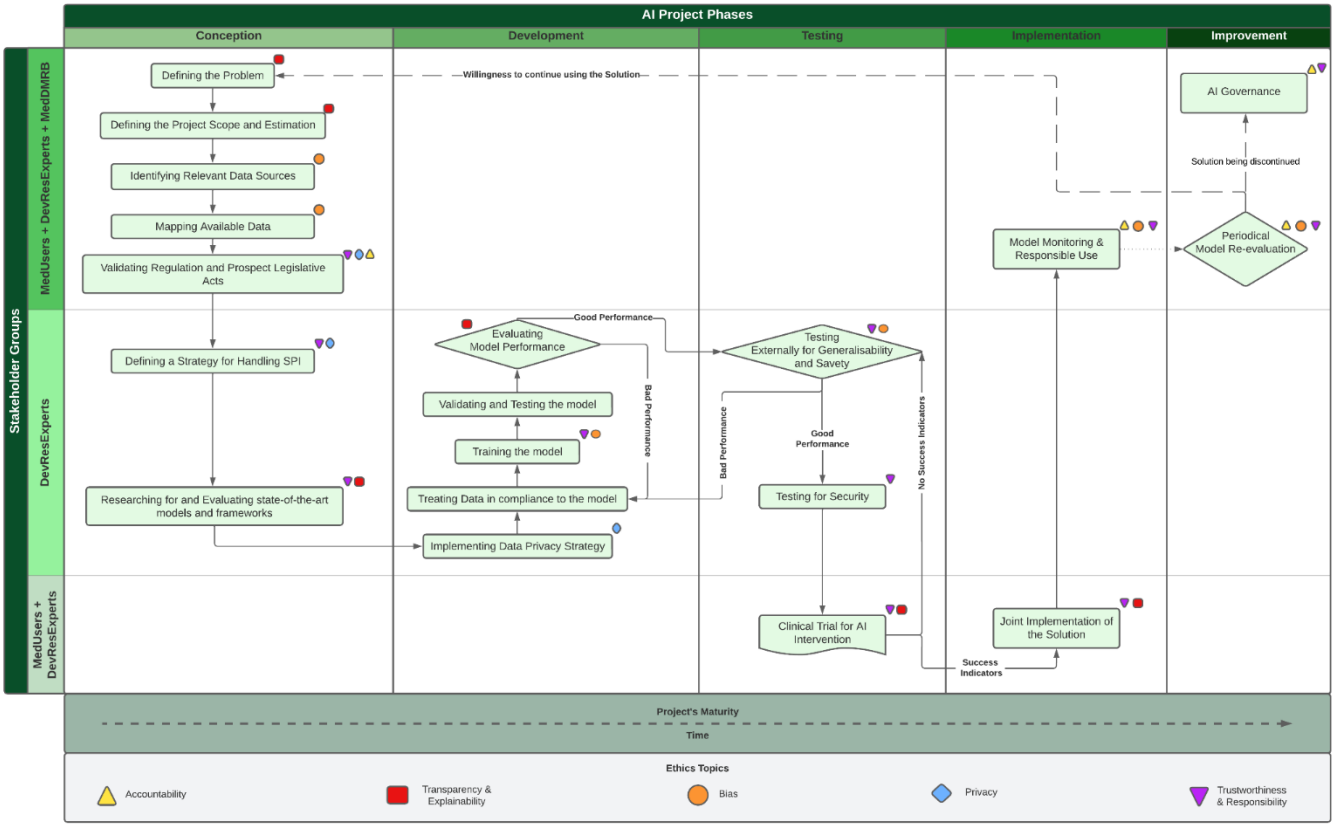

Figure 4.5 - Suggested Framework for AI Developments applied to the HC sector. Divided into four stages and three distinct combinations of stakeholder groups, phases are depicted here as a result of the studied AI state-of-the-art implementations via the approached case studies and the held research exposed in the Literature Review chapter. Adding value to it, ethical topics also feature the framework, along with engaged stakeholders to bring the most benefits to the development and success of the project.

The suggested framework presents four distinct phases (columns) to characterise an AI lifecycle that reflects the beliefs of the authors studied – Conception, Development, Testing, Implementation, and Improvement. Simultaneously across every phase, it is suggested which stakeholders (rows) should be involved in the phase at stake. Assigned to several of the depicted steps (boxes inside the columns/rows), there are ethical topics (colourful symbols at the top-right corner of some of the represented steps) as a way of representing the timing when these concerns should be addressed for projects of this nature. Following the studied approaches in this document to define who the HC stakeholders involved in AI projects are, it is proposed that there are three groups – Medical Users (here abbreviated as MedUsers), Development and Research Experts (abbreviated as DevResExperts), and Medical Decision-Makers and Regulating Bodies (shortened in the framework as MedDMRB). The first group concerns all the HC workforce and practitioners, who will ultimately be the end-users and those with higher interaction with the AI tool. These are also the actors whose work tasks will be more affected and suffer changes. The DevResExperts are the group of actors most knowledgeable about the technologies in place, *e.g.*, solution architects, HC investigators, data scientists, developers, and consultants. These are the ones who will design the approach for the problem at hand, develop and construct the project in an initial phase and be responsible for effectively transferring the knowledge about the tool in a later stage. The last group – MedDMRB – represents physicians that occupy HC institutions' boards or other places of management and oversight bodies, *e.g.,* ERS.

Moving on to the presented phases, Conception is the stage of the project that precedes the development of the AI solution and is when it becomes crucial to engage every stakeholder since themes such as the definition of the clinical problem, scope, and estimation will be addressed, which makes a multidisciplinary team at this stage crucial. Naturally, it is expected that the ones included in the MedUsers group are the ones that will have a better perception of the problem to be solved, its challenges and intricacies. These are also the ones that can better help the development team to locate, retrieve, and identify the needed resources, namely the data. Accompanying this whole process, MedDMRB must be involved to help shape the scope of the project and to negotiate details such as estimations for it *e.g.*, if the model should be interacted with via a UI by the end-users. Although these may seem superficial throughout the process, these actors are indispensable for critical phases, to help maintain a healthy rhythm or to contribute to smothering synergies and collaboration between HC professionals and the technology team. This is crucial throughout the entire project since it is much more challenging to implement a solution and make a project succeed if users don't accept it from the outset or believe that the presence of other teams won't bring any advantage to their routine activities. This reality, which has already been addressed when discussing the general public's current perception of AI, will be explored in greater detail later. Still in the first phase, the current context makes it essential to consider any regulation of legislative acts considering the industry at stake or the type of data one will have to treat, which also connects to the need to define, according to the knowledge gathered in the previous step, a strategy to handle the probable presence of SPI. Multiple sources should be analysed when it comes to these topics, namely the AI Act by the European Parliament (2023), and GDPR (2016) for European countries, or any solution dealing with data from European citizens. In the United States of America, the FDA[59] proposes a regulatory framework (2019) for AI and ML, considering solutions with these components as Software as a Medical Device (SaMD) when used to identify, address, or avoid health issues (IMDRF SaMD Working Group, 2013), needing therefore to comply with specified requirements to obtain legal approval. Like the analysed step, searching and evaluating models, frameworks or technologies considered state-of-the-art is a step that DevResExperts mainly handle. This common practice consists of researching amongst the numerous published models that can be applied to the problem at stake. In the HC sector, considering the work developed during the previous step, the FDA has cleared 521 AI-based Medical Devices (FDA, 2022) that can be leveraged while having the confidence that an approval like this can bring to users and patients.

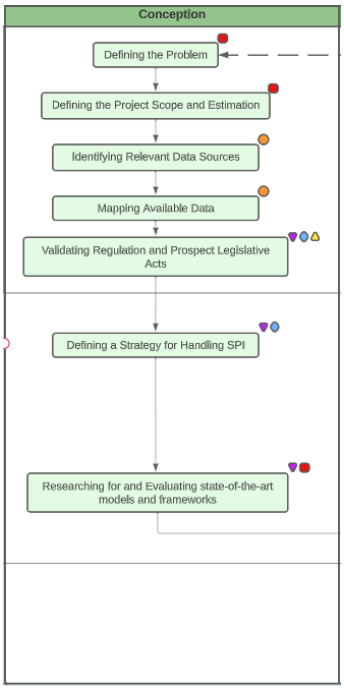

Figure 4.6 - Proposed Framework Focused: Conception phase.

Moving on to the second phase – Development -, mainly assured by DevResExperts, it starts by implementing the defined strategies regarding the privacy of patients whose data belongs to. This step is only relevant when the sources used don't have the data masked, and it becomes possible to identify patients, or when the data is sensitive. If, similarly to the case at this stage in case study II, the data is

[59] The Food and Drug Administration, or FDA, is a government organization under the USA Health Department whose job it is to safeguard the public's health by guaranteeing the efficacy, security, and safety of pharmaceuticals for humans and animals as well as biological products and medical equipment.

already masked to bring anonymization both for the model's input and output, this could be a redundant step. If this is not the case, the governance of the used dataset should come into place, and cryptography (encrypting data) and informed consent (the setting where training tasks are performed is only open to patient data for which consent was granted) are a must. Then, tasks that are related to data treatment are taken. Some include merging data from different files, filtering data incoherencies, managing missing values, labelling the desired outcome (for supervised learning techniques), computing extra variables, and variable selection (e.g., choosing variables with predictive solid power, examining correlations). Next, the model is expected to be trained, validated, and tested, *i.e.*, its reliability is evaluated using unobserved data to evaluate its performance. When testing, methods such as k-fold cross-validation (used in case study III) are widely used. Later on, while evaluating the model's performance, if the chosen metrics (*e.g.*, accuracy and F1-Score), don't show positive results, this process should be reformulated, which is common. In the case of HC models, some metrics can take the lead. For instance, having a model that falsely ignores customers who would adhere to a marketing campaign when predicting customer adherence can make a company lose the chance to convert into money by missing the opportunity to pitch the campaign to these clients. In the medical field, a false negative for a model that helps in diagnosing a specific disease may have a price that can't be calculated by leading to the death of the patient wrongly predicted, since the treatment won't be taking place. A metric that can be used for predictive models with a higher concern for false negatives is recall[60] (or sensitivity).

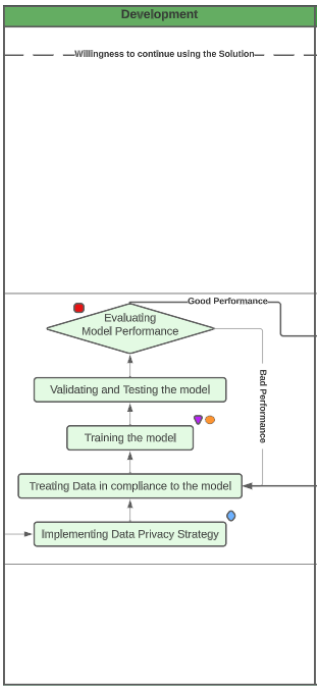

Figure 4.7 - Proposed Framework Focused: Development phase.

If the chosen metrics show satisfactory results for the DevResExperts, a new phase begins – Testing. In the first step of this phase, the intent is to fully understand if the model can be adapted to other scenarios, given the same problem. The intent behind this step is to take into account the possibility of testing the model with other datasets, provided by the organization for which the model is being developed to, or public ones, generated for contests, for instance, as it was the case of the datasets used during the case study II. However, unlike other medical devices, it is important to keep in mind that AI models provide predictions for each of the situations presented *i.e.*, for each patient, which naturally can imply that the model doesn't perform as greatly as in the circumstances it was trained for. Due to this, it is fundamental that, if the aim is to apply it externally, the model must be validated to ensure generalisability, to bring safety to the affected patients in any external environments. Another aspect that must be evaluated to guarantee generalisability in external environments is if the training and validation populations are comparable. In case any of the said tests don't show results as promising as the ones from the phase before, the team may consider retraining the model or altering conceptual aspects. However, if this doesn't happen, as some authors have raised concerns about the level of security and/or robustness that using platforms which host AI solutions may present, it's important to test for security in case there are any attacks. According to Alan Turing Institute (Leslie, 2019), adversarial offensives involve the deliberate manipulation of input data, such as, for cases of CV models, subtle adjustments to a picture's pixels, to lead to inaccurate predictions or

[60] Recall (or sensivity) for one of the labels of a dependent variable is defined as $\frac{\#\,TP}{\#\,TP+\#\,FN}$, where TP stands for True Positives and FN for False Negatives.

misclassifications. The assurance that the solution team can create a safe system that can preserve the integrity of the data is the purpose of security. In addition to maintaining uninterrupted functionality and accessibility for authorized users, a secure system safeguards private and confidential data against adversarial attacks. Model hardening, a sophisticated method that has been researched to thwart adversarial attacks by fortifying the systems' architectural elements, can be applied to do this. It is based on either architectural modification or adversarial training, which involves carefully expanding training data to include antagonistic cases. Other options are available, namely the Adversarial Robustness Toolbox (ART), by IBM (Nicolae *et al.,* 2018/2023) - a Python ML security library designed to give teams the tools they need to protect ML models and applications from adversarial risks including extraction, poisoning, inference, and evasion. This may be a viable option to consider as ART supports a variety of popular ML frameworks, including TensorFlow, Keras, PyTorch, scikit-learn, and XGBoost; all data types, including images, tables, audio, and video; and specific ML tasks, including classification, object detection, speech recognition, generation, and certification. Coming to the final step of the Testing phase, it is important to carry out a clinical trial to better understand how the AI intervention should be done and in what terms. For that, the MedUsers and DevResExperts can resort to the SPIRIT-AI[61] and the CONSORT-AI[62] guidelines, developed by the same working group (Cruz Rivera *et al*., 2020; Liu *et al.,* 2020). In a nutshell, these guidelines are an improvement over previous versions, with the exception that the new version contains criteria for assessing interventions that involve an AI component. To achieve the goal of promoting transparent evaluation of new interventions, CONSORT-AI for elaborating trial reports, and SPIRIT-AI for elaborating clinical trial protocols, work together to ensure that the clinical trial is conducted in a randomized setting and that steps are extensively detailed to enhance replication by others. Additionally, they provide evidence-based recommendations for the minimal set of items to be addressed when determining how to appropriately convey the results of any AI model to the end user, so that they are both actionable and useful.

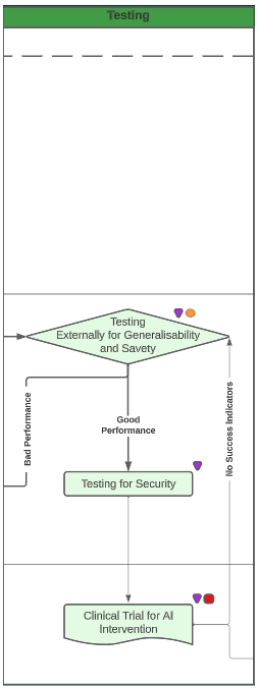

Figure 4.8 - Proposed Framework Focused: Testing phase.

At this point, Implementation can start upon the success of the prior phase. If that's not the case, it might be wise to consider going back to the beginning of the Testing phase and reformulating it. Implementation should start with the collaboration between DevResExperts and MedUsers to jointly bring AI to patients' bedside. This phase is mostly about passing on knowledge to HC professionals about the model and about the best practices for using the solution so that it reaches its full potential and leads to the most beneficial results for patients. This step can take various forms: point-of-situation meetings, with a high training component, following the case study I example, or thematic workshops, to grasp distinct model aspects (*e.g.*, technicalities, limitations, use-cases, best practices). As exposed earlier, this step is as crucial as the development of the model, since unsuccessful training and pitching on how this new component can improve patients' lives and HC professionals' work life can lead to an unused solution because there was no acceptance from its intended users: No AI model will have an impact on any HC reality if its users reject it. Resistance to change and scepticism about technology

[61] SPIRIT-AI, or Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence guideline.
[62] CONSORT-AI, or Consolidated Standards of Reporting Trials–Artificial Intelligence guideline.

can be typical reactions by the HC workforce when the perception is that digital transformation has the purpose of replacing HC professionals. However, these joint meetings should also be an opportunity to shape what the future of patient interaction should be, by promoting moments of education to build an AI-literate workforce. So, to avoid the said fears and to successfully implement the AI solution, a consistent and transparent dissemination of AI model information to end users must occur to uphold the ideals of openness and confidence, which are crucial to winning over HC professionals' confidence. All of this, though, is only feasible if all parties involved work together to create a training and educational setting where workforce knowledge can be reframed. People must view digital HC technology solutions in the context of their own clinical practices and quality improvements in care. This can be achieved through promoting lifelong learning, being open to collaboration and effective codesign, and having a greater understanding of human intelligence. These are all essential components of a culture of learning. On this topic, a framework for digital capabilities is proposed by Topol, E. (2019). Whichever form it takes, encouraging HC worker education is essential to guaranteeing that AI models will be applied securely after they are put into practice, bringing us to the following stage, in which all parties concerned are involved to guarantee the model is used responsibly. An instance of an undesirable circumstance is the use of AI models in a biased manner when training data differs from real-world data because of shifting paradigms related to sickness and care *i.e.*, data shift. Overconfidence in AI, as Kidd (2020) put it, leads to MedUsers "blindly" accepting the outputs of AI models, which is also not desirable, particularly in situations as data shifts. Therefore, contributing to an AI-literate workforce is also what prevents giving too much weight to models' output for clinical decision-making, and can help DevResExperts to detect a data shift situation, a deterioration of the predictions given by the model or if there is still utility in the AI solution to solve the problem thanks to an alarm given by the end-users, that remain alert and conscious when using the said tool.

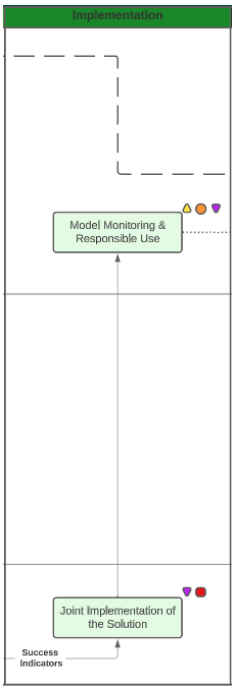

Figure 4.9 - Proposed Framework Focused: Implementation phase.

Faced with the possibility of any of these situations occurring, it is important to periodically review the models, either internally or through an audit, which leads us to enter the Improvement phase that must be accompanied by all groups so far involved. An intermittent arrow shows that there is still no formally defined period for these activities to take place, because this said period would vary greatly depending on the purpose and technologies used. The purpose of this re-evaluation is to provide answers if the evaluation's original questions can be rigorously re-answered or if using the AI solution will continue to be beneficial in the long run. More practically speaking, every specific model or solution in clinical use should be periodically inspected to see if the accuracy of the output differs from the application's prior performance criterion, which is true of all HC devices. Following this reassessment, if there is the desire to continue using the solution, it's wise to review the first phases, particularly if the HC team is faced with the same problem as before, and if so: Is the model chosen in previous interactions and its parameters still the best way to go? Has there been any technological advance that can come as resourceful to apply to the problem in-hands? There is a lot to consider at this stage, which justifies the arrow going back to the

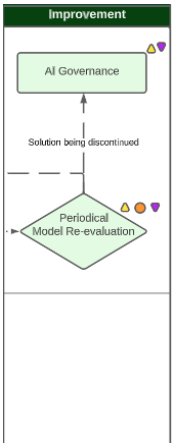

Figure 4.10 - Proposed Framework Focused: Improvement phase.

beginning of the framework. However, if there is the desire to discontinue the use of the solution, or terminate maintenance or support, action should be taken, particularly on the subject of AI governance. AI governance comes into play since aspects such as model accountability, version control, adaptability, data security, data quality, data access, and overall data accountability are all covered under it. This phase is particularly important since it allows to keep the associated assets, reasoning, findings and documentation even when the model has been decommissioned while advocating for patients' safety. These holdings must be retained for several years, according to IMDRF SaMD Working Group (2015), for eventual requests, as legislation may dictate that the data must be appropriately archived, and not destroyed, for eventual requests related to patient safety or systems environment *e.g.,* protecting patient data and any other confidential information (which may involve data removal, transferring patients' data to a new product or solution, or safely archiving user information); notifying users of significant milestones in a timely manner so they have time to research, assess, and approve potential alternatives; and submitting a plan to the MedDMRB for gracefully ending maintenance and support of the AI solution.

Having explored extensively and exhaustively each step and phase of the proposed framework, it's now time to present the reasoning behind the inclusion of the most frequent ethical topics in the analysed literature – accountability, transparency and explainability, bias, privacy, trustworthiness and responsibility. In the framework, they are symbolically associated with steps in the process that, when carried out, lead to teams approaching, inspecting and integrating an ethical vision combined with a technical one - something that is particularly relevant in the HC industry, as it has been proven so far. Additionally, the ethical approach taken in each of the case studies will be critically addressed again, this time having the framework as guideline.

Accountability, as exposed during the literature review chapter, is a somewhat controversial, debatable and poorly regulated topic when it comes to solutions that use AI for their operation. It's also recognised that the main issue regarding this topic is who the main "culprits" might be when there are bad outcomes for those who are supposed to benefit from the presence of these devices, whether they are patients or HC staff. Although no "adoption of legal framework conditions" (Hagendorff, 2020) is in place to compensate for possible harms caused by AI systems, the proposed framework addresses an approach also suggested by Leslie (2019) – accountability by design, meaning that for complete answerability and auditability, a human-in-the-loop methodology must be used at all times. Being so, this is a topic that gets to be addressed through many steps - validating regulation and prospect legislative acts (Conception phase), model monitoring and responsible use (Implementation phase), periodical model re-evaluation, and AI governance (Improvement phase) – where activity monitoring protocols are produced and kept. It should be noted that for all of these steps, all the stakeholders are involved. This is not in vain, since accountability demands the support of virtue ethics from all parties concerned, especially the IT communities. This view is supported by the belief in distributed responsibility, where Floridi (2016) describes how duties can be assigned using the backpropagation technique from DL. However, in this case, backpropagation is applied in networks of distributed responsibility, where it is intended to demonstrate how every actor has to be held accountable or accept some degree of responsibility for the consequences of their actions. Automated decisions are not self-justifiable, leading to an accountability gap if there isn't this approach of having everyone in the loop from the beginning. This way, even though there may be complex AI processes and/or teams, a responsible and engaged approach builds an accountable attitude. During the study of the three mentioned cases in the methodology chapter, the ways of approaching accountability were through

the engagement of stakeholders at the conception and implementation phases and the fomentation of responsible use, by appealing to the critical sense of each user (case study I), an exhaustive report of results and reasoning behind architectural and data choices (case study II), and through the associated scientific institutions, that seem to back up the presented developments (case study III). It is noticeable how there are both common steps, and steps here suggested that one has no information if they were held or not, namely the validation of regulation and prospect legislative acts. Besides this, there isn't information if any decommissioning took place.

Transparency and explainability, as mentioned earlier, is seen as a quality requirement for an AI system, due to three perspectives that coexist when it comes to defining (and putting into practice) this ethical value - virtue, relation and system. The way transparency and explainability are introduced in the proposed framework reflects exactly this coexistence in order to bring sustainability and quality to endeavours that rely on AI. Starting with the Conception phase, when the process of defining the problem, project scope and estimations is held with the input of every involved actor, transparency is exercised when every stakeholder has the opportunity to communicate intents, concerns or particular objectives. This exercise allows for independent judgment and decision-making when it comes to backing the uttered goals or knowing if this is the best methodology to achieve them *i.e.*, using an AI model or solution. In the same phase, researching for and evaluating state-of-the-art models and frameworks is a step that perhaps leads to further discussion, since it is the time when an approach is decided, leading to the so-called "black box" due to several authors supporting the trade-off that can exist between explainability and the model's complexity. As an example, one can try to picture exposing the particularities of a linear regression versus exposing the reasoning behind a NN. However, it's not uncommon having situations where a more complex model provides better results. For these cases, Leslie (2019) produced guidelines to consider when using such AI techniques that (1) carefully contemplate weigh up risks and effects; (2) have into account alternatives for further interpretability that confer a simultaneously domain-appropriate and implementation-consistent semantic explanation, in order to (3) create an interpretability action plan for non-technical actors *i.e.*, MedDMRB, MedUsers, to brief the system's decisions or behaviours, or reasoning. These guidelines are based on the author's belief that transparency and explainability are based on three critical actions: justifying processes, justifying outcomes and having clear content and outcome explanations. Evaluating the model's performance is adjacent to the previous step and is another clear opportunity to transparently report results. To achieve that, for the case of prediction models, Collins *et al.* (2015) suggest the TRIPOD statement, and more recently, TRIPOD-AI (2019) that is reportedly being developed. This 22-item checklist[63] promotes thorough and open reporting that accurately reflects the conduct and design of the study. Because transparency as a system needs auditability, it is imperative to conduct clinical trials for AI interventions or maintain assets for AI governance. AI systems in HC must be built with an explainable architecture, in line with cognitive human decision-making processes that the MedUsers are familiar with, and directly connected to clinical data - any stakeholder who utilizes the output to guide clinical decisions must be able to understand how an AI model or solution functions and what it produces for them to assess if the system is likely to fulfil its stated goals. To guarantee these features, Char *et al*. (2020) auditability should include aspects of the development phase *e.g.*, training data and process, and of the implementation phase *e.g.*, for clinical trials, pre-

[63] The TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) checklist can be consulted at https://www.tripod-statement.org/wp-content/uploads/2020/01/Tripod-Checlist-Prediction-Model-Development.pdf.

specification of study design, or outcome measures, and analysis are required to enable a potential audit. Finally, when it comes to jointly implementing the model, much of what has been pointed out so far contributes to the success of this phase and can also be applied here. At this point, it is essential to communicate AI model information understandably and consistently to foster confidence and transparency. To guarantee that everyone using the tool has the necessary knowledge and awareness, clarifications such as if the AI used is "locked" or "continuously learning" are the kinds of information that should be shared at this point to ensure everyone involved has the needed knowledge and awareness when handling the tool. Leslie (2019) provided four explanatory tactics to help technical teams assess various methods of successfully doing their knowledge transfer, and the *What-If* tool[64], made available by Google People+AI Research, aims for a user-friendly interface for deepening one's comprehension of regression or classification “black box” ML models. When we look at the analysed case studies, no formal guidelines or frameworks were reported to be followed to guarantee explainability. However, some good practices as producing a user-friendly UI (case study I) – that enables distinct presentations of outcomes, leading to a deeper understanding of processes, or several materials to have more ways of explaining outcomes (case study II), have shown to be good practices. Further considerations regarding transparency worth mentioning are that this openness can, on one hand, contribute for stakeholders to weigh potential conflicts of interest better, but on the other hand, must be balanced against the protection of intellectual property of the developed work.

Bias is a topic that has proven to have serious consequences when the best decisions are not taken to avoid it in AI systems. Proof of such was when racial bias was being sustained by AI-assisted jail sentencing rules in the United States of America (Angwin *et al.,* 2016) or, in the HC case, when decades later, it was noticed how the Framingham study[65] data presented bias since it mainly considered data from Caucasian individuals. As Gijsberts *et al.* (2015) have proved, “the magnitude of associations between risk factors (for cardiovascular events) (…) differ between race/ethnic groups”, proving how the former study both overestimated and underestimated the risk of cardiovascular events when predicting it in non-Caucasian populations. Situations such as these can certainly be avoided when teams aim to build a fairness system, following steps comply to it. For this purpose, the steps that feature the framework are for the Conception phase, identifying relevant data sources and mapping available data; for the Development phase, training the model; for the Testing phase, testing externally for generalizability; for the Implementation phase, model monitoring and responsible use; and for the Improvement phase, periodical model re-evaluation. The idea behind the mentioned steps is to shed light on the impact that these can have to achieve fairness since data-driven technologies, which collect and identify data based on societal dynamics, easily reproduce, reinforce, perpetuate, and amplify the patterns of marginalization, inequality, and discrimination that already exist in these societies. In addition to this reality, technical teams dictate processes like feature selection, modelling, and characterization; giving these teams the ability to readily include implicit or explicit prejudice or preconceptions into their decisions, either intentionally or unintentionally, even in activities that appear to be morally neutral at first look, which may be due to a “group effect”. Moreover, hidden bias frequently arises because the available data utilized in the training step doesn't appear to be as representative of the populations from which inferences are being drawn as they should be.

[64] The *What If* tool is available at https://github.com/PAIR-code/what-if-tool.

[65] The Framingham study is population-based, observational cohort research that was started in 1948 with the goal of examining the epidemiology and risk factors for cardiovascular disease prospectively by the US Public Health Service (Boston Medical Center, n.d.).

Altogether, because the data being fed into the algorithms is faulty from the beginning, this generates significant possibilities of biased and discriminatory outputs that are suboptimal for specific populations. In contrast to bias observed in conventional research (such as selection bias), bias in AI models can also be classified as algorithmic and social bias. These biases might result from various aspects like race, gender or measurement mistakes and teams must be ready to completely abstain from performing certain tasks that are deemed unethical, even when output is ready for decision-making (*i.e.*, responsible use of the model), and to act when there is a periodical re-evaluation and a data shift is occurring to guarantee that all sectors of the aimed population are represented, in which "external" testing can help. Due to all of this, some approaches are suggested to overcome such challenges. For example, using the Principle of Discriminatory Non-Harm (Leslie, 2019), teams are focusing on bringing fairness in data and design to the AI systems or models, where the already mentioned aspects are ensured in outcomes - advocating that these structures shouldn't affect lives in an unfair or discriminatory manner -, and in implementation – when it is ensured that AI assemblies are implemented by people who have received enough training to do so, both ethically and impartially. Another approach is to leverage the Prediction Model Risk of Bias Assessment Tool (PROBAST) made by Wolff *et al.* (2019) which is organized into four domains - participants, predictors, outcome, and analysis – and comprises a total of twenty signalling questions to enable a targeted and transparent method to analyse the risk of bias while assessing the applicability of research that design, validate, or update prediction models for individualized predictions. Moreover, many IT firms currently provide solutions for these purposes, such as IBM (Bellamy *et al.*, 2019), which has released the *AI Fairness 360* open-source toolkit[66]. Looking now at what has been done regarding fairness in the three case studies, very few considerations are made, making it unclear whether it was a concern. Therefore, it may be wise to consider the previous guidance.

Regarding privacy, some practical measures have already been pointed out for the steps in the framework where such an ethical topic is involved – validating regulations and prospect legislative acts; defining a strategy for handling SPI; and implementing a data privacy strategy. Nonetheless, it is important to further notice that AI systems pose threats to privacy as a result of their implementation as well as their design and development processes. Since the foundation of AI projects is data structuring and processing, personal data will often be used to create AI technologies, especially since ML demands a lot of training data. The deployed AI may, in some cases, be interpreted as violating the right of the data subject to privacy if the data is being collected and extracted without the appropriate consent or if it is handled in a way that exposes or jeopardizes the reveal of personal information since it raises concerns about data privacy and data ownership. This is especially relevant for AI used in HC, as training data will probably come from personal demographic data and data collected from specific patients during routine clinical care (like laboratory test results, biopsy results, or diagnostic images) or from specific health insurance plan enrolees (like medical diagnoses from encounters or patterns of HC utilization). When it comes to the considerations made throughout the case studies, the authors rely on the chosen processes and entities, and further formal approaches aren't mentioned. Just like the previous ethical topic, it may be wise to consider implementing the proposed steps in their frameworks.

Lastly, multiple steps are presented as an opportunity to achieve a sense of trustworthiness and responsibility for the project being held. Trustworthiness, as exposed during the literature review, is

[66] The AI Fairness 360 toolkit is available at https://github.com/Trusted-AI/AIF360.

an ethical topic that benefits from the remaining ones, and can only fully be achieved with their presence, particularly the explainability and transparency ones. As exposed by Ribeiro *et al.* (2016), “determining trust in individual predictions is an important problem when the model is used for decision making. When using ML for medical diagnosis (…) predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.”. With this setting, the LIME algorithm arose. LIME, to determine whether to believe a prediction, select a model, enhance an unreliable classifier, and determine the reasons behind a classifier's lack of trustworthiness, is an algorithm that, by locally approximating it with an interpretable model, may faithfully explain the predictions of any classifier or regressor. This objective was set since it is thought that building human confidence and fostering efficient usage of ML requires a thorough explanation of predictions. As a result, by "explaining a prediction," LIME displays textual or visual items that offer a qualitative comprehension of the connection between the model's prediction and the instance's constituent parts (such as words in text or patches in an image). Reflected in the case studies, trust and responsible use of AI were always concerns since thorough reasoning was made throughout the published papers, and institutions stood by the developed systems, ultimately sponsoring them or allowing them into said institutions. However, regarding each of the case studies, much more work can be done in this area, and it is believed the framework can be a useful tool for achieving future successful real-world implementations.

Having carried out an exhaustive analysis of the produced framework at this point, a few parallel themes will be emphasised. Although they are not included in the framework on their own, they are central to its usability. Firstly, when five ethical topics were selected (Appendix, Table 1), many relevant ones weren’t mentioned until this point. However, producing a value-upholding AI artefact that values and promotes human autonomy and has environmental and workforce impacts into consideration throughout the process, while keeping in mind educational rights, seemed to also occupy a very relevant space on the reasons why much scientific efforts don’t cross the HC facilities’ doors. Although these concerns are not featured in the produced framework, they weren’t forgotten and were incorporated into the strategies provided throughout this work to overcome the five most cited (and chosen) ethical topics. The human-in-the-loop, present particularly when advocating the engagement and presence of all stakeholders in multiple phases of the process and present where explainability and transparency were strongly encouraged and demonstrated to be crucial, is part of the needed efforts to achieve an AI solution that promotes humans’ autonomy and that works towards their values. More recently, the emergence of the LLM, which will undoubtedly be part of multiple solutions, including in the HC sector, also presents a clear drawback regarding environment sustainability. This is an additional concern that, although more recent, due to the brutal difference when compared to traditional methods on environmental costs, must not be forgotten, as a call for computational efficiency must take place by promoting techniques such as sparsity, which, according to Patterson *et al*. (2021) require less computational efforts without sacrificing performance.

Another highly mentioned concern is regarding both the future of patient interaction and the work impact, and whether these technologies will promote job replacements. Regarding these aspects, the answer lies in promoting education for an AI-literate workforce. Although we had the opportunity to analyse that the current societal perception of AI lies on the extremes of this technology, although none exists (ASI), we also saw how stakeholders are promoting AI not as a way of replacing anyone but as a way of bringing more power and time to the workforce to devote their attention exactly to what most probably has brought them to the HC industry – to have a meaningful patient interaction.

However, several aspects are missing, namely the right to educate such workforce on technology. Fear happens mostly when one doesn't or can't understand the feared subject. However, it's crucial that the opportunity to understand is given – education. Education can take many forms, some of which were already referred, and more can be consulted on a study held by the Topol Review (2019) to promote the preparation of the HC workforce to deliver the digital future without fearing change but also by fomenting a critic sense that avoids situations of overconfidence, which according to Kidd (2020), are also easy to fall into since, whenever one is designing a technology that is going to deliver information to people, it is powered by algorithms that, at minimum, can be influencing the order in which results are being presented. This order, from previous psychology research, can make a big difference in what people walk away with. Algorithms can have the power to shape beliefs and change behaviour. However, education for AI literacy can allow technology, people, and responsibility to coexist and bring enormous advances, particularly for the HC industry.

5. CONCLUSION

AI has never been mentioned as much as it is now, and its implications for our society range from social ones to economic, political and, naturally, ethical ones. Its fast pace and disruptiveness have the power to cause different opinions, since few deny how its impact will define many aspects of our lives from now on. Having already changed many industries and processes, HC is not an exception to the rule. With this in mind, the main objective of the present study was to answer the proposed RQ: “How can ethical considerations be incorporated into the development and implementation of AI for HC?”.

To answer this question that has a societal-level relevance, the held research process ranged from understanding both the technical and more general view regarding AI, to understanding the state-or-the-art artefacts that applied this technology in the spec case of the HC industry. In parallel, ethics was also studied to understand the best ways of addressing it in a technological context, and what frameworks had already been produced. Three real-world scenarios were further researched using the case study methodology to understand the best practices of success projects for each of the said scenarios and on ethical terms.

With the gathered knowledge, an answer to the RQ was provided through the production of a framework, which summarizes the several steps that a multidisciplinary team should invest time in if they’re aiming for a sustainable project that leverages AI in a value-upholding manner, for both patients, HC workforce and technical professionals involved.

The developed work allowed to contribute to the solution of a relevant problem, by pointing out critical steps for HC projects while also proposing practical solutions for each of these steps. It also allowed to expose what has been done in Portuguese contexts related to two out of three main HC topics. The produced framework promotes a shared, ethical and step-by-step evolutive approach that is accessible to the understanding of all HC stakeholders. Finally, it also allowed, despite the pointed limitations, for a societal gain that highlights the relevance of governing AI, particularly with its advances and at a time when few AI-specific regulations are widely known to help avoid an “AI winter”. Simultaneously, the present work also contributes to guaranteeing access to a fundamental human right – HC -, which, even as it evolves with AI’s contribution, leaves no one behind, as it strives not to exacerbate inequalities but to be fair, transparent and respectful of autonomy and privacy.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The main limitation of this work is the effect of culture when defining what ethical values are. Although the goal of this thesis is to promote the production of an AI that doesn't exacerbate inequalities, one must be aware that these social impacts are highly influenced by the ethical implications of deploying AI, which is always fuelled by the impression one has of such technologies – always conditioned by the cultural setting. Because of this, the produced framework may not be as effective when employed in locations where the culture isn't similar to the European one, since the studied cases derived from Portugal and the United Kingdom.

Likewise, referring to the used method to ascertain the most frequent ethical topics in the literature, it is relevant to notice that most studies belonged to academia and government bodies, meaning that most presented theoretical guidance and did not mirror cases or opinions based on industry or on practical events.

Moreover, there was the need to resort to many studies from the medical area and crossing them with the ones from the information technology area. This was a needed step because there were few studies that intersected both areas, while also presenting a technical approach of AI techniques. However, addressing them whenever relevant throughout the work was a concern since the present work is aimed at both technical and non-technical (HC) profiles. In demographic terms, most studies reflected opinions of authors from developed countries and there was a low proportion of women amongst the researchers.

Further limitations lie with the unpredictable effect and pace of new technologies, which the suggested framework cannot reduce, only provide some approaches to cope with it. AI is a broad and fast-paced field, meaning that the framework may not be as powerful for some approaches as it is to others, particularly the ones that are yet to come and may even become obsolete depending on the new advances, which are always unpredictable (this can also mean that a new tool or guideline may appear that makes this RQ a "non-problem"). There is also a wide variety of HC stakeholders, meaning that it is highly probable that not all feel greatly represented, although the approach of the study is also to provide an insight into the technical choices that are frequently decided aside from non-technical profiles.

Due to the stated limitations, believe future work could address the effect that different cultures can have when producing such framework, which naturally includes involving more diverse and global point-of-views, and is possible when the language barrier is surpassed. Connected to this last point, collecting the opinion and cooperating with health stakeholders would most likely lead to a much more enriched and complete framework by implementing their vision, and not only the vision of a technical profile. As it was demonstrated, the role of health stakeholders is vital to lead any project to success, and it is only possible when there is as much involvement from them as from technical profiles.

Both faced as a limitation and an opportunity for future work is the fact that the present framework has no reported evidence concerning its effectiveness besides the theoretical indications. Since this work has a theoretical scope and was limited on time, it becomes impossible to test its conclusions in any real-case project, since it would require both to accompany a project end-to-end, and to have some management role to be able to dictate its phases.

REFERENCES

Abreu, P., Santos, D., & Barbosa-Povoa, A. (2023a). Appendixes A (Exploratory Data Analysis) and B (Performance of forecasting models) to Data-driven forecasting for operational planning of emergency medical services. https://ars.els-cdn.com/content/image/1-s2.0-S0038012122002993-mmc1.pdf.

Abreu, P., Santos, D., & Barbosa-Povoa, A. (2023b). Data-driven forecasting for operational planning of emergency medical services. Socio-Economic Planning Sciences, 86, 101492. https://doi.org/10.1016/j.seps.2022.101492.

ACSS. (2017). *Benchmarking Hospitais - Grupos e Instituições*. Benchmarking-Acss.min-Saude.pt. https://benchmarking-acss.min-saude.pt/BH_Enquadramento/GrupoInstituicoes.

AI for Everyone. (n.d.). Constraint Programming. In AI Glossary. Retrieved August 2023, from https://www.aiforanyone.org/glossary/constraint-programming.

AI Now Institute. (2023). *2023 Landcape*. AI Now Institute. https://ainowinstitute.org/.

Anaconda.org. (2020). Deepmedic. https://anaconda.org/bioconda/deepmedic.

Anderson, M. (2019). How Should AI Be Developed, Validated, and Implemented in Patient Care? *AMA Journal of Ethics*, *21*(2), E125-130. https://doi.org/10.1001/amajethics.2019.125.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias—There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Artificial intelligence act, 2021/0106 (2023). https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

Aruna, S. (2011). A Novel SVM based CSSFFS Feature Selection Algorithm for Detecting Breast Cancer. International Journal of Computer Applications, 31.

Atkinson, P., Coffey, A., & Delamont, S. (2001). A debate about our canon. Qualitative Research, 1(1), 5–21. https://doi.org/10.1177/146879410100100101.

Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., & Kujala, S. (2022). Transparency and Explainability of AI Systems: Ethical Guidelines in Practice. In V. Gervasi & A. Vogelsang (Eds.), *Requirements Engineering: Foundation for Software Quality* (Vol. 13216, pp. 3–18). Springer International Publishing. https://doi.org/10.1007/978-3-030-98464-9_1.

Bartz-Beielstein, T., Branke, J., Mehnen, J., & Mersmann, O. (2014). Evolutionary Algorithms. *WIREs Data Mining and Knowledge Discovery*, *4*(3), 178–195. https://doi.org/10.1002/widm.1124.

Bartz-Beielstein, T., Preuss, M., Schmitt, K., & Schwefel, P. (2010). *Challenges for Contemporary Evolutionary Algorithms*.

Battalapalli, D., Rao, B. V. V. S. N. P., Yogeeswari, P., Kesavadas, C., & Rajagopalan, V. (2022). An optimal brain tumor segmentation algorithm for clinical MRI dataset with low resolution and non-contiguous slices. BMC Medical Imaging, 22, 89. https://doi.org/10.1186/s12880-022-00812-7.

Baxter, P., & Jack, S. (2015). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. The Qualitative Report. https://doi.org/10.46743/2160-3715/2008.1573.

Bear Don't Walk, O. J., Reyes Nieva, H., Lee, S. S.-J., & Elhadad, N. (2022). A scoping review of ethics considerations in clinical natural language processing. *JAMIA Open*, *5*(2), ooac039. https://doi.org/10.1093/jamiaopen/ooac039.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2023). AI Fairness 360 (AIF360) [Python]. Trusted-AI. https://github.com/Trusted-AI/AIF360 (Original work published 2018).

Ben Glocker (Director). (2016, March 15). DeepMedic—Brain Lesion Segmentation. https://www.youtube.com/watch?v=V68WRK-CYUw.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., & Wang, G. (2008). Algorithmic Prediction of Health-Care Costs. *Operations Research*. https://doi.org/10.1287/opre.1080.0619.

Bhat, S. (2022, September 20). *What are Expert Systems in Artificial Intelligence? 2023*. Great Learning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/expert-systems-in-artificial-intelligence/.

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in Healthcare*, 25–60. https://doi.org/10.1016/B978-0-12-818438-7.00002-2.

Böke, S. S. (2020). *Artificial intelligence in health care: Medical, legal and ethical challenges ahead*.

Bossmann, J. (2016, October 21). *Top 9 ethical issues in artificial intelligence*. World Economic Forum. https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/.

Boston Medical Center. (n.d.). Framingham Study. Boston Medical Center. Retrieved 27 October 2023, from https://www.bmc.org/stroke-and-cerebrovascular-center/research/framingham-study.

Božić, D., Šego, D., Stanković, R., & Šafran, M. (2022). Logistics in healthcare: A selected review of literature from 2010 to 2022. Transportation Research Procedia, 64, 288–298. https://doi.org/10.1016/j.trpro.2022.09.033.

Bracamonte, V. (2019, January 21). *Challenges for Transparent and Trustworthy Machine Learning*. Vrbracamonte-ITUPresentationv1.3; KDDI Research, Inc. https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20190121/Documents/Vanessa_Bracamonte_Presentation.pdf.

Bre, F., & Fachinotti, V. (2017). A computational multi-objective optimization method to improve energy efficiency and thermal comfort in dwellings. *Energy and Buildings*, *154*. https://doi.org/10.1016/j.enbuild.2017.08.002.

Browne, R. (2023, May 15). *Europe takes aim at ChatGPT with what might soon be the West's first A.I. law. Here's what it means*. CNBC. https://www.cnbc.com/2023/05/15/eu-ai-act-europe-takes-aim-at-chatgpt-with-landmark-regulation.html.

Brownlee, J. (2018, August 16). A Gentle Introduction to SARIMA for Time Series Forecasting in Python. MachineLearningMastery.Com. https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/.

Buchanan, B. G., & Smith, R. G. (1988). Fundamentals of Expert Systems. *Annual Review of Computer Science*, *3*(1), 23–58. https://doi.org/10.1146/annurev.cs.03.060188.000323.

Cambridge Dictionary. (2023, July 12). *Ethic*. https://dictionary.cambridge.org/dictionary/english/ethic.

Centro ALGORITMI - Universidade do Minho. (n.d.). ICDS4IM – Intelligent Clinical Decision Support for Intensive Medicine. *Centro ALGORITMI - Universidade Do Minho*. Retrieved 7 April 2023, from https://algoritmi.uminho.pt/projects/icds4im-intelligent-clinical-decision-support-for-intensive-medicine/.

Centro de Estudos de Gestão Instituto Superior Técnico. (n.d.). *Data2Help - Data science for the optimization of emergency medical services*. Retrieved July 1, 2023, from https://cegist.tecnico.ulisboa.pt/~cegist.daemon/projects/data2help-data-science-optimization-emergency-medical-services.

Champion, A. (2022, January 24). *When you should use Constraint Solvers instead of Machine Learning*. Medium. https://towardsdatascience.com/where-you-should-drop-deep-learning-in-favor-of-constraint-solvers-eaab9f11ef45.

Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. *The American Journal of Bioethics : AJOB*, *20*(11), 7–17. https://doi.org/10.1080/15265161.2020.1819469.

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics, 21(2), 167–179. https://doi.org/10.1001/amajethics.2019.167.

Cho, M. K. (2021). Rising to the challenge of bias in health care AI. *Nature Medicine*, *27*(12), Article 12. https://doi.org/10.1038/s41591-021-01577-2.

Chou, J., Murillo, O., & Ibars, R. (2017, October 12). *What The Kids' Game "Telephone" Taught Microsoft About Biased AI*. Fast Company. https://www.fastcompany.com/90146078/what-the-kids-game-telephone-taught-microsoft-about-biased-ai.

Chowdhury, G. G. (2020). Natural Language Processing. *Fundamentals of Artificial Intelligence*. https://doi.org/10.1007/978-81-322-3972-7_19. Conjeti, S. (2023, June 25). Transforming Healthcare: A Step-by-Step Guide to Building and Deploying AI Medical Devices | LinkedIn. https://www.linkedin.com/pulse/transforming-healthcare-step-by-step-guide-building-ai-conjeti/.

Collins, G. S., & Moons, K. G. M. (2019). Reporting of artificial intelligence prediction models. The Lancet, 393(10181), 1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ, 350(jan07 4), g7594–g7594. https://doi.org/10.1136/bmj.g7594.

Copeland, B. (2019, September 19). DENDRAL. Encyclopedia Britannica. https://www.britannica.com/technology/DENDRAL.

Cordeiro, J. V. (2021). Digital Technologies and Data Science as Health Enablers: An Outline of Appealing Promises and Compelling Ethical, Legal, and Social Challenges. *Frontiers in Medicine*, *8*, 647897. https://doi.org/10.3389/fmed.2021.647897.

Costa, A. M. O. de S. da. (2015). *Gestão de Capacidade em Serviços*. https://repositorio.ucp.pt/bitstream/10400.14/21709/1/TFM_ANT%C3%93NIO_COSTA.pdf.

Council of Europe. (2019). *Unboxing Artificial Intelligence: 10 steps to protect Human Rights*. https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64.

Creswell, J. W., & Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed). SAGE Publications.

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., Ashrafian, H., Beam, A. L., Collins, G. S., Darzi, A., Deeks, J. J., ElZarrad, M. K., Espinoza, C., Esteva, A., Faes, L., Ferrante di Ruffano, L., Fletcher, J., Golub, R., Harvey, H., Haug, C., … Yau, C. (2020). Guidelines for clinical trial

protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. The Lancet Digital Health, 2(10), e549–e560. https://doi.org/10.1016/S2589-7500(20)30219-3.

Cuevas, E., Gálvez, J., & Avalos, O. (2020). Fuzzy Logic Based Optimization Algorithm. In E. Cuevas, J. Gálvez, & O. Avalos (Eds.), Recent Metaheuristics Algorithms for Parameter Identification (pp. 135–181). Springer International Publishing. https://doi.org/10.1007/978-3-030-28917-1_6.

Cunha, D., & Ribeiro, S. (2023, May 23). APEMERG - International Congress on Emergency—ICE 2023. Lifesaving, 28. https://issuu.com/lifesaving/docs/lifesaving_28.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

de Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P., & Horrocks, J. C. (1972). Computer-aided Diagnosis of Acute Abdominal Pain. *British Medical Journal*, *2*(5804), 9–13.

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, *3*(6), 100489. https://doi.org/10.1016/j.patter.2022.100489.

Déclaration de Montréal. (2018). MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE 2018. https://5dcfa4bd-f73a-4de5-94d8-c010ee777609.filesusr.com/ugd/ebc3a3_506ea08298cd4f8196635545a16b071d.pdf.

DeepMedic. (2023). [Python]. DeepMedic. https://github.com/deepmedic/deepmedic (Original work published 2016).

DeLapp, K. (n.d.). *Metaethics | Internet Encyclopedia of Philosophy*. Retrieved 15 July 2023, from https://iep.utm.edu/metaethi/.

Deloitte. (2022). *State of Ethics and Trust in Technology Annual report First edition*. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/about-deloitte/us-tte-annual-report.pdf.

Delua, J. (2022, November 15). *Supervised vs. Unsupervised Learning: What's the Difference?* https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.

Denisova, A. (2023). Evaluation of DeepMedic Neural Network for a Region of Interest Eextraction in Medical Image Watermarking. 2023 11th International Symposium on Digital Forensics and Security (ISDFS), 1–6. https://doi.org/10.1109/ISDFS58141.2023.10131753.

Diakopoulos, N., & Friedler, S. (n.d.). *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms: FAT ML*. Retrieved 7 April 2023, from https://www.fatml.org/resources/principles-for-accountable-algorithms.

Docker Hub. (2017). medphys/deepmedic—Docker Image | Docker Hub. https://hub.docker.com/r/medphys/deepmedic.

Dornbach, J. (2018, September 7). *SAP uses Gecode, an award-winning constraint solver, in S/4HANA for advanced variant configuration | SAP Blogs*. https://blogs.sap.com/2018/09/07/sap-leverages-gecode-an-award-winning-constraint-solver-in-s4hana-for-advanced-variant-configuration/.

Dudovskiy, J. (n.d.). *Purposive sampling*. Research-Methodology. https://research-methodology.net/sampling-in-primary-data-collection/purposive-sampling/#_ftnref1.

Dugdale, D. C., Epstein, R., & Pantilat, S. Z. (1999). Time and the Patient–Physician Relationship. *Journal of General Internal Medicine*, *14*(Suppl 1), S34–S40. https://doi.org/10.1046/j.1525-1497.1999.00263.x.

EARTO. (n.d.). *INESC TEC – KnowLogis Efficient Healthcare Logistics | EARTO*. Retrieved 1 July 2023, from https://www.earto.eu/?post_type=rto-innovation&p=11672.

Entidade Reguladora da Saúde. (2020). Tempos Máximos de Resposta Garantidos (TMRG). https://www.ers.pt/pt/utentes/perguntas-frequentes/faq/tempos-maximos-de-resposta-garantidos-tmrg/.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056.

European Commission. (2019, April 8). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

European Commission. (2021a). *Study on eHealth, Interoperability of Health Data and Artificial Intelligence for Health and Care in the European Union Lot 2: Artificial Intelligence for health and care in the EU Final Study Report.*

European Commission. (2021b). *Study on eHealth, Interoperability of Health Data and Artificial Intelligence for Health and Care in the European Union Lot 2: Artificial Intelligence for health and care in the EU Final Study Report - Country Factsheets* (pp. 130–134).

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112.

Food and Drug Administration. (2022). Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices.

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

Fundação para a Ciência e a Tecnologia (FCT). (2020). Research in Data Science and Artificial Intelligence applied to Public Administration (p. 35). Fundação para a Ciência e a Tecnologia. https://www.fct.pt/wp-content/uploads/2022/06/Brochura_ResearchinDataScienceandAIappliedtoPA.pdf.

Future Advocacy. (2018). *ETHICAL, SOCIAL, AND POLITICAL CHALLENGES OF ARTIFICIAL INTELLIGENCE IN HEALTH A report with*. https://wellcome.org/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf.

Gates, B. (2023, March 28). *Here's what the age of AI means for the world, according to Bill Gates*. World Economic Forum. https://www.weforum.org/agenda/2023/03/heres-what-the-age-of-ai-means-for-the-world-according-to-bill-gates/.

GECODE - An open, free, efficient constraint solving toolkit. (n.d.). Retrieved 25 August 2023, from https://www.gecode.org/.

George, T. (2023, March 10). What Is Participant Observation? | Definition & Examples. Scribbr. https://www.scribbr.com/methodology/participant-observation/.

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, 295–336. https://doi.org/10.1016/B978-0-12-818438-7.00012-5.

Gibney, E. (2020). The battle for ethical AI at the world's biggest machine-learning conference. *Nature*, *577*(7792), 609. https://doi.org/10.1038/d41586-020-00160-y.

Gijsberts, C. M., Groenewegen, K. A., Hoefer, I. E., Eijkemans, M. J. C., Asselbergs, F. W., Anderson, T. J., Britton, A. R., Dekker, J. M., Engström, G., Evans, G. W., de Graaf, J., Grobbee, D. E., Hedblad, B., Holewijn, S., Ikeda, A., Kitagawa, K., Kitamura, A., de Kleijn, D. P. V., Lonn, E. M., … den Ruijter, H. M. (2015). Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events. PLoS ONE, 10(7), e0132321. https://doi.org/10.1371/journal.pone.0132321.

Glintt. (2020). *"QUAL O POTENCIAL DAS TECNOLOGIAS EMERGENTES NA SAÚDE?"* https://apdsi.pt/wp-content/uploads/2020/12/3-Ricardo-Gil-Santos-GLINTT-KnowLogis.pdf.

Glintt. (2022a). *Consultoria de Eficiência*. https://www.glintt.com/pt/o-que-fazemos/mercados/healthcare/Consultoria/Paginas/Consultoria-de-Eficiencia.aspx.

Glintt. (2022b). *Consultoria de Gestão em Centros/Unidades Hospitalares*. https://www.glintt.com/pt/o-que-fazemos/ofertas/BusinessConsulting/Paginas/Adjust.aspx.

Glintt, & APAH. (2019). *Barómetro da Saúde Digital da Adoção da Telessaúde e de Inteligência Artificial.* https://apah.pt/portfolio/barometro-telessaude-inteligencia-artificial/.

Glintt, & APAH. (2022). *Barómetro da Saúde Digital da Adoção da Telessaúde e de Inteligência Artificial.* https://www.glintt.com/pt/o-que-somos/noticias/Documents/Relat%C3%B3rio-Completo-Bar%C3%B3metro-Sa%C3%BAde-Digital-2022.pdf.

Google People+AI Research. (n.d.). What-If Tool. Retrieved 28 October 2023, from https://pair-code.github.io/what-if-tool/.

Google People+AI Research. (2023). What-If Tool [HTML]. PAIR code. https://github.com/PAIR-code/what-if-tool (Original work published 2018).

Graber, M. L., Franklin, N., & Gordon, R. (2005). *Diagnostic Error in Internal Medicine*. ResearchGate. https://www.researchgate.net/publication/298348382_Diagnostic_Error_in_Internal_Medicine.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal, 26(2)*, 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.

Griffis, J. C., Allendorfer, J. B., & Szaflarski, J. P. (2016). Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. *Journal of Neuroscience Methods*, *257*, 97–108. https://doi.org/10.1016/j.jneumeth.2015.09.019.

Grilo, M., & Barros, F. (2023, August). Artigo INEM - Inteligência Artificial no Pré-Hospitalar. Lifesaving, 29. https://www.chualgarve.min-saude.pt/lifesaving/.

Guillot, J. D. (2023, May 11). *AI Act: A step closer to the first rules on Artificial Intelligence | News | European Parliament*. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence.

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients. *Healthcare Informatics Research*, *19*(2), 121–129. https://doi.org/10.4258/hir.2013.19.2.121.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8.

Hanbanchong, A., & Piromsopa, K. (2012). SARIMA based network bandwidth anomaly detection. https://ieeexplore.ieee.org/abstract/document/6261934.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0.

Heeley, E., Anderson, C. S., Huang, Y., Jan, S., Li, Y., Liu, M., Sun, J., Xu, E., Wu, Y., Yang, Q., Zhang, J., Zhang, S., & Wang, J. (2009). Role of Health Insurance in Averting Economic Hardship in Families After Acute Stroke in China. *Stroke*, *40*(6), 2149–2156. https://doi.org/10.1161/STROKEAHA.108.540054.

HIMSS Analytics. (2019). *Health-IT predictions for Europe*. https://europe.himssanalytics.org/sites/himssanalytics_europe/files/eHealth%20TRENDBAROMETER%20-%20HIMSS%20Analytics%20Annual%20European%20eHealth%20Survey%202019.pdf.

HIMSS Analytics. (2021). *HIMSS Annual European Digital Health Survey | HIMSS*. (2021, January 20). https://www.himss.org/resources/himss-annual-european-digital-health-survey.

House of Lords' Select AI Committee. (2018). *AI in the UK: ready, willing and able?* https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Hromkovič, J. (2013). *Algorithmics for Hard Problems: Introduction to Combinatorial Optimization, Randomization, Approximation, and Heuristics*. Springer Science & Business Media.

Huang, H., Jiang, M., Ding, Z., & Zhou, M. (2019). Forecasting Emergency Calls With a Poisson Neural Network-Based Assemble Model. https://ieeexplore.ieee.org/document/8632890.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, *15*(1), 41–51.

Huang, Y., Bert, C., Sommer, P., Frey, B., Gaipl, U., Distel, L. V., Weissmann, T., Uder, M., Schmidt, M. A., Dörfler, A., Maier, A., Fietkau, R., & Putz, F. (2022). Deep learning for brain metastasis detection and segmentation in longitudinal MRI data. Medical Physics, 49(9), 5773–5786. https://doi.org/10.1002/mp.15863.

IBM. (n.d.-a). *What is Deep Learning? | IBM*. Retrieved 26 May 2023, from https://www.ibm.com/topics/deep-learning.

IBM. (n.d.-b). *What is Strong AI? | IBM*. Retrieved 26 May 2023, from https://www.ibm.com/topics/strong-ai.

IBM. (2023). *AI Ethics*. https://www.ibm.com/impact/ai-ethics.

IBM. (2023, May 25). *Generative AI: The state of the market*. IBM. https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/generative-ai-data-story.

IBM Technology (Director). (2023, April 14). *Risks of Large Language Models (LLM)*. https://www.youtube.com/watch?v=r4kButlDLUc.

IMDRF SaMD Working Group. (2013). Software as a Medical Device (SaMD): Key definitions. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf.

IMDRF SaMD Working Group. (2015). Software as a Medical Device (SaMD): Application of Quality Management System. https://www.imdrf.org/sites/default/files/2021-09/imdrf-cons-samd-aqms-150326.pdf.

Imperial College London. (2017). DeepMedic – BioMedIA. https://biomedia.doc.ic.ac.uk/software/deepmedic/.

Infosys. (2018). AI for Healthcare: Balancing Efficiency and Ethics. *AMPLIFYING HUMAN POTENTIAL - TOWARDS PURPOSEFUL ARTIFICIAL INTELLIGENCE*.

INOV. (2019). *KnowLogis—GLINTT | INOV*. https://inovglintt.com/projetos/knowlogis/.

Ischemic Stroke Lesion Segmentation Challenge. (2015). *ISLES: Ischemic Stroke Lesion Segmentation Challenge 2015*. Www.isles-Challenge.org/ISLES2015/. http://www.isles-challenge.org/ISLES2015/.

Islam, M. S., Hasan, M. M., Wang, X., Germack, H. D., & Noor-E-Alam, M. (2018). A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining. *Healthcare*, *6*(2), 54. https://doi.org/10.3390/healthcare6020054.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.

Joint Research Centre (European Commission), Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). *AI watch: Defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence*. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/382730.

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., & Glocker, B. (2015). Multi-Scale 3D Convolutional Neural Networks for Lesion Segmentation in Brain MRI.

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., & Glocker, B. (2016). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78. https://doi.org/10.1016/j.media.2016.10.004.

Kaplan, J. (2016). *Artificial intelligence: What everyone needs to know*. Oxford University Press.

Kavlakoglu, E. (2022, January 19). *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?* https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.

Khamis, A. (2023). Optimization Algorithms—AI techniques for design, planning, and control problems. Manning. https://www.manning.com/books/optimization-algorithms.

Kidd Lab (Director). (2020, January 19). *Celeste Kidd | How to Know | NeurIPS 2019 [Full talk with slides]*. https://www.youtube.com/watch?v=MX5cqgUVkQE.

Kok, J. N. (n.d.). Artificial Intelligence: Definition, Trends, Techniques and Cases. *ARTIFICIAL INTELLIGENCE*.

Koskela, T.-H., Ryynanen, O.-P., & Soini, E. J. (2010). Risk factors for persistent frequent use of the primary health care services among frequent attenders: A Bayesian approach. *Scandinavian Journal of Primary Health Care*, *28*(1), 55–61. https://doi.org/10.3109/02813431003690596.

Kulikowski, C. A. (2015). An Opening Chapter of the First Generation of Artificial Intelligence in Medicine: The First Rutgers AIM Workshop, June 1975. *Yearbook of Medical Informatics*, *10*(1), 227–233. https://doi.org/10.15265/IY-2015-016.

Kusiak, A., & Chen, M. (1988). Expert systems for planning and scheduling manufacturing systems. *European Journal of Operational Research*, *34*(2), 113–130. https://doi.org/10.1016/0377-2217(88)90346-3.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. Zenodo. https://doi.org/10.5281/ZENODO.3240529.

Lev-Ari, A. (2013, May 13). *Vinod Khosla: "20% doctor included": speculations & musings of a technology optimist or "Technology will replace 80% of what doctors do"*. Leaders in Pharmaceutical Business Intelligence (LPBI) Group. https://pharmaceuticalintelligence.com/2013/05/13/vinod-khosla-20-doctor-included-speculations-musings-of-a-technology-optimist-or-technology-will-replace-80-of-what-doctors-do/.

Lippert, F., Cheng, B., Golsari, A., Weiler, F., Gregori, J., Thomalla, G., & Klein, J. (2018). Exploring DeepMedic for the purpose of segmenting white matter hyperintensity lesions. 10575, 105752F. https://doi.org/10.1117/12.2292809.

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., Beam, A. L., Chan, A.-W., Collins, G. S., Deeks, A. D. J., ElZarrad, M. K., Espinoza, C., Esteva, A., Faes, L., Ferrante di Ruffano, L., Fletcher, J., Golub, R., Harvey, H., Haug, C., … Yau, C. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. The Lancet Digital Health, 2(10), e537–e548. https://doi.org/10.1016/S2589-7500(20)30218-1.

Lohr, S. (2012, October 30). *I.B.M.'s Watson Goes to Medical School*. Bits Blog. https://archive.nytimes.com/bits.blogs.nytimes.com/2012/10/30/i-b-m-s-watson-goes-to-medical-school/.

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift Für Medizinische Physik*, *29*(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002.

Lutkevich, B. (2020, August). What is framework? | Definition from TechTarget. WhatIs.Com. https://www.techtarget.com/whatis/definition/framework.

Lynch, S. (2023, April 3). *2023 State of AI in 14 Charts*. Stanford HAI. https://hai.stanford.edu/news/2023-state-ai-14-charts.

Madelin, G., & Lahrichi, N. (2021). Modeling and improving the logistic distribution network of a hospital. *International Transactions in Operational Research*, *28*(1), 70–90. https://doi.org/10.1111/itor.12697.

Magrabi, F., Ammenwerth, E., McNair, J. B., De Keizer, N. F., Hyppönen, H., Nykänen, P., Rigby, M., Scott, P. J., Vehko, T., Wong, Z. S.-Y., & Georgiou, A. (2019). Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications: A Position Paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. *Yearbook of Medical Informatics*, *28*(01), 128–134. https://doi.org/10.1055/s-0039-1677903.

Maier, O., Menze, B. H., von der Gablentz, J., Ḧani, L., Heinrich, M. P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Götz, M., Haeck, T., Halme, H.-L., Havaei, M., … Reyes, M. (2017). ISLES 2015—A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis, 35, 250–269. https://doi.org/10.1016/j.media.2016.07.009.

Malagon-Maldonado, G. (2014). Qualitative Research in Health Design. HERD: Health Environments Research & Design Journal, 7(4), 120–134. https://doi.org/10.1177/193758671400700411.

Manquinho, V., Tiam-Lee, T. J., Henriques, R., Costa, J., & Galhardas, H. (2022). Consolidation of massive medical emergency events with heterogeneous situational context data sources.

Marmé, P. (2018, October 24). 3,8 milhões de euros para levar Inteligência Artificial aos serviços públicos. Motor 24. https://www.motor24.pt/sites/welectric/38-milhoes-euros-levar-inteligencia-artificial-aos-servicos-publicos/433688/.

Marques, I., Captivo, M. E., & Barros, N. (2019). Optimizing the master surgery schedule in a private hospital. Operations Research for Health Care, 20, 11–24. https://doi.org/10.1016/j.orhc.2018.11.002.

Marturano, A. (2002). The role of metaethics and the future of computer ethics. *Ethics and Information Technology*, *4*, 71–78. https://doi.org/10.1023/A:1015202319899.

Mayoh, B. (1994). Constraint Programming and Artificial Intelligence. In B. Mayoh, E. Tyugu, & J. Penjam (Eds.), *Constraint Programming* (pp. 17–50). Springer. https://doi.org/10.1007/978-3-642-85983-0_2.

McCarthy, J. (n.d.). *WHAT IS ARTIFICIAL INTELLIGENCE?.*

Melichov, M. (2022, September 27). Time Series Anomaly Detection With LSTM AutoEncoder. Medium. https://medium.com/@maxme006/time-series-anomaly-detection-with-lstm-autoencoder-b13a4177e241.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., … Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging, 34(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694.

Metz, C., & Schmidt, G. (2023, March 29). Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society'. *The New York Times*. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Greg Miller, W., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S., & Knaus, W. A. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in Biology and Medicine*, *36*(12), 1351–1377. https://doi.org/10.1016/j.compbiomed.2005.08.003.

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics*, *22*(1), 14. https://doi.org/10.1186/s12910-021-00577-8.

Naik, N., Hameed, B. M. Z., Shetty, D. K., Swain, D., Shah, M., Paul, R., Aggarwal, K., Ibrahim, S., Patil, V., Smriti, K., Shetty, S., Rai, B. P., Chlosta, P., & Somani, B. K. (2022). Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Frontiers in Surgery*, *9*, 862322. https://doi.org/10.3389/fsurg.2022.862322.

Natrup, S. (2022). *The Landscape of Artificial Intelligence Ethics: Analysis of Developments, Challenges, and Comparison of Different Markets* [MSc Thesis]. https://run.unl.pt/bitstream/10362/134702/1/TGI0568.pdf.

Ng, M. Y., Kapur, S., Blizinsky, K. D., & Hernandez-Boussard, T. (2022). The AI life cycle: A holistic approach to creating ethical AI for health decisions. *Nature Medicine*, *28*(11), 2247–2249. https://doi.org/10.1038/s41591-022-01993-y.

Nicolae, M.-I., Sinn, M., Tran, M., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., & Edwards, B. (2023). Adversarial Robustness Toolbox (ART) v1.16 (1.16.0) [Python]. IBM. https://github.com/Trusted-AI/adversarial-robustness-toolbox (Original work published 2018).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342.

Oliveira, E. (2020, May 30). *INESC TEC ajuda a prevenir rotura de stocks em hospitais e farmácias*. Notícias U.Porto. https://noticias.up.pt/inesc-tec-ajuda-a-prevenir-rotura-de-stocks-em-hospitais-e-farmacias/.

Open Data Science (Director). (2020, January 16). *Michael I. Jordan on the Future of AI and Machine Learning (Full Video) | ODSC West 2019*. https://www.youtube.com/watch?v=SRF4bXKOGSI.

OpenAI. (2023). *About*. https://openai.com/about.

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY*, *35*(4), 927–936. https://doi.org/10.1007/s00146-020-00965-5.

Oun, M. A., & Bach, C. (2014). *Qualitative Research Method Summary*. *1*(5).

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon Emissions and Large Neural Network Training. https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf.

Paul, Y., Hickok, E., Sinha, A., Tiwari, U., & Bidare, P. M. (2018). *Artificial Intelligence in the Healthcare Industry in India*.

Pearson, D. (2023, April 24). *Moderna throws in with IBM over quantum computing, generative AI for healthcare*. https://aiin.healthcare/topics/artificial-intelligence/moderna-throws-ibm-over-quantum-computing-generative-ai-healthcare.

Peixoto, H., Guimarães, T., & Santos, M. F. (2020). A New Architecture for Intelligent Clinical Decision Support for Intensive Medicine. *Procedia Computer Science*, *170*, 1035–1040. https://doi.org/10.1016/j.procs.2020.03.077.

Pengtao Xie (Director). (2021, February 10). *AAAI21 Workshop—Trustworthy AI for Healthcare*. https://www.youtube.com/watch?v=mJK53b150eM.

Pérez Malla, C. U., Valdés Hernández, M. del C., Rachmadi, M. F., & Komura, T. (2019). Evaluation of Enhanced Learning Techniques for Segmenting Ischaemic Stroke Lesions in Brain Magnetic Resonance Perfusion Images Using a Convolutional Neural Network Scheme. *Frontiers in Neuroinformatics*, *13*. https://www.frontiersin.org/articles/10.3389/fninf.2019.00033.

Perrault, R., Clark, J., Wald, R., Shoham, Y., Parli, V., Niebles, J. C., Ngo, H., Manyika, J., Lyons, T., Ligett, K., Etchemendy, J., Brynjolfsson, E., Fattorini, L., & Maslej, N. (2023, April). *AI Index Report 2023 – Artificial Intelligence Index*. https://aiindex.stanford.edu/report/.

Pfohl, S. R., Foryciarz, A., & Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, *113*, 103621. https://doi.org/10.1016/j.jbi.2020.103621.

Pope, C., & Mays, N. (Eds.). (2020). *Qualitative research in health care* (Fourth edition). Wiley-Blackwell.

Portugal, A. (2030). *AI PORTUGAL 2030 PORTUGUESE NATIONAL INITIATIVE ON DIGITAL SKILLS*.

Prakash, S., Balaji, J. N., Joshi, A., & Surapaneni, K. M. (2022). Ethical Conundrums in the Application of Artificial Intelligence (AI) in Healthcare—A Scoping Review of Reviews. *Journal of Personalized Medicine*, *12*(11), 1914. https://doi.org/10.3390/jpm12111914.

Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Radley-Gardner, O., Beale, H., & Zimmermann, R. (Eds.). (2016). Fundamental Texts On European Private Law. Hart Publishing. https://doi.org/10.5040/9781782258674.

Rais, A., Alvelos, F., Figueiredo, J., & Nobre, A. (2018). Optimization of logistics services in hospitals. *International Transactions in Operational Research*, *25*(1), 111–132. https://doi.org/10.1111/itor.12370.

Rangasamy, S., Nadenichek, R., Rayasam, M., & Sozdatelev, A. (2018, December 6). *Natural language processing in healthcare | McKinsey*. https://www.mckinsey.com/industries/healthcare/our-insights/natural-language-processing-in-healthcare.

Rehme, A. K., Volz, L. J., Feis, D.-L., Bomilcar-Focke, I., Liebig, T., Eickhoff, S. B., Fink, G. R., & Grefkes, C. (2015). Identifying Neuroimaging Markers of Motor Disability in Acute Stroke by Machine Learning Techniques. *Cerebral Cortex*, *25*(9), 3046–3056. https://doi.org/10.1093/cercor/bhu100.

Reis, J., Santo, P., & Melão, N. (2020). Impact of Artificial Intelligence Research on Politics of the European Union Member States: The Case Study of Portugal. *Sustainability*, *12*(17), 6708. https://doi.org/10.3390/su12176708.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). '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*, 1135–1144. https://doi.org/10.1145/2939672.2939778.

Rigby, M. (2019). AMA Journal of Ethics ® FROM THE EDITOR Ethical Dimensions of Using Artificial Intelligence in Health Care. *AMA Journal of Ethics, 21(2)*, 121–124. https://journalofethics.ama-assn.org/sites/journalofethics.ama-assn.org/files/2019-01/fred1-1902_1.pdf.

Rish, I. (2001). *An empirical study of the naive Bayes classifier*.

Rokach, L., & Maimon, O. (2005). Decision Trees. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). Springer US. https://doi.org/10.1007/0-387-25465-X_9.

Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., & Summers, R. M. (2014). A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations. In P. Golland, N. Hata, C. Barillot, J. Hornegger, & R. Howe (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014* (pp. 520–527). Springer International Publishing. https://doi.org/10.1007/978-3-319-10404-1_65.

Saenger, A. K., & Christenson, R. H. (2010). Stroke Biomarkers: Progress and Challenges for Diagnosis, Prognosis, Differentiation, and Treatment. *Clinical Chemistry*, *56*(1), 21–33. https://doi.org/10.1373/clinchem.2009.133801.

Sam Altman [@sama]. (2023, February 1). *We know that ChatGPT has shortcomings around bias, and are working to improve it. But directing hate at individual OAI employees because of this is appalling. Hit me all you want, but attacking other people here doesn't help the field advance, and the people doing it know that.* [Tweet]. Twitter. https://twitter.com/sama/status/1620927983627427840.

Samuel, A. L. (n.d.). *Some studies in machine learning using the game of checkers*.

Santos, N. (2023, March 12). Projeto piloto nas urgências permite ao hospital "conhecer" doente enquanto este é transportado pelo INEM | TVI Jornal. Televisão Independente (TVI). https://tviplayer.iol.pt/programa/tvi-jornal/63ef5eb50cf2665294d5f87a/video/640dce570cf2cf9224fd0e79.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & Lecun, Y. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. International Conference on Learning Representations (ICLR) (Banff).

Setzler, H., Saydam, C., & Park, S. (2009). EMS call volume predictions: A comparative study. Computers & Operations Research, 36(6), 1843–1851. https://doi.org/10.1016/j.cor.2008.05.010.

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), Article 12. https://doi.org/10.1038/s41591-021-01595-0.

Shahid, N., Rappon, T., & Berta, W. (2019). Applications of artificial neural networks in health care organizational decision-making: A scoping review. PLOS ONE, 14(2), e0212356. https://doi.org/10.1371/journal.pone.0212356.

Shaw, J. A., Sethi, N., & Block, B. L. (2021). Five things every clinician should know about AI ethics in intensive care. *Intensive Care Medicine*, *47*(2), 157–159. https://doi.org/10.1007/s00134-020-06277-y.

Shiraishi, J., Li, Q., Appelbaum, D., & Doi, K. (2011). Computer-Aided Diagnosis and Artificial Intelligence in Clinical Imaging. *Seminars in Nuclear Medicine*, *41*(6), 449–462. https://doi.org/10.1053/j.semnuclmed.2011.06.004.

Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975a). Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, *8*(4), 303–320. https://doi.org/10.1016/0010-4809(75)90009-9.

Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975b). Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, *8*(4), 303–320. https://doi.org/10.1016/0010-4809(75)90009-9.

SICAS Medical Image Repository. (n.d.). ISLES2015 - SICAS Medical Image Repository. Www.smir.ch. Retrieved September 19, 2023, from https://www.smir.ch/ISLES/Start2015.

Simons, H. (2009). Case Study Research in Practice. SAGE.

Singer, P. (2023, July 11). *Ethics—Machiavelli, Morality, Politics | Britannica*. https://www.britannica.com/topic/ethics-philosophy.

Skirpan, M., & Yeh, T. (2017). Designing a Moral Compass for the Future of Computer Vision Using Speculative Analysis. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1368–1377. https://doi.org/10.1109/CVPRW.2017.179.

Sonnenberg, F. A., Hagerty, C. G., & Kulikowski, C. A. (1994). An Architecture for Knowledge-based Construction of Decision Models. *Medical Decision Making*, *14*(1), 27–39. https://doi.org/10.1177/0272989X9401400104.

Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The Ethics of Computing: A Survey of the Computing-Oriented Literature. *ACM Computing Surveys*, *48*(4), 1–38. https://doi.org/10.1145/2871196.

Stanfill, M. H., & Marc, D. T. (2019). Health Information Management: Implications of Artificial Intelligence on Healthcare Data and Information Management. *Yearbook of Medical Informatics*, *28*(01), 056–064. https://doi.org/10.1055/s-0039-1677913.

Stanford University HAI. (2023). *Artificial Intelligence Index Report 2023 Introduction to the AI Index Report 2023*. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf.

Sujan, M., Furniss, D., Grundy, K., Grundy, H., Nelson, D., Elliott, M., White, S., Habli, I., & Reynolds, N. (2019). Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health & Care Informatics*, *26*(1), e100081. https://doi.org/10.1136/bmjhci-2019-100081.

Sunarti, S., Fadzlul Rahman, F., Naufal, M., Risky, M., Febriyanto, K., & Masnina, R. (2021). Artificial intelligence in healthcare: Opportunities and risk for future. *Gaceta Sanitaria*, *35*, S67–S70. https://doi.org/10.1016/j.gaceta.2020.12.019.

Syntetos, M., Boylan, J., & Croston, J. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society*, *56*. https://doi.org/10.1057/palgrave.jors.2601841.

Tai, M. C.-T. (2020). The impact of artificial intelligence on human society and bioethics. *Tzu-Chi Medical Journal*, *32*(4), 339–343. https://doi.org/10.4103/tcmj.tcmj_71_20.

Tamir, M. (2020, June 26). *What Is Machine Learning? - I School Online*. UCB-UMT. https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/.

Taylor, J., & Hern, A. (2023, May 2). 'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation. *The Guardian*. https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning.

Topol, E. (2019). *Preparing the healthcare workforce to deliver the digital future An independent report on behalf of the Secretary of State for Health and Social Care*. https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf.

Tracxn. (2023, March 30). AI in Healthcare Startups in Portugal. https://tracxn.com/explore/AI-in-Healthcare-Startups-in-Portugal.

Treasury Board of Canada Secretariat. (2021, March 22). *Algorithmic Impact Assessment Tool* [Guidance]. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

Turing, A. M. (1950). *Computing Machinery and Intelligence. Mind, LIX(236)*, 433–460. https://doi.org/10.1093/mind/lix.236.433.

UK Central Digital and Data Office. (2020). *Data Ethics Framework*.

Umoren, I. J., E., U., P., A., & C., K. (2021). Healthcare Logistics Optimization Framework for Efficient Supply Chain Management in Niger Delta Region of Nigeria. *International Journal of Advanced Computer Science and Applications*, *12*(4). https://doi.org/10.14569/IJACSA.2021.0120475.

UNESCO Digital Library. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000380455.

van de Sande, D., Van Genderen, M. E., Smit, J. M., Huiskens, J., Visser, J. J., Veen, R. E. R., van Unen, E., BA, O. H., Gommers, D., & van Bommel, J. (2022). Developing, implementing and governing artificial intelligence in medicine: A step-by-step approach to prevent an artificial intelligence winter. *BMJ Health & Care Informatics*, *29*(1), e100495. https://doi.org/10.1136/bmjhci-2021-100495.

Vanian, J. (2018). *Unmasking A.I.'s Bias Problem*. Fortune. https://fortune.com/longform/ai-bias-problem/.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689.

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). *Heart Diseases Detection Using Naive Bayes Algorithm*. *2*(9).

Vogel, L. (2017). What "learning" machines will mean for medicine. *Canadian Medical Association Journal*, *189*(16), E615–E616. https://doi.org/10.1503/cmaj.1095413.

Volkov, M., Hashimoto, D. A., Rosman, G., Meireles, O. R., & Rus, D. (2017). Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 754–759. https://doi.org/10.1109/ICRA.2017.7989093.

Whitby, B. (n.d.). *The Ethical Implications of Non-human Agency in Health Care*.

Winters, B., Custer, J., Galvagno, S., Colantuoni, E., Kapoor, S., Lee, H., Goode, V., Robinson, K., Nakhasi, A., Pronovost, P., & Newman-Toker, D. (2012). Diagnostic errors in the intensive care unit: A systematic review of autopsy studies. *BMJ Quality & Safety*, *21*, 894–902. https://doi.org/10.1136/bmjqs-2012-000803.

Wirtz, B., Weyerer, J., & Sturm, B. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, *43*. https://doi.org/10.1080/01900692.2020.1749851.

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal Medicine, 170(1), 51–58. https://doi.org/10.7326/M18-1376.

Wolford, B. (2018, November 7). *What is GDPR, the EU's new data protection law?* GDPR.Eu. https://gdpr.eu/what-is-gdpr/.

World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health Ethics And Governance of Artificial Intelligence for Health*.

Zhou, X., Chen, S., Liu, B., Zhang, R., Wang, Y., Li, P., Guo, Y., Zhang, H., Gao, Z., & Yan, X. (2010). Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artificial Intelligence in Medicine*, *48*(2–3), 139–152. https://doi.org/10.1016/j.artmed.2009.07.012.

APPENDIX

Table 1 - Mapping of ethical topics to isolate those most referred to in the analysed literature.

	Title	Year	Author(s)	Keywords	Ethic's Topics Frequence									
					Accountability	Transparency & Explainability	Bias (*1)	Value-upholding use (*2)	Privacy (*3)	Trustworthy & Responsible (*4)	Workforce Impact & Sustainability/ Educational Rights	Security (*5)	Sustainability	Overconfidence
1	The impact of artificial intelligence on human society and bioethics	2020	Tai, M.	Artificial Intelligence; Bioethics; Principles of artificial intelligence bioethics	X	X		X						
2	Artificial Intelligence in the Healthcare Industry in India	2018	Paul, Y. et al.	-	X	X	X	X	X	X				
3	The Landscape of Artificial Intelligence Ethics: Analysis of Developments, Challenges, and Comparison of Different Markets	2022	Natrup, S.	Machine Learning Ethics; Artificial Intelligence Ethics; Machine Learning Bias; Machine Learning Discrimination; Trustworthy AI; Artificial Intelligence Principles; Artificial Intelligence Guidelines			X			X				
4	AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media	2020	Ouchchy, L., et al.	Artificial intelligence; Ethics; Media; News; Public discourse; Public policy		X		X	X	X	X	X		
5	Artificial intelligence for good health: a scoping review of the ethics literature	2021	Murphy, K., et al.	Artificial intelligence; Ethics; Healthcare; Public and population health; Global health	X		X		X	X		X		
6	Developing, implementing and governing artificial intelligence in medicine: a step-by- step approach to prevent an artificial intelligence winter	2022	van de Sande, D., et al.	-		X	X	X						
7	Digital Technologies and Data Science as Health Enablers: An Outline of Appealing Promises and Compelling Ethical, Legal, and Social Challenges	2021	Cordeiro, J.	Digital Health; Ethics; Law; Artificial Intelligence; Telemedicine; Big Data; Patient–doctor Relationship		X	X	X	X	X		X		
8	Ethical and legal challenges of artificial intelligence-driven healthcare	2020	Gerke, S., et al.	-		X	X		X			X		
9	Health Information Management: Implications of Artificial Intelligence on Healthcare Data and Information Management	2019	Stanfill, M., et al.	Artificial Intelligence; Health Information Management; Automation; Medical coding; Health workforce	X				X	X				
10	The Topol Review: Preparing the healthcare workforce to deliver the digital future	2019	NHS	-			X	X			X	X		
11	The Battle to Embed Ethics in AI Research	2020	Gibney, E.	-			X							
12	The Ethical Implications of Non-human Agency in Health Care	n.d.	Whitby, B.	-	X					X	X			X
13	What "learning" machines will mean for medicine	2017	Vogel, L.	-		X								
14	How to Know	2019	Kidd, C.	-										X
15	The potential for artificial intelligence in healthcare	2019	Davenport, T., et al.	Artificial Intelligence; Clinical Decision Support; Electronic Health Record Systems	X	X			X					
16	Artificial intelligence in health care: medical, legal and ethical challenges ahead	2020	Böke, S.	-	X	X	X		X	X	X	X		
17	Artificial intelligence in healthcare: opportunities and risk for future	2020	Sunarti, S., et al.	Artificial Intelligence; Healthcare; Opportunities; Risk	X	X	X		X	X				
18	Ethical Dimensions of Using Artificial Intelligence in Health Care	2019	Rigby, M.	-		X		X	X					
19	Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector	2019	Leslie, D.	-	X	X	X					X	X	
20	A scoping review of ethics considerations in clinical natural language processing	2022	Bear Don't Walk, et al.	Natural Language Processing; Bias; Fairness; Ethically informed			X							
21	Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications	2019	Magrabi et al.	Artificial Intelligence; Machine Learning; Clinical Decision Support, Evaluation studies; Program evaluation		X	X			X		X		
22	The AI life cycle: a holistic approach to creating ethical AI for health decisions	2022	Ng, M., et al.	-	X	X	X							
23	The Ethics of Computing: A Survey of the Computing-Oriented Literature	2016	Stahl et al.	Computer Ethics; Information Ethics; Responsible Research and Innovation	X			X	X	X				
24	Data Ethics Framework	2020	Government Digital Service	-	X	X	X							
25	Ethical Conundrums in the Application of Artificial Intelligence (AI) in Healthcare—A Scoping Review of Reviews	2022	Prakash, S., et al.	Artificial Intelligence; Machine Learning; Deep Learning; Ethics; Medical Ethics; Ethical complications; Autonomy; Artificial Intelligence in Healthcare; Legal and Ethical guidelines; Application in Healthcare	X	X	X		X	X				
26	Identifying Ethical Considerations for Machine Learning Healthcare Applications	2020	Char, D., et al	Machine learning; Artificial Intelligence; Safety; Effectiveness; Test characteristics; Ethics	X	X		X	X					
27	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	n.d.	Diakopoulos, N., et al.	-	X	X	X			X				
28	AI Ethics' Pillars	2023	IBM	-		X	X		X			X		
29	State of Ethics and Trust in Technology	2022	Deloitte	-	X	X	X		X	X		X		
30	The Ethics of AI Ethics: An Evaluation of Guidelines	2020	Hagendorff, T.	Artificial intelligence; Machine learning; Ethics; Guidelines; Implementation	X	X		X	X			X		
31	AI in the UK: ready, willing and able	2018	House of Lords, Select Committee on Artificial Intelligence	-			X	X	X		X			
32	Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?	2022	Naik, N., et al.	Artificial Intelligence; Machine Learning; Ethical issues; Legal issues; Social issues		X	X		X	X				
33	Ethics and Governance of Artificial Intelligence for Health	2021	WHO	-	X	X	X	X		X		X	X	
34	Recommendation on the Ethics of Artificial Intelligence	2021	UNESCO	-	X	X	X	X	X	X	X	X		
35	Unboxing Artificial Intelligence: 10 steps to protect Human Rights	2019	Council of Europe	-		X	X		X		X			
36	The global landscape of AI ethics guidelines	2019	Jobin, A, et al.	-		X	X	X	X	X				
37	How Should AI Be Developed, Validated, and Implemented in Patient Care?	2019	Anderson, M., et al.	-		X	X							
38	Can AI Help Reduce Disparities in General Medical and Mental Health Care?	2019	Chen, I., et al.	-			X							
39	Declaration for a Responsible Development of Artificial Intelligence	2018	Université de Montréal	-			X	X	X	X		X		
40	2023 Landscape	2023	AI Now Institute	-	X				X	X	X		X	
41	Machine learning in medicine: Addressing ethical challenges	2018	Vayena, E., et al.	-	X	X	X		X	X				
42	Five things every clinician should know about AI ethics in intensive care	2021	Shaw, J. A., et al.	-		X	X			X		X		
43	The practical implementation of artificial intelligence technologies in medicine	2019	He, J., et al.	-		X			X			X		
					21	30	30	15	25	22	8	16	3	2

(*1) This column, intitled "Bias", includes mentions of the following topics: Bias; Inclusiveness; Equity; Discrimination; Exclusion; Fairness; Inequalities; (Social) Justice; Impartiality; Unrepresentativeness.

(*2) This column, intitled "Value-upholding use", reflects in particular mentions to human's autonomy.

(*3) This column, intitled "Privacy", includes mentions of the following topics: (Patient/Data) Privacy; Consent (to use health/patients' data).

(*4) This column, intitled "Trustworthy & Responsible", includes mentions of the following topics: Liability; Responsability; Reliability; Trustworthy.

(*5) This column, intitled "Security", includes mentions of the following topics: Security; Safety; Robustness.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa