NEW PATHWAYS OF APPLIED ETHICS

Leslye Denisse Dias Duran

Machine Learning in Medical Diagnosis

A Framework for a Normative Evaluation of Benefits and Risks

1682

J.B. METZLER

Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics

In der Ethik beschäftigte man sich lange Zeit nicht mit konkreten moralischen Fragestellungen. Die Angewandte Ethik gehört deshalb zu den wichtigen neueren Entwicklungen der normativen Ethik. Sie vermag dann überzeugende Beiträge zu leisten, wenn sie sich mit Problemstellungen befasst, die sich gut ein- und abgrenzen lassen, was die relevanten Sachfragen und normativen Anforderungen anbelangt. Dies ist aber bei den großen Herausforderungen unserer Zeit, wie z. B. Klimawandel, Weltarmut und Migration, Bewältigung von Pandemien, der fehlenden Nachhaltigkeit des Wirtschaftswachstums, der digitalen Revolution oder der Fragilität der Finanzmärkte, nicht gegeben. Die Komplexität der vorliegenden Sachverhalte überfordert oftmals unsere Wissensmöglichkeiten. Mit dieser Entwicklung scheinen die Methoden der Angewandten Ethik nicht mitgehalten zu haben. Es wird mit begründungslogisch bescheidenen Ansätzen gearbeitet, sofern man nicht ganz auf philosophische Theoriebildungen zugunsten von allein empirisch ausgerichteten Untersuchungen verzichtet. Die Reihe „Neue Wege der angewandten Ethik“ füllt diese Lücke. Sie ist ein Ort für Monographien und Sammelbände, die diese Themen sowie die damit zusammenhängenden Grundlagen- und Methodenprobleme gezielt angehen und so exemplarisch zur Lösung der komplexen Herausforderungen unserer Zeit beizutragen versuchen. Dabei sollen grundbegriffliche und begründungslogische Aspekte nicht ausgespart, sondern als wesentlicher Teil einer auf Anwendungskontexte reflektierenden Ethik ernst genommen werden: einer Angewandten Ethik, die das Reflexionspotenzial der normativen Ethik nutzt und damit auf der Höhe der aktuellen Problemlagen argumentiert.

New Pathways of Applied Ethics

For a long time, ethics was not concerned with concrete moral questions. Applied ethics therefore constitutes one of the more important recent developments of normative ethics. So far it has been able to make convincing contributions if it is dealing with topics which are clearly circumscribed regarding the involved factual and normative issues. However, this is not the case with respect to the great challenges of our times, such as, for example, climate change, world poverty and migration, the prevention and overcoming of pandemics, the unsustainability of economic growth, the challenges of the digital revolution, or the fragility of the financial markets. The complexity of pertinent factual issues within these domains outstretches our abilities to know. The methods of applied ethics have not kept pace with these developments and challenges. The justificatory approaches that are chosen tend to be plain and modest, and often there is even a complete abandonment of philosophical theorizing in favor of purely empirically oriented investigations. The book series New Pathways of Applied Ethics is prepared to overcome these shortcomings. It is a place for treatises and edited volumes devoted to challenging topics and connected foundational and methodological problems. It thus aims to contribute to the solution of the great challenges of our times while not eschewing the investigation of basic concepts and justificatory possibilities. Rather, they will be taken up as an indispensable part of an ethics that reflects on complex contexts and applications. It is an applied ethics that draws upon the full potential of philosophical reflection to address current problems and situations.

Leslye Denisse Dias Duran

Machine Learning in Medical Diagnosis

A Framework for a Normative Evaluation of Benefits and Risks

J.B. METZLER

Leslye Denisse Dias Duran
Ruhr Universität Bochum
Bochum, Germany

Approved dissertation for the award of the academic degree of Doctor of Philosophy.
Faculty of Philosophy and Educational Research.
Ruhr University Bochum.
First reviewer: Prof. Dr. Klaus Steigleder.
Second reviewer: Prof. Dr. Philipp Richter.
Disputation on 04.12.2024

ISSN 3059-2550 ISSN 3059-2569 (electronic)
Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics
ISBN 978-3-662-71356-3 ISBN 978-3-662-71357-0 (eBook)
https://doi.org/10.1007/978-3-662-71357-0

This J.B. Metzler imprint is published by the registered company Springer-Verlag GmbH, DE, part of Springer Nature.
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

Acknowledgements Throughout these -almost- four years of working on this dissertation, I learned quite a lot of things from my craft, from my field, and from myself. However, the one that will perhaps stay with me for the rest of my life is the humbling power of the relationships we form with others. While this journey was often lonely, I was never left to fend for myself in the moments when I did not know what to do or how to move forward. Every time I reached out for help, there were people willing to lend me a hand in many ways. I truly could not have completed this step into becoming a doctor on my own.

The first person I want to express my deepest gratitude to is my first supervisor and mentor, Klaus Steigleder. I honestly do not know what blessing of the universe led me to have him as my academic guide throughout this process. He not only supervised my master's thesis but invited me to make a PhD at his chair at the Ruhr University. He encouraged me to be independent but also was always available to listen to my bouts of ranting, to read every single piece of writing I produced and to write every report and recommendation letter I needed (which were quite a lot, to be honest). However, most importantly, he also became a discussion partner. At his office and surrounded by books, I had some of the most interesting discussions about philosophy (and a lot else) I have ever had. Every time I left, I was full of ideas and went back to my writing, feeling recharged and motivated. I truly believe that a graduate student's path depends a great deal on their advisor, and Professor Steigleder was the best one could hope for.

I want to thank Professor Phillip Richter, who accepted to be my second supervisor, and who made insightful comments and gave me quite helpful feedback on my work. In particular, I am thankful for his encouragement to make the diagrams in Chap. 5 to visualize the intricate nature of a relational, rights-based approach. It was a challenging and fun process, and I learned to see problems from a different perspective. I also thank the late Professor Christoph Bambauer, who helped me with recommendation letters and reports for the application as a doctoral student and to several funding institutions.

I am also indebted to the members of the doctoral colloquium at the chair of Applied Ethics of the Institute of Philosophy I, who enthusiastically read many pages about machine learning and medical diagnosis and always gave me insightful feedback and raised questions that helped me shape the philosophical chapters of the dissertation. To Raoul Scheppat, Alina Pfleghardt, Julia Weinheimer, Tobias Vogel, Andrea Lautenschläger, Vanessa Bieber, Marie Göbel, Johannes Graf Keyserlingk and Friederike Asche, many, many thanks.

I need to thank the team at the international department of the Konrad Adenauer Foundation. The funding I received as an international doctoral student made it financially possible to dedicate my time to the dissertation without worries. In particular, I want to express my gratitude to my referent, Kerim Kudo, who is devoted to support all the international students in every way he can. His patience and kindness are one of a kind. Also, I want to thank the members of the PhD Group of Students of AI Ethics, in particular Bec Johnson, Tricia Griffin, Giada Pistilli, Natalia Menendez and Enrico Panai, with whom I wrote my first academic paper and learned a lot about academic collaboration.

There is a tendency for graduate students to make the dissertation the axis around which every other aspect of their lives revolves. This was often the case for me, and so I would like to thank my parents, Nidia and William, for constantly reminding me to take walks, breathe deeply, and enjoy the process as much as possible. They managed to be present even from hundreds of miles away. I also have to thank my sister, Maria Camila, who was always there to cheer me up and share every experience with me. I also would like to thank my closest friends, Marissa, Estefa, Ceci, Katie, Ricardo, and Victoria, who throughout the years have supported me and cheered for me at every step. I also need to thank the family I found in the second country I count as home: Margaret, Paolo, Katrin, and Tony. You helped in ways you do not even know.

Finally, I would like to thank my fiancée, Giovanni, from the bottom of my heart. This process has been a roller coaster at times and an uphill battle at others, but you have held me in place when I felt like losing my head, and pushed me forward with your humor, kindness, and love when I felt I did not if I could go on. This dissertation is the result of many late nights, library visits, and work weekends that you chose to spend right next to me. You have encouraged me and challenged my ideas. Our discussions while watching the news and taking walks shaped not only the final product, but also the person who wrote it. You helped me get through the days of doubt, celebrated every milestone, and gave me the reassurance that no matter what, there is a way, and we will find it together. I dedicate this dissertation to you.

Competing Interests The author has no competing interests to declare that are relevant to the content of this manuscript.

Introduction

"Regulation without ethics is blind, ethics without regulation is weak."

Klaus Steigleder, 2024 (Alluding to Kant)

What is the role of ethics in applied contexts? This is perhaps one of the first questions that someone working in the field of applied ethics is confronted with, and it is often a difficult question to answer. While other disciplines may have more or less straightforward definitions and boundaries, applied ethics is found at the intersection of other disciplines, which usually means that one has to go into some detail about them before being able to explain the meaning of the ethical work. Applied ethics is a discipline within the field of philosophy. It is not just the description of ethical challenges in concrete contexts, nor is it just conceptual clarification with philosophical rigor. Both tasks can be included in the work of an applied ethicist, but the work that distinguishes it is the normative evaluation of considerations with moral relevance in a particular context.

The emergence of digital technologies like artificial intelligence (AI), their implications and consequences are matters that warrant this kind of philosophical reflection and evaluation. They raise questions that go beyond the rush to comply with regulatory requirements or produce internal governance structures that paint a good company image. These questions require that we reflect upon how we perceive and approach problems of moral significance. Are these technologies going to aid or hinder us to fulfill our moral obligations to others? Other questions confront us about the tradeoffs we are making or will have to make as a result of introducing these technologies in our routines. For instance, what skills, habits or opportunities we might be discarding in favor of the affordances obtained by

using these tools? What are we expecting to gain from it, and is it worth it? In areas of application like in healthcare, the use of AI systems demands that we consider how *we* are changing, and how our ways of relating and connecting with others can be impacted. Are there new moral duties that we owe to other agents because of these technologies? Are there rights that emerge and need protection?

As these technologies become increasingly ubiquitous, as moral agents at the center of any process in which they operate, we have a responsibility to inquire more deeply about the role or function we should assign to them, the effects they can have, and the ways in which we ought to prepare to deal with any potential consequences. These normative tasks are underscored by an appeal to the value of slowing down. Technology companies and developers in the field of artificial intelligence have adopted a pace of development that often does not allow sufficient time to consider the potential risks or drawbacks of their products at different levels. This strains the ability of societies to adapt to the changes, which in some cases are significant, and can lead to a general feeling of fatigue that distracts attention from the potential benefits.

It is a fact that the integration of artificial intelligence in healthcare is considered one of the most promising areas for positive change and societal benefit. Companies and technical experts have suggested for over a decade now that the technical capabilities of AI systems could significantly improve diagnostic accuracy, facilitate early detection of diseases and reduce medical error[1]. As of now, there are also applications approved commercially that have demonstrated evidence of fulfilling these promises through randomized clinical trials[2]. However, the rapid adoption of these technologies still raises critical ethical, technical, and socio-normative concerns that require careful consideration.

There are several research problems that this dissertation aims to address. First, there is a lack of clarity about how to deal with the normative issues raised by the implementation of AI systems. Existing approaches to AI ethics mostly focus on the technical and regulatory challenges and use methodologies based on principles

[1] Thomas Davenport and Ravi Kalakota, "The Potential for Artificial Intelligence in Healthcare," *Future Healthcare Journal* 6, no. 2 (June 2019): 95–97, https://doi.org/10.7861/futurehosp.6-2-94.

[2] Alexander Seager et al., "Polyp Detection with Colonoscopy Assisted by the GI Genius Artificial Intelligence Endoscopy Module Compared with Standard Colonoscopy in Routine Colonoscopy Practice (COLO-DETECT): A Multicentre, Open-Label, Parallel-Arm, Pragmatic Randomised Controlled Trial," *The Lancet Gastroenterology & Hepatology* 9, no. 10 (October 1, 2024): 916–22, https://doi.org/10.1016/S2468-1253(24)00161-4.

or values and how to translate and operationalize them in models[3]. There is common agreement that AI-driven technologies should be used for the common good and not to harm people's rights. However, there is less agreement on *how* to reach these overarching goals. Second, the focus on principles-based approaches has proven largely ineffective on two important tasks of applied ethics: providing actionable advice in situations where there is a conflict between principles and making ethical work more enforceable[4]. Third, there is a marked gap in attitudes about what the concrete aims of these implementation processes should be. As this is an interdisciplinary context, the normative aims of different involved actors converge, but there is no work in the existing literature that analyzes them from a normative perspective.

This dissertation seeks to navigate the rift between the optimism surrounding machine learning (ML) driven healthcare innovations and the urgent need for a robust, normative framework for assessing their risks and benefits. There are two main objectives of this dissertation. First, to make an analysis of the normative implications and considerations that emerge at the intersection of ML models and medical diagnosis; and second, to propose a normative framework for the evaluation of the distribution of benefits and risks, using a rights-based approach together with elements from relational theory.

This framework is conceptualized as an "ecosystem of moral constellations" due to its foundation in relational theory. It identifies the actors and the relationships between them that form at different points of the intersection of AI and healthcare. The analysis conducted at each constellation establishes the normative aspects and goals among these relationships that must be considered for an evaluation of the distribution of benefits and risks of implementing ML models in the diagnostic process. I argue that this normative approach can be well suited to conducting a practical assessment of the distribution of benefits and risks. However, I limit my work in this dissertation to establishing this framework, leaving its application to concrete scenarios of implementation to future work.

---

[3] Jessica Morley et al., "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics* 26, no. 4 (August 1, 2020): 2141–68, https://doi.org/10.1007/s11948-019-00165-5.

[4] Jess Whittlestone et al., "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES'19 (New York, NY, USA: Association for Computing Machinery, 2019), 196–97, https://doi.org/10.1145/3306618.3314289.

The proposed rights-based approach builds on the philosophical work of Klaus Steigleder on a rights-based theory of risk ethics[5] and the moral philosophy of Alan Gewirth[6]. This approach has four features, called "pillars" in the dissertation, that offer a practical way to deal with situations where multiple rights conflict. This approach is complemented by the integration of some elements of relational theory from Vangie Bergum[7] working on the ethics of care in health, Jennifer Nedelsky[8] and Marina Oshana[9] from feminist theories applied to law and ethics, and Abeba Birhane[10] from decolonial approaches to AI ethics. Although these approaches generally criticize and, sometimes even reject, the notion of rights because of its common individualistic interpretation in rights theories, I argue in this dissertation that rights are inherently relational and should be considered from this perspective to evaluate the distribution of benefits and risks of AI technologies in health care.

This approach is further supported by two additional conceptualizations. The first is the definition of AI as a socio-technical system (STS), which considers the risks and benefits of technology from a networked perspective involving different actors, institutions, physical infrastructures, and ethical and regulatory frameworks. This approach emphasizes that the problems and challenges of technology in society cannot be solved by technical fixes (in contrast to techno-solutionist approaches) but require interdisciplinary collaboration and the acceptance of trade-offs rather than complete solutions[11]. Second, understanding the diagnostic

[5] Klaus Steigleder, "Risk and Rights: Towards a Rights-Based Risk Ethics" (Bochum, 2012).

[6] Alan Gewirth, Reason and Morality (Chicago: University of Chicago Press, 1978) and Alan Gewirth, *The Community of Rights* (Chicago, IL: University of Chicago Press, 1996), https://press.uchicago.edu/ucp/books/book/chicago/C/bo3645650.html.

[7] Vangie Bergum, "Discourse—Ethical Challenges of the 21st Century: Attending to Relations," *Canadian Journal of Nursing Research Archive*, 2002, https://cjnr.archive.mcgill.ca/article/view/1761.

[8] Jennifer Nedelsky, "Reconceiving Rights as Relationship," in *Explorations in Difference*, ed. Hart, Jonathan and Bauman, Richard W., 1st ed. (Routledge, 1996), 67–88.

[9] Marina Oshana, "Relational Autonomy," in *International Encyclopedia of Ethics* (John Wiley & Sons, Ltd, 2020), 1–13, https://doi.org/10.1002/9781444367072.wbiee921.

[10] Abeba Birhane, "Algorithmic Injustice: A Relational Ethics Approach," *Patterns* 2, no. 2 (February 2021): 100205, https://doi.org/10.1016/j.patter.2021.100205.

[11] Günter Ropohl and Society for Philosophy and Technology, "Philosophy of Socio-Technical Systems," *Society for Philosophy and Technology Quarterly Electronic Journal* 4, no. 3 (1999): 186–94, https://doi.org/10.5840/techne19994311.

process not just as a cognitive exercise of assigning the right label to a set of symptoms, but as a "situated process", as proposed by Mark L. Graber[12].

Chap. 1 aims to lay the foundation for the main goal of the dissertation, based on the actual capabilities of ML models in medical diagnosis. It is critical to begin with this task because, while ML models can be powerful tools, setting unrealistic expectations -whether for improved efficiency or a significant reduction in medical error- can lead to disappointment or, worse, dangerous consequences for patient care. From the early days of expert systems designed to help doctors diagnose infections and analyze molecular structures to today's state-of-the-art ML algorithms, the journey of AI in healthcare has been one of progress punctuated by a problematic pattern: the hype surrounding AI often outpaces its real-world capabilities. Therefore, a solid foundation on the actual capabilities of ML is crucial. The development of the chapter begins with a first section that traces the evolution of AI and ML in healthcare, highlighting key developments that have shaped the field. The second and third sections focus on describing the basic terms and concepts of AI and healthcare, respectively. The final section discusses the benefits of implementing ML models in diagnostic settings according to these technical possibilities.

Chap. 2 is concerned with presenting and analyzing the justifications of the dissertation. It seeks to answer the questions: why is an assessment of benefits and risks needed at all, and why is a relational, rights-based, approach suitable for this task? In order to provide answers, this chapter is divided into two main sections. The first offers a conceptualization of four normative tensions between disciplines and approaches in AI ethics as four gaps: the epistemic gap, the responsibility gap, the conceptual gap, and the implementation gap. The second section provides an overview of the other most common methodological approaches to AI ethics, highlighting their weaknesses and the reasons why a relational, rights-based approach may be appropriate to address the challenges posed by the implementation of ML models in medical diagnosis.

Chap. 3 builds on the arguments provided in the previous chapter and presents the normative foundations of the proposed rights-based approach and the relational theory used to ground the normative framework developed in Chapter 5. It outlines four "pillars": that everyone has an equal claim to the conditions necessary to be able to lead life, that rights form a hierarchy, that rights are

[12] Mark L. Graber, "Progress Understanding Diagnosis and Diagnostic Errors: Thoughts at Year 10," *Diagnosis* 7, no. 3 (August 27, 2020): 151–59, https://doi.org/10.1515/dx-2020-0055.

simultaneously negative and positive, and that these rights require effective protection. The relational aspect is conceptualized as a transversal "floor" where the pillars rest. The second part of the chapter aims to examine the impact of the emergence of ML in healthcare on existing rights and considers the justifications or issues with proposals to recognize new rights that emerge from these new technologies. In particular, the chapter addresses the right to an explanation and the right to a human decision.

Chap. 4 shifts the focus to the risks associated with ML in medical diagnosis. First, the chapter proposes to consider three phenomena that pose serious risks to the rights of patients and clinicians, as well as to healthcare institutions: diagnostic error rates, demographic change, and workforce shortages. It then examines emerging risks, including discrimination, data breaches, and environmental impacts due to the resource-intensive nature of ML models. The chapter argues that an assessment of the balance of benefits and risks must weigh the challenges posed by existing healthcare problems against the potential risks posed by the implementation of ML models.

Finally, *Chap. 5* synthesizes the clinical, technical, and normative discussions of the previous chapters to propose a relational rights-based framework for evaluating the implementation of ML in medical diagnosis, according to the distribution of benefits and risks. The framework is constructed as an "ecosystem of moral constellations" that conceptualizes the relationships between moral agents involved in the diagnostic process, the technical development of ML models, regulatory processes, and other relevant intersections in these areas. The ecosystem is divided into four key constellations: clinical, operational, technical, and regulatory. Each of them addresses relevant normative aspects derived from each relationship and evaluates them according to the proposed rights-based approach. It also addresses the normative considerations regarding the intersection of normative goals and concludes with an exercise that aims to show how the normative assessment of the distribution of potential risks and benefits could be conducted, integrating all the elements developed throughout the dissertation.

Contents

Abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
AAAQ	Availability, accessibility, acceptability and quality
ACM	Association for Computing Machinery
AGI	Artificial General Intelligence
AI	Artificial intelligence
AIES	Artificial Intelligence, Ethics and Society
AMD	Advanced Micro Devices, Inc
ASI	Artificial superintelligence
CAD	Computer assisted diagnosis
CDSS	Clinical decision support system
CIA	Confidentiality, integrity, and availability
CNN	Convolutional neural network
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CT	Computer tomography
CUDA	Compute unified device architecture
DDSS	Diagnostic decision support system
DL	Deep learning
ECG	Electrocardiogram
EEG	Electroencephalogram
EHDS	European Health Data Space
EHR	Electronic health records
EMA	European Medicines Agency
ER	Emergency Room
FAccT	Conference on Fairness, Accountability, and Transparency
FDA	Foods and Drug Administration

GDPR	General Data Protection Regulation
GPD	Gross domestic product
GPT	Generative Pretrained Transformer
GPU	Graphics processing units
HIC	High-income country
HIPAA	Health Insurance Portability and Accountability Act
HITL	Human-in-the-loop
IBM	International Business Machines
ICESCR	International Covenant on Economic, Social and Cultural Rights
ICU	Intensive care unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IOM	Institute of Medicine
LLMs	Large language models
LMIC	Low- and middle-income country
ML	Machine learning
MRI	Magnetic resonance imaging
NLP	Natural language processing
OECD	Organisation for Economic Co-operation
OOD	Out of distribution
PACES-MRCPUK	Practical Assessment of Clinical Examination Skills of the Federation of Royal Colleges of Physicians of the UK
PCP	Primary care physician
PHI	Protected health information
PII	Personal identifiable information
PUE	Power use effectiveness
SSI	Safe superintelligence
STS	Socio-technical systems
TDP	Thermal design power
TNM	Tumor, nodes and metastasis
TTD	Time-to-die
UDHR	Universal Declaration of Human Rights
VSD	Value sensitive design
WHO	World Health Organization
WUE	Water usage effectiveness

List of Figures

List of Tables

Machine Learning in Medical Diagnosis: The Chances

1

The development and implementation of new technologies in healthcare has always been a complex task to undertake. Clinical settings are spaces where multiple convergences occur. There is a multiplicity of stakeholders with a plurality of values and potentially conflicting interests, intricate regulatory protocols, and frameworks, financial constraints, and incentives from public and private actors, and the sensitive matter of dealing with human health and life, which adds a particularly strong normative relevance to existing and emerging challenges. As such, there is an increasing necessity to take a step aside from the excessive enthusiasm that sometimes surrounds the field of medical AI and undertake a critical assessment of the benefits and risks of these technologies. Fortunately, in parallel with the technical advancement of AI systems over the last decade and the rush of capital being invested in AI-driven technologies for healthcare, the ethical assessment of challenges and potential risks has also become a matter of general interest. This comes not only from regulators, academic institutions, or healthcare organizations, but from technology companies in the business of developing AI systems as well as from healthcare providers who are interested in integrating these technologies into their products and services.

A critical assessment of the implementation of ML in medical diagnosis in normative terms, as proposed in this dissertation, requires identifying the strengths and weaknesses of the use of the models in relation to the relevant moral claims that moral agents have regarding their rights and duties. The first aspect of this assessment will be conducted in this chapter, which consists of establishing a foundation grounded on the factual possibilities of ML models in medical diagnosis. The reasoning behind this assessment, as it will be further developed in Chap. 2, is that the tension between the excessive enthusiasm

L. D. Dias Duran, *Machine Learning in Medical Diagnosis*, Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics,
https://doi.org/10.1007/978-3-662-71357-0_1

and uncritical skepticism regarding the benefits and risks of AI has formed a broad gap that distracts the attention from important normative considerations with practical repercussions in the short term like the infringement or violation of people's rights. Setting unrealistic expectations can also jeopardize harnessing the aspects of these models or types of applications that could be beneficial in areas of science or civil society, where technology could be a part of the solutions needed to tackle certain existing challenges[1]. From a business perspective, boosting the capabilities of AI without a factual basis is not only morally irresponsible, but can lead to a so-called "AI winter", a period marked by a generalized sense of disillusionment born of the unmet expectations. During an AI winter, the private investment dwindles, academic research on the subject is less likely to receive further funding and mass media focuses its attention elsewhere. For companies working in this field, such a prospect can be disastrous, as a significant source of their initial investment comes from venture capital.

Therefore, the first task in order to help close this gap is to have a solid understanding of what ML models are and how they operate in the diagnostic process. This is essential for the assessment of benefits and risks because proper normative work, i.e., one that considers whether certain actions are permissible, requires having as much solid factual knowledge as possible about the context in which the assessment is conducted. Without such a foundation, any normative work would be mostly devoid of actual usefulness, for instance, to inform policymaking, offer ethical guidance to companies or organizations, or to educate the public about matters of general interest. Based on this factual information, I seek to present an informed overview of the realistic possibilities of these technologies in medical diagnosis that will serve as a basis for the identification of their risks and benefits that will be carried out in the later chapters of this work.

The structure of this chapter will be organized as follows: Sect. 1.1 will provide a short overview of the history of ML in medical diagnosis through generative models. Sect. 1.2 will investigate the relevant aspects of ML models frequently developed for diagnostic purposes. Similarly, Sect. 1.3 aims to establish a common ground of basic concepts and elements about medical diagnosis that will be discussed throughout the dissertation. Finally, Sect. 1.4 is focused on the benefits of the application of ML models from technical and non-technical perspectives.

[1] Kevin LaGrandeur, "The Consequences of AI Hype," *AI and Ethics* 4, no. 3 (August 1, 2024): 653–56, https://doi.org/10.1007/s43681-023-00352-y.

1.1 History of Artificial Intelligence in Medical Diagnosis

It is a difficult task to pinpoint the emergence of the general notion of artificial intelligence, broadly understood as the simulation of human intelligence by automated entities or machines. Depending on the perspective at hand, various accounts of artificial automated beings are scattered throughout history, from myths to philosophical speculations to the modern notions of robots and automated systems. A first trace of such an idea can be found in ancient Greece and Egypt where sacred mechanical statues were thought to have minds. There are similar stories from ancient China where a mechanical human-like artifact was presented to a king[2] and from texts in the Jewish mystical tradition that speak of golems, anthropomorphic beings created from mud[3]. During the Age of Enlightenment there were further advancements that were grounded in philosophy and mathematics as in the case of Blaise Pascal, who invented the first mechanical calculator in 1642 and of Gottfried Leibniz who formulated the "alphabet of human thought", a catalogue of basic elements or concepts that humans are born with and whose combinations comprise all human thoughts in 1679[4] which is considered one precursor of computational linguistics. In 1837, Charles Babbage and Ada Lovelace proposed the first digital mechanical general-purpose computer called the "analytical engine"[5] and in 1912 Leonardo Torres Quevedo built an automaton capable of playing chess[6].

Although one could most likely find that such historical accounts play a role in understanding the reach and limitations of a notion of AI, it is safe to say that for the field of healthcare this conception appeared thanks to the convergence of

[2] Joseph Needham, *Science and Civilisation in China*, vol. 2 (Cambridge: Cambridge university press, 1991), 53.

[3] Hillel J. Kieval, "Pursuing the Golem of Prague: Jewish Culture and the Invention of a Tradition," *Modern Judaism* 17, no. 1 (1997): 1–23.

[4] Anna Wierzbicka, "The Alphabet of Human Thoughts," in *The Alphabet of Human Thoughts* (De Gruyter Mouton, 2011), 23–52, https://doi.org/10.1515/9783110857108.23.

[5] Pamela McCorduck, Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence, 2nd ed., An A K Peters Book (Boca Raton: CRC Press, 2018), 31–32.

[6] Francisco González De Posada and Francisco A. González Redondo, "En Torno al «Astra-Torres XIV», El «autómata Ajedrecista» y Los Ensayos Sobre Automática: Leonardo Torres Quevedo, 1913–2013," *Llull: Revista de La Sociedad Espanola de Historia de Las Ciencias y de Las Tecnicas* 36, no. 78 (December 2013): 456–65.

research from different fields in the 20th century[7] and, famously, the inquiries of Alan Turing about the theoretical feasibility of thinking machines[8]. Based on his research that theorized that any calculation, in all likelihood, could be represented digitally, Turing focused on the cognitive capacity of machines and devised a test that aimed to discover whether machines could carry a conversation with a human evaluator over a teleprinter without being recognized as a machine. In short, Turin wanted to know if it was possible to build a machine that could not be distinguished from a human by semantic rules. This is widely known as the "Turing test".

Another evidence of the hypotheses about machines built to perform actions that could be classed a manner of thinking comes from Descartes, earlier than the advent of computer science. In his *Discourse of the Method,* contrary to Turing's attempt at building a machine that could fool a person from trying to distinguish it from a real human, Descartes suggested a machine could in all certainty not pass as human because of two characteristics. First, due to its inability to interact appropriately to contextual communication as a person could do. Although Descartes envisioned it would be possible that machines be built to emit sounds that resembled human communication, he did not think plausible that they could understand meaning[9]. The second way to distinguish a machine is that they cannot, according to Descartes, act in a conscious manner. Descartes accepts machines can be made with a sufficiently complex disposition to perform certain actions, and do so even better than humans, but he argues no machine can be built with all the dispositions to perform all actions that humans do throughout their lives. He says that in order for machines to act seemingly like humans, their parts (he calls them "organs") have to be arranged specifically for each action and because of the complexity of this, he considers unfeasible that a machine would contain sufficient parts and be of such a complexity that it could perform like a human in every context[10,11].

[7] Among the fields that contributed in various ways stand out neurology with the discovery of electronical pulse signals among neurons and the theory of computation that hypothesized that any computation could be done by a machine.

[8] Alan Turing, "I.—Computing Machinery and Intelligence," *Mind* LIX, no. 236 (October 1, 1950): 433–60, https://doi.org/10.1093/mind/LIX.236.433.

[9] René Descartes, *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*, trans. Ian Maclean, Oxford World's Classics (Oxford; New York: Oxford University Press, 2006), 46.

[10] Descartes, A Discourse on the Method, 47.

[11] Although the technical performance of state-of-the-art models might seem to have proven Descartes skepticism wrong, his reflections allow to establish a key question regarding the

From the viewpoint of computer science, it is generally agreed that the term "artificial intelligence" was officially coined by John McCarthy along with other colleagues for the famous Dartmouth workshop at Dartmouth College in 1956 (formally known as the "Dartmouth Summer Research Project on artificial intelligence"), at which AI as a research discipline was officially founded. McCarthy defined AI as "the science and engineering of making intelligent machines"[12], while Marvin Minsky, also a participant at the conference, defined it as "… the science of making machines do things that would require intelligence if done by men"[13]. Expectedly, there was a great amount of excitement derived from these ideas that fueled what is now known as the first AI spring, spanning from 1956 to early 1970. During this time, in 1959, Arthur Samuel, an engineer at IBM (International Business Machines) corporation, coined the term "Machine Learning" to describe the process behind a computer program for the IBM 701 computer that calculated the chances of winning at the game of checkers and incorporated techniques like rote learning, in which the computer program stores the moves already made and uses them against new data i.e., new moves[14].

After the Dartmouth workshop, the focus of AI development was on machines that could be able to follow instructions. The uptake of such AI-driven devices and systems in different industries and academic research rapidly occurred. For instance, General Motors employed the first industrial arm called "Unimate" in its assembly line in 1961 to transport die castings and welding them into car bodies, a job with significant risk of harm to assembly line workers. Other famous projects were an interactive program able to carry short bouts of dialogue called "ELIZA" in 1964 and a robot built by scientists at the Massachusetts Institute of Technology in 1969 called "Shakey" that could follow sets of complex instructions by breaking them into simpler blocks.

---

actual capabilities of these technologies: can be considered intelligence the arrangement of specific internal functions that outwardly look on par with human cognitive abilities? Although the matter of intelligence as such will not be researched in depth in this dissertation as it is a highly complex subject that requires a whole dissertation of its own, this question is relevant because it enables an important distinction that will be discussed later on between the notions of strong or general, and weak or narrow artificial intelligence.

[12] John McCarthy, "What Is Artificial Intelligence?" (Stanford University, 2004).

[13] Marvin Lee Minsky, Preface to *Semantic Information Processing*, Reprint (Cambridge (Mass.) London: MIT Press, 1988), v.

[14] Arthur L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development* 3, no. 3 (July 1959): 210–29, https://doi.org/10.1147/rd.33.0210.

In contrast, the scenario of development of AI for the healthcare sector was much slower. The first decades after the Dartmouth College workshop were mostly occupied by digitization of medical data and from these efforts surged clinical research databases like Medlars (the predecessor of Medline) in 1963 and also the foundation for what would be the electronic health records of today[15]. Although by the late 1960s and beginning of 1970s the excitement about the possibilities of AI in computer science has significantly decreased as a result of the unfulfilled expectations and lack of progress[16], and the first AI winter had settled in, a notable advancement for AI in science kick-started the first remarkable period of the history of medical AI. In 1960 a team of computer scientists, chemists, and microbiologists developed a software program called "Dendral" employed to help organic chemists identify unknown organic molecules using an approach that will be later named as expert system[17].

Expert systems are a type of AI-driven computer programs that aim to solve complex problems by using logical inference based on IF-THEN rules derived from the domain knowledge of human experts. They have four main components: first, a knowledge base that stores the specific domain knowledge and the rules by which the knowledge is organized. Second, an inference engine that receives "queries" and returns "answers" from the knowledge base. Third, a user interface that is used to communicate the results of the queries to the user, and fourth, an explanation facility that consists of the module used to explain the input-outcome process to the user. An important feature of expert systems is that they use both symbolic and heuristic methods[18]. Symbolic reasoning is, in very simple terms, based on rules set and coded by humans and it is distinct from ML, which learns the rules by establishing patterns in the data independently

[15] Vivek Kaul, Sarah Enslin, and Seth A. Gross, "History of Artificial Intelligence in Medicine," *Gastrointestinal Endoscopy* 92, no. 4 (October 1, 2020): 808, https://doi.org/10.1016/j.gie.2020.06.040.

[16] This disillusionment was also fueled by a report commissioned in 1973 by the UK parliament about the state of AI development and conducted by the mathematician James Lighthill. The negative results of the report occasioned universities across the UK to cease almost all research on the field.

[17] There is some disagreement among scientist regarding this statement as expert systems later on were more methodologically refined and when we speak about them we tend to refer to the models like MYCIN or CADUCEUS. Nevertheless, the Dendral project is responsible for establishing a new approach for problems where domain expertise is highly necessary and not easily programmable.

[18] B G Buchanan and R G Smith, "Fundamentals of Expert Systems," *Annual Review of Computer Science* 3, no. 1 (June 1988): 23–58, https://doi.org/10.1146/annurev.cs.03.060188.000323.

from human intervention. The concept of heuristics describes different criteria, principles, rules, or methods for finding the most effective course of action from a variety of possibilities[19]. The use of heuristics, in much of the same fashion as humans, allows the model to produce a faster answer and simplify complex decision-making processes.

One of the most notable expert systems was MYCIN, a consultation system developed in 1974 with the aim of helping physicians identify bacteria that caused severe infections. It used a backward chaining inference method. This means that it starts from the goal, i.e., the infection, and uses simple yes or no textual questions to guide the user to a potential cause. The system could provide a list of potentially responsible bacterial pathogens and even a recommendation for treatment adjusted to the patient's weight[20]. Another early example of expert systems was the CASNET model developed by Rutgers University in 1976. This system was intended to test the feasibility of applying AI to biomedical problems and it was mostly known for its application as a consultation model to help the diagnosis and treatment of glaucoma[21]. It is relevant to remark that these early expert systems were not built for specific medical problems. The fundamental value of these systems was to represent the knowledge about a complex process, such as the medical knowledge about a disease, successfully, in a manner that could be used as a tool for reasoning. In other words, that they could assist physicians during the diagnostic process of a certain condition.

Despite their seemingly resounding success during the 1970s and 1980s, expert systems are not at the forefront of the applications of health information technology in recent years, although they are still present in other industries like business management and operations, and they have evolved with the advent of state-of-the-art AI systems. The reasons for their decline in healthcare come from a combination of technical and economic factors, however, a transversal aspect to them is the discrepancy between the expectations set regarding their possibilities and the results they delivered. Initially, expert systems were described as capable

[19] Judea Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, The Addison-Wesley Series in Artificial Intelligence (Reading, Mass: Addison-Wesley Pub. Co, 1984), 3.

[20] Edward H. Shortliffe et al., "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System," *Computers and Biomedical Research* 8, no. 4 (August 1, 1975): 303–20, https://doi.org/10.1016/0010-4809(75)90009-9.

[21] Sholom Weiss, Casimir A. Kulikowski, and Aran Safir, "Glaucoma Consultation by Computer," *Computers in Biology and Medicine* 8, no. 1 (January 1, 1978): 25–40, https://doi.org/10.1016/0010-4825(78)90011-2.

of solving difficult problems as well as or even better than human experts. It was said that they could interact with humans effectively, could handle uncertainty and function with errors in the data, were able to grasp the complexities behind competing hypothesis and could justify their conclusions[22].

However, there were major drawbacks that heavily hindered their effective implementation in healthcare. The first hurdle is the matter of knowledge acquisition or elicitation. As explained before, expert systems rely entirely on the knowledge represented in the database to operate, which is modeled from the knowledge of human experts. Thus, one of the central tasks in building the knowledge base of these systems was to find experts interested in participating in the project, which was not an easy task. One reason was that the experts with the knowledge needed to model these complex problems might not be readily available or willing to work with the computer scientists. Another reason was that the knowledge itself might be so complex that the domain experts could not communicate it in a way that could be represented in computer language[23]. An expert system in healthcare is ultimately an attempt at framing vast amounts of medical knowledge from different sources into well-defined rules. Therefore, if a system does not have sufficient domain knowledge modeled with clear rules, the inference part could generate errors, which would mean the failure of the system. Expert systems, unlike ML architectures, are unable to produce results that do not correspond to the exact knowledge that has been introduced into the knowledge base. Conversely, even if the experts were available, acquiring the knowledge and then modeling it according to the rules for the inference process was expensive and lengthy, which made expert systems financially challenging.

A second hurdle was that there was not enough computing power or even hardware resources to accommodate the weight of the information required at the time of their invention, not to mention the fact that the efforts to digitize data, while already significant, were not comprehensive enough for the complexity of many medical challenges. Finally, a third hurdle was the issue of user acceptance. Similar to the current situation with ML models in healthcare, the developers of expert systems faced challenges to convince clinicians that they were safe to use in clinical practice, that they would not end up replacing medical professionals, and that there was a problem that would be solved with integrating these technologies in clinical workflows. Moreover, there was a lack of standardized

[22] Pamela K. Coats, "Why Expert Systems Fail," *Financial Management* 17, no. 3 (1988): 77–78, https://doi.org/10.2307/3666074.

[23] Michael Z. Bell, "Why Expert Systems Fail," *Journal of the Operational Research Society* 36, no. 7 (July 1985): 613, https://doi.org/10.1057/jors.1985.106.

assessment measures for these systems and studies to validate their algorithms for clinical practice were seldom confirmed[24].

After the decline of expert systems that led to a second AI winter until the late 1990s, the interest in ML methods was renewed during the first decade of the 21st century. Several corporations became involved with the development of ML models in healthcare. For instance, IBM, the company that in 2007 had created the AI system "Watson", famous for having defeated the world champions in the quiz game *Jeopardy!* in 2011, announced a partnership with the Memorial Sloan-Kettering Cancer Center and the Cleveland Clinic to use Watson to assist medical professionals in processing large quantities of patient data to improve predictions and diagnosis of cancer. In 2016, they started an ambitious project called "Watson for oncology" along with Manipal Hospitals, a chain of hospitals in India, that sought to provide clinicians with information to treat patients with cancer. However, the expectations of hospitals were not fulfilled and after an investigation carried out by health news website STAT that reported that Watson had recommended unsafe and incorrect cancer treatments[25], Watson Health, the entire branch specialized in medical problems, was sold to a private equity firm.

The next remarkable event in the development of AI that impacted the implementation of these technologies in healthcare was the popularization of deep learning algorithms (which will be defined in the next section) occurred in 2012 as a result from an annual competition organized by the ImageNet project, a large database of images created for visual object recognition research. ImageNet is composed of over 14 million hand-annotated (i.e., labeled by humans manually) images classified in approximately 20.000 categories. The competition, called the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), consisted in achieving the highest accuracy rate possible in a series of image identification and classification tasks. In 2012, a program named AlexNet created by computer scientist students Alex Krizhevsky, Ilya Sutskever and supervised by computer scientist Geoffrey Hinton achieved a margin of error of only 15.3%[26]. This was a revolutionary result as the minimum error margin at the time was

[24] Juri Yanase and Evangelos Triantaphyllou, "A Systematic Survey of Computer-Aided Diagnosis in Medicine: Past and Present Developments," *Expert Systems with Applications* 138 (December 2019): 112821, https://doi.org/10.1016/j.eswa.2019.112821.

[25] Casey Ross Swetlitz Ike, "IBM Pitched Its Watson Supercomputer as a Revolution in Cancer Care. It's Nowhere Close," *STAT* (blog), September 5, 2017, https://www.statnews.com/2017/09/05/watson-ibm-cancer/.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM* 60, no. 6 (May 24, 2017): 84–90, https://doi.org/10.1145/3065386.

approximately 25%. The success of the model rested on the application of a type of DL algorithm called convolutional neural network (CNN), which had already been formulated in the late 1980s but that had been made feasible by the improved processing capacity of the recently launched CUDA (Compute Unified Device Architecture) a parallel computing platform that uses graphics processing units (GPU) from Nvidia[27]. These technical advantages paired with the increasing availability (in terms of collection methods and storage possibilities) and improved quality of digital data, in this case digital images, started what is considered the 21st century "AI summer" or "AI boom".

As a result of the success of CNNs, this new set of ML algorithmic architectures gained significant popularity and displaced other forms of symbolic reasoning AI[28]. Over the course of a decade, the number of ML models in computational research exploded, facilitating the development of other subfields such as natural language processing (NLP). In 2019, thanks to the application of DL algorithms and the increasing capacity of GPUs, the first models based on the transformer architecture, NLP and trained on large language datasets appeared. They consolidated a new subfield that has given way to state-of-the-art large language models (LLMs) and generative AI models such as ChatGPT or Midjourney[29]. Since 2020, the integration of LLMs in healthcare has again increased the expectations of the possibilities of AI in medical diagnosis and patient management. Developments included domain-specific models such as BioBERT, which was fine-tuned on biomedical texts to improve entity recognition and medical language understanding, and applications as decision support systems like Google's Med-PaLM.

This opened the door for a revitalized interest in applying the new model architectures and methods to medical problems. Applications of these models in medical diagnosis highlight their potential to increase the efficacy of diagnostic procedures by reducing diagnostic error, support physicians with immediate and simple access to up-to-date medical literature, enable healthcare delivery to areas

[27] A GPU is an electronic circuit designed to process digital images. However, due to their parallel structure, they were found suitable to be applied to data processing for ML development.

[28] The name "Deep learning" refers to the layers that compose the structure of the algorithm. This will be explained later in this Chapter.

[29] Stefan Feuerriegel et al., "Generative AI," *Business & Information Systems Engineering* 66, no. 1 (February 2024): 111–26, https://doi.org/10.1007/s12599-023-00834-7.

of difficult access or lack of personnel, among others[30]. Although this seems like an exciting prospect, it will be the aim of the upcoming chapters to show that it is necessary to evaluate whether these are realistic possibilities grounded in evidence both at the technical, socioeconomic and ethical levels.

1.2 Artificial Intelligence and Healthcare: The Fundamentals

The overview of the history of the development of AI and ML in medical diagnosis helped to substantiate the argument highlighted at the beginning of this chapter, namely that the setting of exaggerated expectations without evidence leads to disillusionment, and this can negatively affect the course of the development of the field of AI and its potential positive impact on, among other things, the process of diagnosing diseases and conditions. The underlying importance of having a solid grasp of the technical possibilities of ML models lies in the Kantian maxim that the material "can" has a significant impact on the normative "ought". In other words, it is assumed that the technical materiality of ML models and the way they operate have a direct bearing on the normative criteria to decide if using them is justified or not, in which scenarios or under which considerations, and what potential normative conflicts might arise.

This assumption is sustained in the conceptualization of state-of-the-art AI as socio-technical systems, which acknowledges the interrelatedness of social and technical factors at every step in the process of ideation, development, implementation, and use of these technologies. The key implication of such conceptualization is that the challenges and risks that derive from implementing these technologies cannot be addressed without considering both the technical and the social factors that play a role. Otherwise, there is a chance of overlooking potential risks to human rights or ignoring the underlying reasons or causes of these challenges. From this perspective, the establishment of this factual basis also serves as a common ground for the interdisciplinary work required to approach the assessment of the distribution of benefits and potential risks of ML in diagnostics that this dissertation aims to provide.

[30] Jan Clusmann et al., "The Future Landscape of Large Language Models in Medicine," *Communications Medicine* 3, no. 1 (October 10, 2023): 1–8, https://doi.org/10.1038/s43856-023-00370-1.

1.2.1 Artificial intelligence, Machine Learning and Deep Learning

Although the development of AI has had a not insignificant impact on most areas of human life over the past few decades, there is still some degree of confusion about the meaning of the discipline's fundamental concepts and how they differ from one another. Of course, this is not a straightforward endeavor as there are numerous definitions of these concepts that have changed throughout the decades along with the evolution of the technologies and the field as such. For instance, in 2023 the Organization for Economic Co-operation (OECD) presented a revised definition to the one presented in 2019 that aimed to reflect the significant changes occurred in posterior years, like the role of human decision-making and their impact in physical environments aside from digital ones[31]. Another everyday use of the term AI is as a field of research that studies the general question of intelligence (not just human) and how to apply it to machines and computational systems. Despite the disagreements, there are common elements, such as the fact that AI does not refer to a specific model or system but is used to designate a wide range of computational strategies and techniques that aim to solve complex tasks imitating -or attempting to imitate- human intelligence[32], or that some of the most relevant tasks that AI is designed for include knowledge representation, planning, natural language processing, and symbolic reasoning.

One important distinction to make at this point is between "general" and "narrow" intelligence. Whilst at present most of the attention is directed towards models like ChatGPT, in fact, the commercial use of AI is decades old. Known examples of this are email filtering software, search engines, chatbots, language translation, AI assistants like Google's Siri or Amazon's Alexa, and recommender systems in social media and streaming services. All these applications, including

[31] OECD, "Explanatory Memorandum on the Updated OECD Definition of an AI System" (Paris: OECD, March 5, 2024), https://doi.org/10.1787/623da898-en.

[32] I want to make a minor conceptual distinction between AI systems and ML models as it could seem that they are being used interchangeably throughout the dissertation. When using the concept of AI systems, I am referring to all types of AI methods including ML and DL but not only about a particular ML model or a specific DL architecture. Instead, I employ it to speak of general risks derived of AI implementation. Conversely, I use the wording "ML models" to denote particular programs that were trained on a dataset for a specific purpose in healthcare or medical diagnosis. The reasoning behind is that the evaluation I am proposing in the dissertation does not attempt to apply to all AI systems as a blanket evaluation like the principlist approach does (see Sect. 2.3.1) but instead acknowledges that there are particular normative considerations to models under specific conditions like relationships between moral agents.

LLMs and generative AI, belong to the categorization of "weak" or "narrow" AI. The main characteristic of these applications is that they can perform a well-defined computational task, such as prediction or classification, but are unable to extrapolate beyond that task. Although extrapolation is a mathematical concept, when applied to human action, it is the ability to estimate, expand, or project prior experience or knowledge about a known context to an unknown scenario or problem. Essentially, this means that all the existing applications of AI, be it ML or DL, cannot create something from scratch because they lack the ability to extrapolate. One can then think of the generative abilities of large language or image generation models based on state-of-the-art DL. What they *can* do is generalize from their input data but not extrapolate. Generalization is the ability of a model to make predictions from data that was not included in the training dataset but that share similar characteristics or values. The ability to generalize is unique to ML algorithms and is not possible for other AI methods. This means, that the texts from GPT-4, for instance, in the format of short stories, are not original works created by the AI model but instead the result of an order or "prompt" that the model generated based on the millions of data examples that comprise the datasets that were used by computer scientists at the training phase.

The concept of "strong" or "general" AI, conversely, describes a system that is capable of solving tasks without being explicitly programmed to do so. Beyond this moderate definition of strong AI, there are conceptualizations of hypothetical machines capable of achieving human intelligence, attaining consciousness or sentience, or surpassing human intelligence altogether. Although this dissertation will not be concerned about these hypothetical scenarios in depth, Sect. 2.2.1 discusses some normative issues derived from them. Nevertheless, the distinction between weak and strong AI is crucial in normative terms because it helps to establish that AI models do not "think" or "reason" for themselves, despite what media articles or some scientists depict. Therefore, assigning them moral qualities in contexts with decision-making processes of normative weight is not only *not* possible from an ethical point of view but also careless as it generates confusion about on whom should the responsibility or accountability be allocated if there are unwanted consequences and what is reasonable to demand from the performance of the models, in other words, what we can realistically expect that they achieve.

The concept of ML, as has been hinted at, is a subfield of AI that develops algorithms able to "learn" from data and generalize without explicit programming. Traditional approaches, like the expert systems of the 1990s, use explicit rules programmed by computer scientists to process input data and obtain specific, desired outputs. In ML, the focus is shifted from those strict programmed rules with direct input commands to learning techniques that can map patterns

across large amounts of data instead[33]. Despite the improvements in ML techniques, they were limited in their ability to process data in its unstructured form (see the next subsection on data), so computer and data scientists had to go through the complex process of transforming this type of data into data that could be effectively and comprehensively processed by the algorithm to produce accurate results. DL methods have revolutionized this process in the last decade. DL refers to a specific set of ML algorithmic architectures[34] and are known for their representation learning functionality[35], i.e. they can process unstructured data and discover the representations needed for tasks through a set of layers of algorithms and mathematical functions. Representation learning is perhaps one of the most important technical advancements of ML for medical diagnostic purposes because it enabled ML models to use unstructured data and with remarkable accuracy rates. However, to understand this in depth, it is necessary to clarify how data is used to train these models.

1.2.2 Data

The availability of large quantities of quality data, also known as "big data," is one of the essential elements for the success of ML models in the latest AI boom[36]. As was mentioned in Sect. 1.1, after the establishment of AI as a research field, many efforts in the healthcare sector were spent in developing digital repositories of medical literature and the first versions of electronic

[33] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane, "Machine Learning in Medicine," New England Journal of Medicine 380, no. 14 (April 4, 2019): 1348, https://doi.org/10.1056/NEJMra1814259.

[34] As such, when I speak of ML models I am almost always including DL algorithmic architectures as well.

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521, no. 7553 (May 2015): 436–44, https://doi.org/10.1038/nature14539.

[36] Although it is common to see the concept of big data closely associated with the field of AI, this has led to the misconception that big data exists only in relation with it. However, big data is an industry in its own right that predates the emergence of applicable ML models and has its own set of normative considerations. Big data is typically defined by three features: volume, velocity, and variety, reflecting its size, speed of generation, and diversity. However, this concept is expanded by additional traits like exhaustivity and scalability, which enhance its capacity for detailed, multi-scale analysis. For more see: Rob Kitchin and Gavin McArdle, "What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets," *Big Data & Society* 3, no. 1 (June 1, 2016): 2053951716631130, https://doi.org/10.1177/2053951716631130.

health records (EHR). However, the use of health data in ML requires that the data is accessible by known keys or attributes that the model can process. There are, broadly speaking, two types of data according to this criterion: Structured, which means data that is organized in a standardized, clearly defined and consistent manner and format, for instance, like an excel sheet, and unstructured data denotes all other data that does not fit into the structured definition, such as images or videos.

Data in healthcare is collected from different sources like the patients themselves or their proxies, EHR, health devices, and wearables, pharmacies, clinical trials, genetic testing, insurance companies, governmental health agencies, senior homes, hospitals, or clinics, clinical laboratories, and private praxis. It comes also in diverse formats like text (EHR, medical observations, administrative data, claim data health surveys), video or voice recordings (patient-physician interactions or surgeries), medical imaging (X-rays, MRI, and CT scans, ultrasounds, endoscopies, ECG, and EEG, among others), biological samples (tissue or blood tests) and wearable health devices or apps (activity levels, heart rate, calorie intake, sleep quality). The challenge is that, as estimated by data experts[37], most of the health data generated is unstructured, which means that it is not suitable for traditional ML methods.

However, DL models with their representation-learning architecture, like the CNN mentioned in Sect. 1.1, can use unstructured data and make highly accurate predictions and classifications. For example, a grayscale x-ray image is essentially a grid of pixels that are represented by a particular number depending on the intensity of the color. What the DL algorithm is processing is not the image as we know it, but a grid of numbers that represents the features of the image in an accessible, well-defined way. The model then analyzes the pixel intensity patterns in the image, automatically extracting relevant features, such as edges, textures, and shapes. From there, it progressively learns hierarchical representations, from simple features such as lines or curves, to more complex ones, like lungs or bones, and from this analysis, the model is trained to detect abnormalities in the images. While turning unstructured data into structured data manually is technically possible, it is extremely difficult, time-consuming and prohibitive in terms of cost. Therefore, since most of the available health data is unstructured

[37] Michael O'Reilly, "Council Post: The Unseen Data Conundrum," Forbes, accessed August 7, 2024, https://www.forbes.com/sites/forbestechcouncil/2022/02/03/the-unseen-data-conundrum/.

and other ML methods are unable to process it, DL has become arguably one of the most popular ML approaches for the development of healthcare models.

There are two main approaches to train a ML model to learn to identify the patterns in the data. The first one is called "supervised learning". In this approach, there are two key elements: the feature or features that are relevant for the prediction task that are known as "input value" and the "labels" or desired outcome that is known as "output value". What the ML model does in supervised learning is learn which features are associated with which labels. The main characteristic of this approach is that the inputs are paired with the outputs from the start and what the model learns is the logic behind these associations so it can apply it to unseen data, i.e., generalize to new data. As such, the success of the model in supervised learning depends completely on the labels, also known as the "ground truth", being correct. There are several challenges concerning the accuracy and reliability of the labels. The first one is the differing opinions among medical experts about what constitutes the ground truth. Since labels are manually annotated by human experts, it is not uncommon, especially for complex diagnostic cases, that the annotators have different opinions about the location, size, or severity of an abnormality on an image, which can result in a label that is not sufficiently accurate and thus, reliable. This is one of the examples in which human bias can be introduced in the data samples, as it will be further discussed in Chap. 4. Another challenge is the difficulties of annotating images or tests that have poor quality and that might lead to errors in the data. Other issues are the lack of domain expertise to do the annotations as occurred in expert systems, as well as the costs and time required to annotate the labels for large datasets.

In supervised learning the algorithms focus on two types of tasks: classification, in which the algorithm sort classifies the dataset into the categories indicated by the label, and regression, in which the objective is to predict continuous values by determining the underlying relationship between the input and output values. For instance, a ML model using an algorithm based on regression analysis can predict or calculate the risk score for cardiovascular disease based on factors like age, cholesterol levels, blood pressure, family history, habits like drinking and smoking[38].

The second approach is "unsupervised learning". Contrary to supervised learning, this approach removes the need to have labeled outputs and instead it maps the patterns in the data independently from human intervention. This approach is

[38] Kevin Cullen et al., "Multiple Regression Analysis of Risk Factors for Cardiovascular Disease and Cancer Mortality in Busselton, Western Australia—13-Year Study," *Journal of Chronic Diseases* 36, no. 5 (January 1, 1983): 371–77, https://doi.org/10.1016/0021-9681(83)90169-8.

usually applied to DL algorithms. The tasks that are best suited for it are clustering, i.e., grouping data points, and association. One of the key advantages of unsupervised learning, aside from using unstructured data, is that it can find useful associations in the data that are not apparent or even visible to human experts. This has been useful for early detection of tumors, for instance. However, while in most cases, well-programmed unsupervised learning can produce a model capable of achieving a high accuracy rate in grouping the data, the downside of this approach is that the clustering may lack any meaning at all.

Another relevant technical aspect of the training process, be it with supervised or unsupervised approaches, is the structure of the data during the training phase. Normally, a data set is divided into a training set, usually a validation set, and a test set. A training set is a set of examples that is compiled so the model maps the patterns in the data from input to output label. However, there is a common misconception that a "good function" (a good algorithm) is one that can just accurately link the input value to the output value. However, the objective of ML is to produce a model that, after being trained, can process new inputs, i.e., new data, and yield the same accuracy. Therefore, the reason to set aside a portion of the data to test after the training phase, is to check whether the model is indeed effective at the task that is meant to achieve and can generalize outside of the data that was used to train it. This is the truly complex task, as will be expanded on further down, because there is a significant gap between a model performing under ideal conditions during the training phase and a model deployed in a real-world clinical setting and achieving the same results[39]. The validation process is an extra step that is not always included in the training phase. It sets aside a sample of data with the aim of periodically performing a task called "hyperparameter tuning". Hyperparameters are values chosen by the computer scientists prior to training the model that determines the learning process, or in other words, the way the model plots the patterns in the data. The tuning of hyperparameters is crucial because the end accuracy of the model depends on it. The more precise hyperparameter values are, the higher the performance the model will have against the test set.

[39] Christopher J. Kelly et al., "Key Challenges for Delivering Clinical Impact with Artificial Intelligence," *BMC Medicine* 17, no. 1 (December 2019): 195, https://doi.org/10.1186/s12916-019-1426-2.

1.3 Medical Diagnosis and Artificial Intelligence: The Fundamentals

In 2014 Warner Slack, an American physician, and pioneer in health information, said that "Any doctor who could be replaced by a computer should be"[40]. Although this may seem like a problematic statement to make at first, it can also lead to open valid questions about what are the conditions that could lead to doctors being so disconnected from the patients and the aims of their professions that a machine could take over.

The process of determining which disease or condition is the origin of a patient's set of symptoms is one of the most challenging aspects of medical practice. Symptoms can have different causes and can -or not- be correlated with other symptoms or particular conditions of a patient. In the diagnostic process, numerous aspects play a role. From quantifiable test results, physical examination, the patient's own perceptions and the physician's experience, knowledge, and intuition. As a result, performing a diagnosis is not a streamlined process. Despite the extensive history of medicine and the advancements it has made throughout the centuries, there is not a fully error-proof technique that can be taught to medical students. Moreover, social and cultural contexts also have a significant influence over how the diagnostic process is conducted.

Nevertheless, all diagnostic methods are dependent at least on two essential factors: the information or data available about the patient and his ailment, and the experience or knowledge of the diagnostician. The process starts with the first contact between the patient and the physicians. Here, a first aspect of normative weight must be recognized: that the patient and the physician are over two single individuals taking part in the process of receiving/giving a service. The patient comes to the primary care physician (PCP) in the role of a person with an illness, but also as part of a network of interrelated relationships with other people and institutions, and within specific contexts that may affect the course of the clinical encounter (this will be developed further in Chap. 5).

Similarly, the physician enters the clinical encounter in his role as a medical practitioner and brings along his medical knowledge, professional experience, and deontological code. However, he also encompasses his role as a representative of the medical profession, his particular institutional affiliation and also his own set of interconnections (with colleagues, other patients, his own family and personal relations, etc.), moral values and social context. Thus, the clinical encounter does

[40] Robert Wachter, The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age, 2017, 93.

not occur as an isolated event that lasts 20–30 minutes in a consulting room, but as an event that is situated in a larger context in which the interests, values, and choices of patients and physicians are influenced and may influence the interests, values, and choices of other moral agents. This is particularly relevant to an approach that considers certain rights as necessary conditions for living, as will be conceptualized in sect. 3.1.

The clinical encounter between the patient and the PCP is also the point at which the fiduciary relationship begins to form. The fiduciary concept comes from legal rulings that assigned a certain party the obligation to act on behalf of the interests of the other party. While initially this conceptualization was not necessarily grouped with relationships where the element of trust was present, when we speak of the fiduciary relationship between physicians and patients, we almost always include this aspect as essential. According to Ludewigs and colleagues, fiduciary relationships share common characteristics such as that the fiduciary party is entrusted with power over the legal or practical interests of the beneficiary, the services they provide are social in nature and require specialized expertise and that the beneficiary is in a position of vulnerability because of their epistemic position, i.e. the knowledge they possess about the service they require, so they are required to place their trust in the fiduciary party to act in their best interests[41]. The relevance of acknowledging these relational aspects is that, in order to understand and analyze the complex normative implications of implementing disruptive technologies such as ML models in the diagnostic process—and in health care in general -, it is helpful and even necessary to understand the moral agents involved not as isolated individuals. This framing is part of the methodological approach of a rights-based evaluation of the distribution of potential risks and benefits that will be presented in Chap. 3.

During the first and subsequent clinical encounters, the next task is to gather information about the patient and the symptoms of the disease or condition from which he or she is suffering. This includes standard practices such as interviewing the patient to gather as much information as possible about the patient's experience, which may include questions about his or her habits, recent abnormal events, past illnesses, and relevant family history. At this point in the clinical encounter, the physician also would perform an initial physical examination to contrast with the information provided by the patient. Finally, the physician would

[41] Sophie Ludewigs et al., "Ethics of the Fiduciary Relationship between Patient and Physician: The Case of Informed Consent," *Journal of Medical Ethics*, December 23, 2022, 1, https://doi.org/10.1136/jme-2022-108539.

begin evaluating which answer might provide an explanation for all the information gathered, in short, the cognitive exercise of making a diagnosis. This can lead to three different outcomes: first, that the physician makes a conclusive or tentative diagnosis and after communicating it with the patient, he or she is sent home with instructions (this can be a treatment plan or to remain observant in case the symptoms develop i.e., worsen or some other symptoms appear); second, that there is not a conclusive diagnosis, and the patient requires further specialized testing or a referral to physicians at other levels of care; or third, that there is a conclusive diagnosis that requires that the patient be admitted to a hospital or clinic for immediate treatment.

It is relevant to note that levels of care refer to a classification used in clinical settings to distinguish between the complexity of medical cases. Certain resources and clinician's expertise are allocated according to the levels of care, and as such, each level of care entails a unique relationship between the medical provider and the patient, as will be further developed in Sect. 5.2.1 and 5.4.2. The levels of care are usually divided into three categories: primary care, secondary care, and tertiary care. Primary care is the first point of contact a patient has access to when there is a health concern and encompasses a wide array of health professionals like general practitioners, nurses, allied health professionals, and group-specific primary care specialists like gynecologists, geriatrist, and pediatricians. Moreover, primary care also covers most aspects of preventive medicine, health promotion strategies, rehabilitation, and primary palliative care. The IOM Committee defines primary care as: "(…) the provision of integrated, accessible health care services by clinicians who are accountable for addressing a large majority of personal health care needs, developing a sustained partnership with patients, and practicing in the context of family and community."[42].

Secondary care includes clinicians in medical specialties such as cardiology, oncology, ophthalmology, etc. Unlike primary care clinicians, secondary care clinicians usually do not come into contact with the patient until after the PCP (primary care physician) has determined that the patient's evaluation warrants it and makes a referral. Finally, tertiary care encompasses the care of patients who have been admitted to a hospital, usually for the treatment of complex conditions that require highly specialized and often intensive care and equipment. This level includes medical specialties such as cardiac surgery, cancer treatment, and palliative care, etc.

[42] Institute of Medicine (US) Committee on the Future of Primary Care, "Defining Primary Care," in *Primary Care: America's Health in a New Era*, ed. Molla S. Donaldson et al. (National Academies Press (US), 1996), https://www.ncbi.nlm.nih.gov/books/NBK232631/, 1.

Although no golden method exists to conduct the cognitive part of the diagnostic process, the advancement of medical practice has allowed the refinement of techniques that work better than others. The differential diagnosis is a proven method that allows physicians to arrive at a satisfactory explanation about a patient's set of symptoms. It is a process that requires considerable skill, and it is sometimes considered having a bit of art in it. Diseases present themselves in a myriad of ways. It is not uncommon that the symptoms are manifestations of two or more different conditions. Thus, although they may respond to common markers, there can be significant challenges to determine which symptom corresponds to which disease. This is also not counting that one condition can be the cause of the other and therefore the physician needs to focus on the causation to resolve the other one.

The process of diagnosing a patient can also turn into an "iterative hypothesis testing" that evolves as there is more information and some hypotheses are ruled out. Robert Wachter explains that the diagnostic reasoning conducted by physicians is underlined by the logic behind the Bayes theorem, which in very simple words explains that common things occur commonly. This is conceptualized as "system 1" thinking based on heuristics and is defined by Eric Topol as "the reflexive, mental shortcuts that bypass any analytical process, promoting rapid solutions to a problem"[43]. The advantage of this thinking system is that it is quick and intuitive, which is useful under straining conditions present in healthcare like medical emergencies or a shortage of time to go through lengthy process of verification. The success of this kind of thinking relies heavily on the physician's experience and evidence-based knowledge gained over time. Intuition is not a skill that comes from nowhere, which makes this system rather difficult for new graduates or physicians at the beginning of their practice careers. The downside, besides the need for several years of active bedside practice to acquire experience and knowledge, is that this method, while statistically effective, is by no means infallible. As such, there is a not insignificant rate of diagnostic error, which carries the risk of negative clinical outcomes that can cause the deterioration of the patient's overall well-being, disability, or even death.

"System 2" thinking, conversely, is an analytical, reflective, more reliable process that meets a standard of scientific rigor and, for these reasons, is also slow. Although this approach to the diagnostic process may intuitively seem more appropriate given the complexity of diagnosing in the first place, in fact, System 1 thinking tends to be the immediate course of action for most physicians.

[43] Eric J. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, First edition (New York, NY: Basic Books, 2019), 43.

Moreover, system 2 does not resolve that more inexperienced physicians still do not have the experience and knowledge from practicing medicine for a period of time, even if they slow down to reflect on the possible causes of the symptoms.

The problem with this conceptualization of the diagnostic process mainly as a cognitive exercise performed by the clinician is that it does not consider the relational approach established earlier. As such, it is proposed here to expand this notion to one where the diagnostic process is "situated" within a context and shaped by the interactions with others and the surrounding environment[44]. A situated perspective on the diagnostic process engages with different views about what occurs at the doctor's office and contemplates that instead of the isolated event that was mentioned before, the process occurs with the influence and participation of others aside from patient and physician. For instance, it would be considered how many hours of shift has the physician had until that moment? What resources does he or she have at hand to achieve the diagnostic process, including the availability of supporting staff like nurses, technicians, pharmacists, lab workers, etc.? Such factors, which are usually not taken into account, can and do influence the development of the process and the way in which physicians deal with it. Another advantage of a situated perspective is its potential to help explain and address diagnostic errors. For example, if a patient has lost a job or a close family member is ill or has recently died, this may indicate insomnia, decreased activity levels, or even early signs of depression.

This expansion of the conceptualization of the diagnosis from a cognitive activity assigned exclusively to the medical professional in charge, to a broader process framed within a situated and relational approach, allows a broader and more comprehensive analysis of the impact of, since it facilitates the inclusion of agents, tools, and contexts that have been left out in traditional approaches. It will allow us to contemplate the advantages and disadvantages of ML-driven tools that, although not designed to conduct diagnostic tasks strictly speaking, contribute indirectly to the success of the diagnostic process, for instance, like with summarizing reports for referrals or insurance claims. Another benefit of this broader view is that the question of what the potential benefits and risks of introducing these technologies are would no longer be limited to those directly affecting the patient in question. Instead, the inquiry would extend to how physicians, other medical professionals, and other patients might benefit, or perhaps be exposed to potential risks. A more thorough development of these questions

[44] Mark L. Graber, "Progress Understanding Diagnosis and Diagnostic Errors: Thoughts at Year 10," *Diagnosis* 7, no. 3 (August 27, 2020): 151, https://doi.org/10.1515/dx-2020-0055.

and the argumentation itself will be conducted in Chap. 5 with the help of the relational, rights-based approach outlined in Chap. 3.

1.3.1 The Role of Machine Learning in Medical Diagnosis

It has been established that making an accurate diagnosis is a highly complex endeavor. Even under the most favorable conditions, such as the availability of sufficient resources like specialized personnel, medical equipment, test infrastructure, and enough time, there are no guarantees of a correct and timely diagnosis. On one side, it is almost impossible to account for every disease, virus, or condition in existence as they depend on environmental, geographical, sociocultural, ethnical, and other external factors. Even experienced practitioners are unable to know all the causal and correlational relationships between symptoms and health problems.

There are two types of technology systems that have been designed for use in clinical settings: computer-aided detection (CAD) systems and clinical decision support systems (CDSS). CAD systems appeared in 1980 during the second AI spring and were specially designed for diagnostic tasks that required image analysis, like radiology. The primary goal of these systems was to help clinicians avoid overlooking abnormal structures, or areas of abnormality on the image, during routine medical image analysis by increasing the consistency of image interpretation and improving the overall accuracy of image examinations. In this sense, CAD systems were not intended to replace clinicians in any sense, but to act as a "second opinion" instead. The initial CAD systems were developed using the ML methods available at the time, but due to the still limited availability of data and computing power they could only be trained on small datasets which meant, among other things, that the CAD systems could not distinguish well between harmful lesions and other visible patterns in the images. Thus, despite the initial enthusiasm that a computerized system could significantly improve the clinician's work in terms of improved time efficiency and reduced error rate, the performance of CAD systems unfortunately did not meet the expectations of practitioners. An evaluation even estimated that some CAD systems for mammography screening had actually reduced radiologist performance by 19%[45].

[45] Philip M. Tchou et al., "Interpretation Time of Computer-Aided Detection at Screening Mammography," *Radiology* 257, no. 1 (October 2010): 40–46, https://doi.org/10.1148/radiol.10092170.

In this regard, it is important to note that although CAD systems and state-of-the-art ML are used for seemingly similar purposes in medical image analysis, there are two major differences between the two: First, as highlighted above, CAD is developed as an adjunct to the clinician's work and not as an automated system. The clinician is still required to review and analyze the images, and if necessary, he or she can turn to the CAD system for clarification or to resolve a doubt. Second, CAD systems are task-specific, whereas DL architectures can adapt to any given dataset. While the use of more advanced ML techniques may improve the performance of CAD systems, they are at different stages of automation, and therefore there are different normative considerations, for example in terms of responsibility and accountability, as well as transparency, since the decisions belong solely to the clinicians, and thus the consequences of those decisions.

CDSS, in contrast to CAD systems, are more advanced models with a broader reach. They are computerized tools that aim to assist directly with clinical decision making at the point of care. They first appeared around 1970 deriving from the early research and development of expert systems and as such, early versions of CDSS consisted of the same components, i.e., a knowledge base, an inference engine, and an interface to transmit the results to the end user. Another difference from CAD systems is that CDSS focus not only on medical imaging interpretation but considers other types of information and aim to overall improve the decision-making process. Aside from aiding diagnosis, they have been used for complementary tasks such as reminders, alerts, treatment plans, patient education. The role of CDSS is not limited to interaction with clinicians, and it can also be introduced as an element in the interaction between patient and physician, and the surrounding healthy environment.

Similar to expert systems, CDSS faced both technical and epistemological challenges, such as being time consuming for clinicians and other medical professionals, and it was difficult to integrate them into the existing workflows properly. On the other side, CDSS took distance from expert systems in that the former offered relevant information for diagnosticians instead of being tasked with providing a single answer. In this respect, as with CAD systems, the developers of CDSS understood that a knowledge base could not contain all the relevant medical knowledge that was necessary for an accurate diagnosis and posterior treatment. As such, physicians or clinicians were responsible for sorting through the information provided by the CDSS, discarding useless data, and supplementing it with their own diagnostic process and experience. The role of the CDSS was to augment the clinician's work, not to replace or automate it. Besides assisting in the diagnostic decision-making process, CDSS have historically been used for a variety of clinical tasks, including patient safety, clinical management, cost

containment, improved documentation, administrative automation, and workflow improvement[46]. Although CDSS were originally built on knowledge bases, more recent versions of these systems have introduced ML methods. The fundamental change in the newer versions is that the inference engine is not based on IF-THEN rules, but on ML algorithms. There are many documented benefits of using CDSS, including preventing medication errors, reducing duplicate lab requests, and improving prescribing practices[47].

In addition to these two forms of ML integrated into the clinical workflow, there are direct-to-consumer applications based on ML algorithms used for diagnostic assessments. This trend responds to new social, economic, and technological factors that have made it increasingly desirable for people to have more control over their health decisions, and for which the digital market has jumped up to offer new products to meet this growing market demand. Among these factors are the shift from a paternalistic model of medical decision-making to one that encourages patient engagement and autonomy, the pervasive presence of sophisticated mobile phone technology, the effect of the COVID-19 pandemic on the use of telemedicine tools and, possibly, the existing difficulties to have timely access to in-person appointments with specialist doctors. Included in this category are apps that track, record, and manage aspects of a person's health and lifestyle, such as calorie intake, water consumption, fitness, or activity levels, sleep quality, sexual health, medication tracking, and even conditions such as asthma and blood sugar levels.

The problems with these technologies include: first, that an increased focus on patient autonomy does not necessarily translate into better medical literacy and critical judgement. Having more information about one's health does not mean that a person is prepared to understand what the information implies, how to distinguish between serious recommendations and misleading statements (that could have a secondary intention of market research and consumer nudging) and what actions to take. Second, it is uncertain whether these apps are built with robust mechanisms to protect the data of the consumers which, depending on the particular app, could be considered health data. Third, apps in these categories might fall into a regulatory loop because they are often not considered medical devices and thus do not fall within the jurisdiction of regulatory entities like

[46] Reed T. Sutton et al., "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success," *Npj Digital Medicine* 3, no. 1 (February 6, 2020): 1–10, https://doi.org/10.1038/s41746-020-0221-y.

[47] Amit Garg, "Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review," *DECISION SUPPORT*, 2005, 16, https://doi.org/DOI:10.1001/jama.293.10.1223.

the Food and Drug Administration (FDA) or the European Medicines Agency (EMA) and as such, they are not tested to ensure their safety and compliance with proven medical evidence. Fourth, since direct-to-consumer applications are not designed to comply with the strict standards of medical devices, there is a high chance that they would generate false positives. Given that consumers are not trained to understand diagnostic results in context and to discern accurate from inaccurate information, a potential negative consequence would be the burden placed on healthcare systems with a surge of unnecessary medical appointments, over-testing and other associated costs[48].

1.4 The Benefits of Machine Learning in Medical Diagnosis

The technical success of ML models, as it has been shown throughout this chapter, is contingent on their technical possibilities. That is one reason why over-inflating the expectations about their performance is short-sighted and harmful for all actors involved that intend for the models to become a viable and sustainable option in the long term. With this caveat in mind, in this section, I will explore the possible benefits or opportunities of these technologies for the diagnostic process according to their technical possibilities. It must be clarified that a satisfactory answer to the overarching matter of the distribution of potential risks and benefits of ML models must consider not only these technical possibilities but also the equally important questions of who exactly are the recipients of the benefits, whether there are responsibilities to other agents who do not benefit from the use of the models that we should also consider, and what we should do if there are agents who are harmed, either directly or indirectly, by the use of the models. These questions are addressed in Chap. 4 and 5.

The potential benefits explored here fall into two main categories: technical and non-technical. Technical benefits are divided into direct and indirect benefits depending on whether the application contributes directly or indirectly to the diagnostic process. Direct technical benefits are the improved detection of disease and the ability to predict the incidence of a disease. Indirect technical benefits are those that, while not directly related to the cognitive aspect of the diagnosis, provide support for tasks that are often overlooked, but still necessary to bring

[48] Boris Babic et al., "Direct-to-Consumer Medical Machine Learning and Artificial Intelligence Applications," *Nature Machine Intelligence* 3, no. 4 (April 2021): 283–87, https://doi.org/10.1038/s42256-021-00331-0.

about the process with efficiency like medical literature search and summarization and administrative work. Finally, non-technical benefits are those which are not connected to the material aspects of ML models and have an impact on the wellbeing of clinicians, patients, and other agents.

1.4.1 Direct Technical Benefits

The direct benefits of the use of ML methods can be classified according to two objectives: to detect the existence of a disease, even before the symptoms become apparent or to predict the incidence of a disease. One of the most significant areas of focus for research and development of medical ML is the detection of diseases. As has been shown throughout this chapter the detection and diagnosis of diseases is not a straightforward process although there are standard procedures and best practices that follow regulations and that are backed by solid evidence like clinical trials and longitudinal studies. Each disease has its own particular process to arrive at or confirm a potential diagnosis. Skin conditions, for instance, are primarily diagnosed visually. The process includes the initial assessment of the dermatologist, followed by techniques like dermoscopy and if needed, a histopathological examination of a tissue sample obtained through a biopsy[49]. Breast cancer is initially detected if an unusual lump or mass appears in the breast area including the armpits and collarbones. After the PCP or gynecologists examines the patient and makes a referral, the next step is to conduct a mammography, a low dose X-ray test that shows the location, size, shape, and distinctiveness of high-density regions or masses. Defined or distinct masses tend to be benign and indistinct and fuzzy ones are an indication of malignant tumors[50].

Each form of medical testing has its own "gold standard" or diagnostic benchmark that is considered the best available under reasonable conditions. For example, an MRI is the gold standard for diagnosing brain tumors because it provides a superior visual definition of soft tissue, allowing the clinicians to identify the abnormalities with a high degree of certainty, and it is not as invasive and

[49] Andre Esteva et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature* 542, no. 7639 (February 2017): 115–18, https://doi.org/10.1038/nature21056.

[50] "Mammography," National Institute of Biomedical Imaging and Bioengineering, accessed September 16, 2024, https://www.nibib.nih.gov/science-education/science-topics/mammography.

costly as a biopsy, which while providing almost definitive evidence and information than an MRI, it has been deemed less preferable in an evaluation of risks and benefits[51]. New tests are compared and tested against the gold standard to assess their validity and performance. If the test is deemed acceptable, it is added to the procedures, and if the new test is shown to outperform the old gold standard, it may become a new test.

The detection of diseases using ML methods is introduced in the diagnostic process once the patient has initiated the clinical encounter and no conclusive diagnosis has been reached after a first or more consultations. When the clinician determines that further testing or data analysis is necessary, an appointment is set, and ML models may be employed to evaluate the patient's test results and other relevant data. The ML model aids in the diagnostic decision-making process by analyzing the patterns in the patient's test result that might not be immediately apparent through conventional methods. The results are sent to the clinician in charge and he or she is responsible for reviewing this new data and further analyze it against previous information of the patient that the model was not provided with (private patient history, EHR, physician's observations, etc.). The clinician then decides based on all the facts gathered and communicates it to the patient to continue with the treatment process. There are notable advancements on the detection of certain types of cancer disease like breast cancer, prostate cancer and even pancreatic cancer that require early detection to be treatable. The FDA has approved devices powered by such algorithms for commercial use with promising results[52].

ML models for predicting the occurrence of a disease can be used in two ways: for predictions useful for public health and for the implementation of personalized medicine techniques. For public health purposes, ML models can greatly enhance disease surveillance and control efforts by analyzing data sets derived from EHRs, social media, and other sources to monitor and predict disease outbreaks. For example, during the COVID-19 pandemic, ML models were used to predict mortality risk and ICU transfer in hospitalized patients[53]. The aim of these

[51] Nelly Gordillo, Eduard Montseny, and Pilar Sobrevilla, "State of the Art Survey on MRI Brain Tumor Segmentation," *Magnetic Resonance Imaging* 31, no. 8 (October 1, 2013): 1426–38, https://doi.org/10.1016/j.mri.2013.05.002.

[52] Alan Priester et al., "Prediction and Mapping of Intraprostatic Tumor Extent with Artificial Intelligence," *European Urology Open Science* 54 (August 2023): 20–27, https://doi.org/10.1016/j.euros.2023.05.018.

[53] Fu-Yuan Cheng et al., "Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients," *Journal of Clinical Medicine* 9, no. 6 (June 2020): 1668, https://doi.org/10.3390/jcm9061668.

models was to assist public health officials in managing healthcare resources and formulating response strategies.

In terms of personalized medicine, ML models provide tailored risk predictions by analyzing individual patient data, including genetic information, lifestyle factors, and clinical history. These models enable clinicians to customize prevention and treatment plans based on a patient's specific risk profile. For instance, ML models are utilized to forecast cardiovascular risk by analyzing data such as blood pressure, cholesterol levels, and imaging results. Models that combine EHR data with real-time monitoring of vital signs can predict cardiovascular events, allowing for personalized interventions and management plans[54]. Furthermore, ML models in personalized medicine also support the precision of treatment plans. For instance, in the management of diabetes, ML algorithms analyze patient data to predict glycemic control and recommend individualized treatment adjustments, such as changes in medication or lifestyle modifications, to optimize patient outcomes. This personalized approach not only improves the efficacy of treatment but also minimizes adverse effects by tailoring interventions to the specific needs of each patient.

1.4.2 Indirect Technical Benefits

While the applications of ML models of disease detection and risk prediction tend to earn the spotlight of the academic community and health media communications, there are other cases where the implementation of ML models can contribute significantly to the efficiency and effectiveness of the diagnostic process. The indirect technical benefits include medical literature management and tools to assist clinicians with administrative tasks related to or affecting the diagnostic process. As discussed earlier, one of the critical requirements for a diagnostic process is the experience and expertise of the medical professionals involved. For them to develop this experience and expertise, besides the time factor, it is also necessary for them to constantly update their medical knowledge by reading the latest studies, attending conferences, and producing medical research. However, this is a challenging task for two main reasons: first, under normal conditions, practitioners are often pressed for time to perform all of these tasks in addition to their normal clinical and administrative work; and second, the speed

[54] Chayakrit Krittanawong et al., "Artificial Intelligence in Precision Cardiovascular Medicine," *Journal of the American College of Cardiology* 69, no. 21 (May 30, 2017): 2657–64, https://doi.org/10.1016/j.jacc.2017.03.571.

at which medical literature is produced is such that it is difficult for practitioners to keep up to date with the latest medical knowledge[55]. As such, tools to retrieve and summarize relevant evidence from medical literature and tools to create and manage alarms to keep up to date with new publications can help clinicians avoid information overload, optimize the search of relevant information and filter the one that does not fulfill the scientific standards of quality.

Similarly, tools designed for administrative work, particularly to draft reports, could significantly reduce the workload of practitioners and impact positively on burnout rates and clinician turnover. Studies on these applications have shown that the use of NLP models and particularly, fine-tuned LLMs could serve these purposes[56] although there are still several notable downsides, for instance, that the quality of the text generated would only be as good as the text used for fine-tuning. If the data examples given to the model are imprecise or erroneous, the model would reproduce this and could generate a closed loop of unreliable data that, in the end, could be useless or even harmful.

1.4.3 Non-Technical Benefits

Non-technical benefits refer to other advantages that cannot be easily measured by clinical benchmarks, cost-effective analyses or other quantitative methods. They have to do with aspects like the overall wellbeing of the patients, the sense of gratification of clinicians and a generalized positive public perception about the healthcare system. It can also be associated with the aims of medicine, not merely as a field of research or a scientific enterprise, but as a humane endeavor. Examples of these benefits include availability of time during the consultations that the physician can dedicate to form or strengthen the fiduciary bond with the patient or an improved work environment that positively impacts the level of professional satisfaction and fulfillment of clinicians. The benefits in this category are what

[55] This does not take into account the rate at which literature is considered outdated and a potential source of liability when new gold standards or benchmarks emerge, nor the new difficulties posed by the decline in the quality of medical literature and clinical trials, the pervasive effect of fake and predatory journals, and, more recently, the effect of LLMs being used to write articles without any scientific support.

[56] Reece Alexander James Clough et al., "Transforming Healthcare Documentation: Harnessing the Potential of AI to Generate Discharge Summaries," *BJGP Open* 8, no. 1 (April 2024): BJGPO.2023.0116, https://doi.org/10.3399/BJGPO.2023.0116.

Eric Topol calls "the gift of time"[57] that facilitates that the practice of medicine is done in a more humane way, that values the empathy of clinicians towards their patients, that fosters a better relationship between them, and that considers the human needs of the medical professionals as priority and acknowledges its limitations.

[57] Topol, *Deep Medicine*, 286.

2 The Gaps at the Intersection of Healthcare, Machine Learning and Ethics

The exploration of the technical possibilities of ML applied to tasks and problems in medical diagnosis conducted in Chap. 1 provides the basis for the evaluation of risks and benefits in subsequent chapters. This evaluation will be made using a methodological approach based on two main elements: a normative account of moral rights and a type of relational theory. The specific characteristics of this approach are discussed in detail in Chap. 3.

In this chapter, I seek to address two questions: why was this methodological approach chosen? And why is an evaluation of the distribution of potential risks and benefits necessary in the first place? This chapter, thus, focuses on the epistemic justification of this dissertation. In the first part of this chapter, corresponding to sect. 2.2, I will focus on four gaps that I have identified, which describe a series of disconnects, tensions, and contradictions that have emerged throughout the history of the development of the field of AI and that have gained strength in recent decades thanks to the emergence of ML models. They are the epistemic, responsibility, conceptual and implementation gaps, and are framed by two overarching aspects: the interdisciplinary nature of the subject at hand and its particular normative convergence. Interdisciplinarity has two key features. First, it aims to promote and facilitate the integration of two or more fields of knowledge; second, it seeks to establish a new research path that produces novel results or outcomes that would not have been possible if the disciplines involved had worked separately. For example, in the topic of this work, the actors working on medical AI, computer and data science bring to the table the technical capabilities of advanced algorithms, computing power and data storage, while the medical field brings the health data required to create datasets for the model's training and the medical know-how required to conceptualize properly the problems to

L. D. Dias Duran, *Machine Learning in Medical Diagnosis*, Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics,
https://doi.org/10.1007/978-3-662-71357-0_2

be solved. Integrating these elements into an efficient, ethical, and sustainable solution requires interdisciplinary collaboration.

The role of interdisciplinarity, however, goes beyond applied computer or data science in healthcare. The societal, ethical and regulatory challenges of implementing ML models in medical diagnosis require that other disciplines contribute as well. Thus, there is a different kind of interdisciplinarity that requires not only different collaborative technical work, but also one that integrates diversified methodological approaches and techniques. The work of legal studies, cognitive science, and philosophy is critical to address the complex challenges and potential consequences of these new applications. Working in a truly interdisciplinary manner requires certain adjustments and compromises among the disciplines. First, there must be an honest interest in going through the processes of conceptual clarification and coordination, for example, to determine what precisely is meant by notions like a transparent system or algorithmic justice. Second, involved actors must acknowledge the need to design strategies that help bridge the conceptual and practical gaps in between.

There is a further aspect that is similarly important, namely, the convergence of multiple, and sometimes conflicting, aims and interests. First, medicine as a human endeavor has a particular *telos* and professional deontology that establishes codes of conduct, highlights the ethical values to maintain and respect, and enables channels of responsibility and accountability. Medicine's telos can be broadly construed as the good of the patient. While the exact meaning of this notion has been contested and evolved since the formulation of the Hippocratic Oath, it is reasonable to establish that the good of the patient has to do with bringing the ailing person back to a state of health. Pellegrino, for instance, agues in favor of the person as a holistic entity for whom the healing occurs in four hierarchical levels, physical, emotional, human, and spiritual[1].

However, the telos of ML is mostly understood only in terms of technical execution. The value of ML models is defined by the accuracy of their performance and, as such, it has an instrumental value, and it is not subordinated to a higher purpose. This distinction is relevant when the urgency of the consequences of a model failing varies between a negative health outcome for a patient or group

[1] Edmund Pellegrino, "The 'Telos' of Medicine and the Good of the Patient," in *Clinical Bioethics*, ed. David C. Thomasma et al., vol. 26, International Library of Ethics, Law, and the New Medicine (Berlin/Heidelberg: Springer-Verlag, 2005), http://link.springer.com/10.1007/1-4020-3593-4_2, 25.

of patients that could mean a violation of their fundamental rights, and the failure to meet a certain threshold of efficiency and financial feasibility that may have important repercussions for the health institution or technology company. A decision-making approach that focuses only on the success of the model in technical terms risks ignoring the complexity of ethical, social, and economic issues. There is also the perspective of healthcare as an industry in which insurance companies, biomedical vendors, suppliers, and for-profit healthcare institutions (such as investor-owned hospitals and outpatient facilities) are required to find a balance between ensuring the safety of the user[2] and the profitability of their businesses. While it could be argued in general terms that the interests of patients are—or should be—paramount, in practice, this is not a straightforward matter. While the interests or entitlements of patients most assuredly have normative strength, it does not necessarily follow that they should be prioritized in all cases or at all costs.

The implications of the gaps analyzed in this chapter have important consequences. For example, despite the enthusiastic interest of healthcare institutions, governments, and sometimes academia in adopting AI systems in healthcare, and the resulting amount of research conducted, and resources invested, healthcare has been slow to adopt these models compared to other sectors. Some gaps contribute to the slow adoption for various reasons, such as the lack of efficient regulatory frameworks and little clarity about the balance between clinicians' expectations and the actual circumstances in which the models are or will be deployed.

The second issue addressed in this chapter, which corresponds to Sect. 2.2 and contributes to the justification of the rights-based normative approach, is an analysis of the strengths and weaknesses of the main methodologies for research in AI ethics. These are the principlist approach, the embedded ethics approach and the value sensitive design methodology. Although each is used in different settings, they all share a number of difficulties in dealing with issues of practical ethical concern, such as how to address and resolve conflicts between either principles or values, and how to make a balanced assessment of the benefits and potential negative consequences of using ML in medical diagnosis. This analysis will support my proposal that a rights-based approach is appropriate for making such an evaluation.

[2] It should be clarified that in these contexts the role of the patient including his rights and duties changes to one of consumer or user. This is a relevant normative matter that will bear impact on the dynamic of rights and duties and that will be further discussed in Chap. 5.

2.1 The Gaps

2.1.1 The Epistemic Gap

The first gap that I have identified is the "epistemic gap" regarding the capabilities of ML models and their consequences. This gap lies between three opposing attitudes. First, the attitude of people who tend to overestimate what ML can do and will be able to do in the near future; second, an attitude that only recognizes the capabilities of AI only from its narrow conception, but still considers that it should not be used in most areas of society; and third, an attitude that views AI systems as useless.

The first attitude is characterized by two stances. On the one hand, the fear of missing out on the potential gains of developing or using tools marketed as "AI-powered" or "ML-driven". This leads to companies either investing money to introduce these technologies into existing workflows without critically considering whether it is truly necessary or even useful or investing resources to develop tools with the AI label on them expecting to make profit without analyzing whether the tool or service adds value to the users or solves a real problem for users. In other words, without confirming that there is a genuine market demand and not merely a passing trend. On the other hand, the attitude of general acceptance that artificial general intelligence (AGI) is possible and even within reach in the next few decades[3]. AGI is the term used to describe a type of AI that has attained human-level intelligence across a broad scope of cognitive tasks previously only feasible for humans. These cognitive tasks include, among others, human-like ability to reason, extrapolation, problem-solving skills, learning from experience, and generalizing knowledge to new situations. Similar terms like superintelligence, artificial superintelligence (ASI) and safe superintelligence (SSI) have emerged from the basic premise of AGI and extend it to describe an AI system that has surpassed human intelligence.

This first attitude can also be divided into two more or less defined groups: those who consider it positive and desirable to build an AGI that enables the existence of ASI, and those who consider that doing this could pose a threat to human existence. In the first place, the posture in favor suggests that there are significant benefits to society that justify AGI as a goal. They include an increased innovative potential, the ability to solve complex global problems and

[3] Jing Pei et al., "Towards Artificial General Intelligence with Hybrid Tianjic Chip Architecture," *Nature* 572, no. 7767 (August 2019): 106–11, https://doi.org/10.1038/s41586-019-1424-8.

enhance economic growth[4]. OpenAI's statement on the matter says that "If AGI is successfully created, this technology could help us elevate humanity by increasing abundance, turbocharging the global economy, and aiding in the discovery of new scientific knowledge that changes the limits of possibility."[5]. On the other hand, there are those who believe that while AGI is possible, its development could pose serious risks for the survival and flourishing of human civilization. These risks include the possibility of AGI systems becoming independent of human control, developing unsafe goals, using their capabilities for harmful activities, and even overtaking human societies to serve their own interests[6].

The second attitude rejects the feasibility of AGI and recognizes AI systems only as computer programs capable of performing specific tasks. In short, narrow AI. However, this attitude views the development of AI systems as highly problematic or even undesirable in many areas of application and demands that these technologies be strictly regulated or even banned. For example, New York public schools have banned ChatGPT on the grounds that the use of this model would be tantamount to cheating on schoolwork, could negatively impact on the learning process of pupils, and should therefore be prevented[7]. Finally, the third attitude believes not only that AGI is not feasible or desirable, but it also considers that AI technologies are plainly useless, and no effort should be made to implement them.

The main problem with all three attitudes is that they risk ignoring the factual evidence about the functioning of ML models and fail to critically assess the potential benefits and risks accurately. This epistemic gap has several potentially harmful consequences for users, developers, and the public at large. First, a stance fixated on either the unrealistic benefits or the unknown catastrophic consequences of AGI and ASI ignores both the realistic short-term harms that arise from the development of models, such as discriminatory outputs that could affect the ability of a person to access healthcare services, and the realistic benefits,

[4] Fitzgerald McKenna, Aaron Brody, and Seth D. Baum, "2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy," 2020, https://www.ssrn.com/abstract=3070741.

[5] Altman, Sam, "Planning for AGI and Beyond," *Open AI Blog* (blog), February 24, 2023, https://openai.com/index/planning-for-agi-and-beyond/.

[6] Scott McLean et al., "The Risks Associated with Artificial General Intelligence: A Systematic Review," *Journal of Experimental & Theoretical Artificial Intelligence* 35, no. 5 (July 4, 2023): 649–63, https://doi.org/10.1080/0952813X.2021.1964003.

[7] Maya Yang, "New York City Schools Ban AI Chatbot That Writes Essays and Answers Prompts," *The Guardian*, January 6, 2023, sec. US news, https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt.

such as the automation of report summarization for station nurses in a hospital that could allow them up to spend more time with patients or to work shorter shifts.

Second, exaggerated expectations that are later not met, either due to technical limitations or contextual constraints such as existing regulations, can lead to disillusionment. There is a well-known curve that illustrates this cycle of "hype" and disappointment that has historically led to AI winters[8], when investment and research focus rapidly diminish and a sense of fatigue sets in. A sharp and sudden drop in investment can be extremely problematic, as existing projects may not be completed and valuable ideas may be wasted.

Third, overstating the technical possibilities of ML models creates a direct risk to users that deposit their trust in these statements. In 2019, Tesla's Chief Executive Officer, Elon Musk, stated that their autonomous cars could allow the driver to fall asleep safely[9]. Since then, however, there have been numerous cases of accidents caused by the lack of attention from drivers operating a car with self-driving technology in the US[10]. Fourth, overlooking the potential of ML models to address certain issues or take over some tasks that benefit people or groups of people can also be irresponsible and harmful. It is important to acknowledge that there are in fact tasks and processes that can and should be left to ML models. For instance, the discovery of new ways proteins fold has already helped scientists to develop drugs for rare diseases or improved versions of old ones. AI-powered telehealth apps and platforms are being introduced in poor-resource settings where physicians are not available, and diagnosis takes too long to be useful[11]. A critical assessment of ML models is imperative to avoid or minimize the potential harms to people, but this also means recognizing the beneficial aspects, and the balance in between.

---

[8] Ozgur Dedehayir and Martin Steinert, "The Hype Cycle Model: A Review and Future Directions," *Technological Forecasting and Social Change* 108 (July 1, 2016): 28–41, https://doi.org/10.1016/j.techfore.2016.04.005.

[9] Robert Ferris, "Elon Musk: Tesla Will Have All Its Self-Driving Car Features by the End of the Year," *CNBC*, February 19, 2019, sec. Autos, https://www.cnbc.com/2019/02/19/elon-musk-tesla-will-have-all-its-self-driving-car-features-by-the-end-of-the-year.html.

[10] National Highway Traffic Safety Administration, "Summary Report: Standing General Order on Crash Reporting for Automated Driving Systems" (U.S Department of Transportation, June 2022).

[11] Brian Wahl et al., "Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings?," *BMJ Global Health* 3, no. 4 (August 1, 2018): e000798, https://doi.org/10.1136/bmjgh-2018-000798.

The tension between inflated expectations and the realistic possibilities is often exacerbated by affirmations from high end computer scientists such as that ML "might be slightly conscious"[12], to the claim that AI will replace doctors, or that AI will eliminate medical error. There are several issues with these affirmations aside from those already discussed. For instance, that even though ML models can truly be perfectly calibrated, that has no impact on an appropriate delivery of care -which, as argued before, is the telos of medicine. An accurate predictive outcome in medical diagnosis does not tell the clinicians what decision to make to change the outcome or how to balance the patient's goals with the clinical goals[13]. As such, critically assessed and constantly evaluated expectations ought to be the course of action when publicly discussing the possibilities of ML models.

2.1.2 The Responsibility Gap

The second gap lies at the center of the question of who ought to be responsible and accountable in decision-making processes regarding the implementation of ML in health care. This is an important normative issue as the prevalence of ML models in clinical settings increases and the reliance of healthcare professionals on their purported capabilities becomes more pronounced. However, it is necessary to start with a matter of conceptual clarification to distinguish between responsibility and accountability, as they are often used interchangeably or in reference to each other. For example, responsibility is defined as "moral, legal, or mental accountability" by the Merriam-Webster dictionary. However, I hold that there is a significant normative difference between these terms. First, responsibility can be defined as a duty assigned to someone to perform a task satisfactorily according to certain standards. For responsibility to be, in fact, feasible, the responsible person must be the one in control of the tools, mechanisms, resources, and authority to make decisions and take action. Without this control, being responsible is merely a formality and has little normative significance. The sphere of responsibility exists *ex-ante*, i.e., when it is assigned to someone but also remains throughout the development of the task and can be referred to *ex-post*, when the task has already concluded. For instance, someone who is

[12] Cuthbertson Amy, "Scientists Warn New AI May Be 'Slightly Conscious,'" *The Independent*, February 18, 2022, sec. Tech, https://www.independent.co.uk/tech/artificial-intelligence-conciousness-ai-deepmind-b2017393.html.

[13] Jonathan H. Chen and Steven M. Asch, "Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations," *The New England Journal of Medicine* 376, no. 26 (June 29, 2017): 2507–9, https://doi.org/10.1056/NEJMp1702071.

responsible for writing the reports of a project will be acknowledged as such throughout the process of writing the reports and after they are submitted, the person will be considered the one who was responsible, even retrospectively. Accountability, on the other hand, occurs only after a task or process has been completed and an adverse outcome has occurred. As such, there is a legal, moral, or normative imperative to assign blame and perhaps even punishment to someone. For example, if the reports were tampered with or contained an error, the organization would seek to determine who to blame and, if necessary, penalize.

Another important difference between responsibility and accountability is that while the sphere of responsibility can be shared, i.e., a group of people can be in charge of leading a project, accountability is generally assigned either to the person, persons, or legal entity that holds the authority or decision-making power and that can and ought to face the legal or normative consequences of the actions taken or the decisions made[14]. This highlights that accountability, in addition to its normative importance, is also closely related to the notion of liability that has mainly legal implications. Furthermore, although it is common that whoever is held accountable tends to be the one who was assigned responsibility, it is not necessarily always the case. Continuing with the example, if the employee responsible for writing the reports was new in the company and did not yet have the expertise to conduct this task and was nonetheless set to it by his supervisor, the one who would be held accountable would be the latter as he had the actual knowledge about the discrepancy between the employee's abilities and the requirements for the successful completion of the task.

Given the possibility of shared responsibility, the confusion between this concept and accountability can lead to the common issue of "diffused responsibility" or the problem of "too many hands"[15] which ultimately refers to the difficulty

[14] It must be emphasized that the topic of accountability is full of complexities and in some cases, conceptual inconsistencies, depending on the field that takes on defining it and putting it into practice. I take the differentiation proposed by Schillemans and Bovens (see below) between two forms of accountability: as a mechanism and as a virtue. The former refers to a social mechanism by which an agent has an obligation to explain or justify his conduct. The latter is a normative concept for the evaluation of the behavior of public actors. The type that I focus on is the accountability as mechanism, to which I argue that public actors are not the only ones that can and ought to be hold accountable, but rather that any actor who has enough power or authority to make decisions of normative significance within an organization. Taken from: Thomas Schillemans and Mark Bovens, "The Challenge of Multiple Accountability: Does Redundancy Lead to Overload?," in *Accountable Governance: Problems and Promises*, ed. Melvin J. Dubnick and H. George Frederickson, 1st ed. (Routledge, 2011), 19.

[15] Ibo van de Poel, "The Problem of Many Hands," in *Moral Responsibility and the Problem of Many Hands*, by Van De Poel, Lamber Royakkers, and Sjoerd D. Zwart, 1st ed., Routledge

of allocating the blame to a specific person, persons, or legal entity. The source of this confusion might be because of the conflation between these two terms or because responsibility might be understood from both a forward-looking and a back-ward-looking lens. Nevertheless, what might be called backward-looking responsibility is what is understood here as accountability, as argued so far.

In the subject at hand, the responsibility gap, which in fact includes the questions of both responsibility and accountability, arises because of the uncertainty about who ought to be assigned as responsible for relevant normative tasks and who ought to be held accountable if there is an adverse occurrence derived from ML performing tasks that used to belong to persons. It can be argued that these are two separate matters, although closely connected.

Several considerations need to be taken into account. First, the introduction of ML models means that some tasks of normative significance are being assigned to automated models, such as automated diagnosis or automated driving. Prior to the emergence of these technologies, this responsibility was assigned to humans, as were the consequences of any decisions or actions taken. Although there are well-known cases of authority bias, i.e., an unreasonably high confidence in authority figures, which will be presented and discussed in detail in Chap. 4, responsibility has always been in the hands of moral agents. However, tasks with any degree of normative significance left to ML models raise the second normative consideration, namely that ML models are not moral agents in any sense. They do not fulfill basic requirements for moral agency, such as the capacity for deliberation, the ability to justify the outcomes they produce, to offer any kind of amends to someone who has been harmed by them, such as an apology, and they cannot be the recipients of legal action or punishment.

The third consideration is that the state-of-the-art ML models, for the most part, are difficult to scrutinize due to their architectural design. This is known as the "black box" problem and it describes the phenomenon in which a model's behavior, i.e., the outputs that it generates, cannot be explained by looking at its internal technical structure. The problem with this characteristic opacity is that it applied to the AI developers and non-experts. This poses several practical and normative issues as those related to trust, responsibility, and accountability. Algorithmic opacity essentially makes impractical, or even impossible, to know how the model generated a specific output, which means that there is no logical form of reasoning or decision-making that can be comprehended and explained even by the developers themselves.

Studies in Ethics and Moral Theory (New York: Routledge, 2015), 51–54, https://doi.org/10.4324/9781315734217.

In our existing processes for assessing responsibility and accountability, explanations have always been an essential aspect. If a physician's decision during a diagnostic process is deemed to be faulty and results in a negative clinical outcome, the physician will, at a minimum, be subject to an internal investigation to determine his or her reasoning in making the decision. If the physician is found to have been negligent, for example, he or she may be penalized with a suspension or revocation of his or her license, or even be sued. The problem with the outputs generated by an ML model is that there is a general belief that there is no moral agent that can be identified as accountable for the derived negative consequences. The model's developers cannot explain how and why the output was generated, and the physicians using the models cannot explain to the patient why a certain diagnostic recommendation was made. However, at the end, the patient still must bear the consequences of the vagueness in the decision-making process (the consequences of the aspect of explainability for the rights of patients will be discussed in detail in Sect. 3.3.2).

Although this is, in fact, a complex scenario, we can approach responsibility through an interdisciplinary lens. First, tech companies and the AI developers working for them are assigned the responsibility, by the health institutions as clients and by society as interested stakeholders, to design and develop ML models, particularly models in critical infrastructure areas, according to standards for safety and security set by governments and other regulatory entities. This form of responsibility responds to a normative demand, but it is made practical in technical terms. Second, regulators and governmental institutions play a critical role because they are charged with the responsibility of ensuring that standards and regulatory frameworks protect and promote the rights of direct, indirect users and beneficiaries. Third, the health care institutions that purchase ML models for diagnostic purposes and the clinicians who use them bear responsibility for the diagnostic process itself, since, as I argued in Chap. 1, the diagnostic process is much more complex than determining the correct label for a set of symptoms, and therefore responsibility for the process should always be assumed by moral agents.

The introduction of ML models must not prevent or persuade clinicians from acquiring and refining clinical skills and critical thinking. If it is a known fact that state-of-the-art ML models are opaque and nonetheless health institutions decide to use them, then this becomes part of the responsibility they have towards the patients regarding AI implementation. This form of shared responsibility does not ensure that gaps will not arise again but highlights two fundamental aspects:

first, that a complex interdisciplinary challenge cannot be addressed by a single actor or entity, and second, that allocation of duties of responsibility must be adapted to the epistemic possibilities of each actor. It is not reasonable to attempt to only hold tech companies accountable for a negative clinical outcome derived from a ML model. If the model is deployed, it must have gone through a process of testing, approval, certification, and monitoring in which other responsible actors were involved. Of course, there are issues of normative consequence that remain. For example, many of the risks of potential harms that will be presented in Chap. 4 may emerge at some point in the future, but they are slow to materialize and thus there is uncertainty about the timeframes between responsibility and consequence. If a harm or risk of harm are too far removed from the source of the risk or the harm, there are a number of difficulties in holding someone accountable for it, as it may be too late to seek redress, for example because the legal timeframe has passed or because the potentially responsible person or persons are no longer alive. Furthermore, there are also challenges regarding the uncertainty about the magnitude of the damage which also brings new questions about how to deal with risks of harm in healthcare like disability or death that are exceedingly difficult to compensate[16].

2.1.3 The Conceptual Gap

The third gap I have identified is the conceptual gap. It is related to the epistemic gap in that there is a confusion of relevant terminology and that this has normative implications. Unlike the first gap, however, this gap arises between different conceptions and expectations of what AI ethics as a subfield is and should do. The disconnect here concerns the different visions of the role and impact of AI ethics as a discipline. This is considered being of relevance based on observations made during the years of work on this dissertation, which showed that academic discussions, media representations, and public and private efforts regarding the ethical challenges and risks posed by AI revolve around a technical and corporate-oriented mindset, rather than a philosophical one. As such, the focus is mostly on quantifiable concerns such as efficiency and safety. However, as this chapter has shown, the introduction of disruptive technologies also raises important questions

[16] Maria Paola Ferretti, "Risk and Distributive Justice: The Case of Regulating New Technologies," *Science and Engineering Ethics* 16, no. 3 (September 2010): 507, https://doi.org/10.1007/s11948-009-9172-z.

about the morality of health care, such as what medical goals should be and what constitutes good care for patients[17].

There are several weaknesses of this mindset that become apparent when trying to inform regulators and provide a sound foundation for policymaking and public discussion. First, there is the generalized view that AI is an inevitability. It does not question whether an AI system or ML model should be developed or implemented in certain domains or even at all. It focuses instead on attempting to make them compliant with notions of trustworthiness or safety, for instance. Second, it does not consider socioeconomic, political and cultural situatedness[18], this means that this mindset frames ML models as isolated technical artifacts and does not take into account the myriads of factors required for their design, development and implementation, and the wider interconnected impacts of the potential consequences, both positive and negative.

Third, the fixation on values like efficiency, performance and profitability represent mainly technical and corporate interests and disregard others with a social focus like justice, dignity, diversity and inclusion, etc. Fourth, the narrative of this form of AI ethics often has a self-referential state that causes redundancy. This means that the principles used by companies or corporations as cornerstones for their guidelines and codes of ethics are composed, inspired or even outright copied from those of other companies, corporations, or institutions that are aligned to similar interests[19]. This creates a bubble effect in which it is difficult to see the challenges derived from the selected principles like value-conflicts and conceptual disagreement and risks turning a blind eye to other perspectives and alternative solutions.

Fifth, there is no meaningful reflection about the risks and potential harms that could ensue from these technologies. Ethical guidelines are sometimes used as marketing or a sort of checklist method for apparent compliance with existing regulations to avoid legal issues or enhance the companies' public image called "ethics-washing". The practice of ethics-washing consists of using the work of philosophers or ethicists as a form of, or part of, a communications strategy to make a company appear proactive and invested in matters of public concern

[17] Bas de Boer and Olya Kudina, "What Is Morally at Stake When Using Algorithms to Make Medical Diagnoses? Expanding the Discussion beyond Risks and Harms," *Theoretical Medicine and Bioethics* 42, no. 5 (December 1, 2021): 260, https://doi.org/10.1007/s11017-021-09553-0.

[18] Thilo Hagendorff, "Blind Spots in AI Ethics," *AI and Ethics* 2, no. 4 (November 1, 2022): 851–67, https://doi.org/10.1007/s43681-021-00122-8.

[19] Hagendorff, "Blind Spots in AI Ethics," 853.

like privacy and fairness[20]. Ethics-washing can also describe actions taken by companies, like voluntarily hiring an ethics committee or investing resources to design a set of ethical guidelines, as a means of avoiding government regulation. This form of ethics-washing is particularly prominent in the AI field. Companies are seen as willing to direct the development of their technologies "in the right direction" and thus avoid the scrutiny of policymakers and the public. However, at the first sight of needing to address the ethics' experts advice, the companies respond by firing the whistleblowers or shutting down the ethics departments[21].

On the other hand, what I call the philosophical ethics discourse, while aiming to provide a solid analytical foundation for the normative work that characterizes this discipline, also has its own weaknesses. First, it sometimes has gotten sidetracked by an excessive focus on far-fetched consequences of the utilization of powerful AI systems like the scenario of AI-powered machines taking over, the extinction of the humanity or reaching the point of singularity, at which the development of AI technologies becomes unstoppable and irreversible, posing a threat to human civilization. The problems with focusing on these scenarios are that a) there is insufficient evidence that such harms are plausible from a technical perspective. The debate over whether AI systems are intelligent or possess human-like qualities, like consciousness, has been fraught with speculation and cherry-picking examples, leading to the fallacy of incomplete information; b) there is a tendency to ignore actual and short-term risks and harms that affect actual persons (in the sense of persons living in actuality) in favor of the concerns related to existential risk. While this may sound innocuous, the fact is that focusing on existential risk diverts public, media, and even government attention and economic resources that could be devoted to understanding and addressing the actual risks and harms that could have a significant impact in the short term.

While this is not to argue that the field of philosophy, as a whole, should abandon theorizing about possible scenarios or pursuing questions about matters that are deemed of existential importance, it should be strongly emphasized that

[20] Elettra Bietti, "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT*'20: Conference on Fairness, Accountability, and Transparency, Barcelona Spain: ACM, 2020), 210–19, https://doi.org/10.1145/3351095.3372860.

[21] Cade Metz and Daisuke Wakabayashi, "Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.," *The New York Times*, December 3, 2020, sec. Technology, https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html; Gerrit De Vynck and Will Oremus, "As AI Booms, Tech Firms Are Laying off Their Ethicists," *Washington Post*, March 30, 2023, https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/.

such form of philosophical work is not normative in the required sense and there should be careful consideration whether it may pass as such. This means that a body of philosophical theorizing that does not provide a serious justification for its plausibility based on available facts should not be used to inform policymaking or as a body of normative contribution for public discussions. When we intend to do normative ethics, it is essential to distinguish critically what the aims of the fields should be from a moral perspective, and according to what criteria these aims should have priority over others in situations of conflict.

Second, the narrative tends to concentrate mostly on the "diagnostic phase" i.e., on evaluating the problems and potential negative consequences and not so much, or with significant shyness, towards its normative tasks like, as argued in the previous point, determining which factors in these contexts have genuine moral relevance, which ones have not, or perhaps only have a sort of derivative significance. Third, there is an almost exclusive focus on the negative aspects and consequences, which overlooks the fact that a serious ethical evaluation of any applied context requires identifying, assessing, and critically evaluating the beneficial and desirable aspects as well.

The implications of this gap leak into the next gap because the lack of conceptual and procedural clarity hinders the successful implementation of ML models. In the case of healthcare, as will be further evaluated in Chap. 4, this can pose a risk to both patients and clinicians. A useful discourse in AI ethics must be prospective, proactive and normatively critical and independent from private or particularistic interests. It cannot be limited to corroborate "business-as-usual" approaches and must not be used as a form of corporate ethics-washing.

2.1.4 The Implementation Gap

The final gap, and perhaps the most prevalent in the literature, is the implementation gap, also known as the theory to practice gap. It describes the set of barriers that challenge the effective deployment and use of ML models in medical diagnosis. Although this gap is not the same as it was at the beginning of this dissertation in 2021, it remains a complex issue and the subject of ongoing discussion in academic, clinical, and corporate contexts. As shown in Chap. 1, the impulse in research and development of ML for diagnostic purposes has grown and advanced over the years. However, its adoption in clinical settings has not progressed at the same pace. Although at first glance this may seem counter-intuitive given the technical capabilities of ML models, this gap highlights the

complexity not only of clinical problems but also of the intricacies associated with medical practice and existing clinical workflows.

As happened with the others in this section, the implementation gap can be analyzed from two different standpoints. On the one hand, the issue can be framed from the perspective of the technical development in terms of a product that is meant to be deployed and implemented in a particular setting. A technical conception of the implementation of ML models underscores aspects of a similar technical nature with the ultimate aim of making said product operate as designed and provide results in the expected manner and timeframe. On the other hand, this issue viewed from the standpoint of clinicians can be framed as a problem of adoption. This notion, although not entirely divorced from technical aspects, makes a particular emphasis on the process of choosing and accepting a new technology in a clinical context. This distinction, though subtle, frames the implementation gap, as it shows that there are different concerns, priorities, and aims at the point at which a ML model is placed in a hospital, clinic, or praxis.

The challenges at the core of the implementation gap therefore can be classified in three general categories: technical, conceptual and clinical. The technical challenges are directly associated with the performance of the model in clinical settings, this excludes issues particular to the technical design and training of the model and includes only those who have an effect to the clinicians or patients at the point of implementation and adoption. The first technical challenge in this group is the discrepancy between the model's training phase and its actual deployment performance. In simple terms, this issue is the following: while in the training phase the ML model displays a high performance according to metrics like accuracy and sensitivity, the performance decreases in the production or deployment phase. Among the potential causes for this model's behavior, there are two salient options: out-of-distribution (OOD) generalization problems and model drift[22].

OOD generalization refers to the model's ability to adapt to data that has not been used during the training phase and that has a different data distribution (see Sect. 1.2.1). An example commonly used to demonstrate this point is to imagine a model built to distinguish between images of polar bears and panda bears. Typically, images of polar bears have icy or snowy backgrounds, and images of panda bears have green forests with bamboo trees around them. If the model's OOD generalization is high, it will be able to correctly classify images of panda

[22] Joseph Paul Cohen et al., "Problems in the Deployment of Machine-Learned Models in Health Care," *CMAJ: Canadian Medical Association Journal* 193, no. 35 (September 7, 2021): E1391–94, https://doi.org/10.1503/cmaj.202066.

or polar bears with any background (like a zoo background of concrete and metal bars), since the goal of the model is to classify the bears and not any other feature. However, a common issue in ML development is that when the model is presented with an image with features that differ from the images used in the training phase (sometimes these differences might be not clearly discernible to the human eye) the model's performance drops significantly. For example, if the model has a poor OOD, when presented with an image of a polar bear on a grassy background, it will misclassify it. The explanation for this is that the model was focusing on the wrong features of the image, e.g., the background, and thus is not able to classify the image with the correct focus, i.e., the bears.

Model drift occurs when the model is already in the deployment phase, i.e., it is being used in a clinical setting and the performance decreases respective to the training phase when applied to real-world patient data collected under different, perhaps not so ideal, circumstances. This variation can occur between two hospitals using machines with different protocols or with different imaging quality (usually, ML models are trained with high-quality definition images). These technical issues have particularly critical implications in healthcare due to the variance in small features present in medical images that depend on the way the tests might be taken in a particular hospital or by a particular machine. Other technical challenges present here are the lack of understanding about the expectations of clinicians from the side of digital entrepreneurs and developing companies, as well as their lack of knowledge on how the integration of new technologies in clinical workflows is carried out[23] and the uncertainty of clinicians regarding their lack of technical knowledge to operate the medical devices properly and understand the model's outputs.

The conceptual challenges, specific to the implementation gap and not to be confused with the conceptual gap developed in Sect. 2.2.3, shows the problems associated with determining or clarifying what the actual problem to solve *is*. This means finding out the nature of the clinical problem and whether there is a part of it that is susceptible to be solved with AI. Cabitza et al. argue that there is a general lack of consensus on how to assess if the implementation of ML in a clinical setting is successful[24]. This helps to highlight the problematic implications of the

[23] Iredia M. Olaye and Azizi A. Seixas, "The Gap Between AI and Bedside: Participatory Workshop on the Barriers to the Integration, Translation, and Adoption of Digital Health Care and AI Startup Technology Into Clinical Practice," *Journal of Medical Internet Research* 25, no. 1 (May 2, 2023): 5 https://doi.org/10.2196/32962.

[24] Federico Cabitza, Andrea Campagner, and Clara Balsano, "Bridging the 'Last Mile' Gap between AI Implementation and Operation: 'Data Awareness' That Matters," *Annals of*

notion of technological solutionism, a phenomenon characterized by the assumption that social phenomena can be delineated as "discrete processes or problems" that can be optimized or solved by sufficiently advanced technologies[25]. Technological solutionism, according to Morozov, also entails a kind of faith in the possibilities of technology[26] but most importantly, it changes how we approach and understand social phenomena[27]. The downsides of this ideology include that the benefits of these solutions may be overestimated, particularly regarding the infrastructure and institutional contexts (e.g., minimum requirements might not be met in terms of appropriate adaptation to certain contexts).

As will be shown throughout the development of this dissertation, the problems arising from the implementation of ML in medical diagnosis are neither discrete nor easily solved by technology alone. Identifying the nature of the clinical problem that AI is meant to help solve must go beyond meeting only technical benchmarks as a means of proving that a model is useful in clinical settings. There is an urgent need to design and develop comprehensive clinical metrics, physician perception studies, and success evaluations that provide meaningful insights into the needs of the actual end users and beneficiaries of the technologies and the risks they may pose to their rights that are not known to the developing companies and AI developers.

Finally, the clinical challenges emphasize the concerns and questions of clinicians regarding the adoption of new, and potentially disruptive, technologies. Here the concerns are not focused on the technical nuances of ML implementation, but on factors related to how the models can be channeled into routine care. For instance, Liberati et al. emphasize the importance of the negotiation for control, in other words, the perceived amount of control clinicians have over the models and their outputs[28]. The idea of perceived control refers to the amount

Translational Medicine 8, no. 7 (April 2020): 501–510, https://doi.org/10.21037/atm.2020.03.63.

[25] John Gardner and Narelle Warren, "Learning from Deep Brain Stimulation: The Fallacy of Techno-Solutionism and the Need for 'Regimes of Care,'" *Medicine, Health Care and Philosophy* 22, no. 3 (September 1, 2019): 364–65, https://doi.org/10.1007/s11019-018-9858-6.

[26] Technological solutionism is a characteristic of the attitude in favor of AGI and ASI outlined in the previous subsection.

[27] Evgeny Morozov, *To Save Everything, Click Here: The Folly of Technological Solutionism*, Paperback 1. publ (New York, NY: PublicAffairs, 2014), 5.

[28] Elisa G. Liberati et al., "What Hinders the Uptake of Computerized Decision Support Systems in Hospitals? A Qualitative Study and Framework for Implementation," *Implementation Science* 12, no. 1 (December 2017): 113, https://doi.org/10.1186/s13012-017-0644-2.

of autonomy clinicians have in the decision-making process when the system or model is introduced and how much liability and accountability they must face if there is an error or mistake made during a process in which a system or model is involved. Petersson et al highlight, aside from the governance and legal aspects, the factors influencing the capacity of healthcare systems at small and large scale for strategic change management and the potential transformation of the medical profession and of medical practice[29].

It seems clear that clinicians agree that ML models and other AI technologies are able to provide valuable information and tools for some aspects of the diagnostic process and other areas of healthcare. However, the arguments presented in this subsection demonstrate that they are also aware that achieving improvements in clinical outcomes depends on an effective and thoughtful approach to the medical problems as such. Technological tools such as ML models may provide solutions to certain types of problems, but in some cases, it may be preferable to look for alternative methods or approaches that do not rely on AI. Such a decision will depend on concrete normative considerations and an assessment of the distribution of benefits and risks, as will be shown in Chap. 5.

Ultimately, from the analysis of these four gaps, the question arises of whether they convey a problem of trust. The notion of trust in relation to AI development and implementation is constantly subject to debate. Many authors argue that trust is an important prerequisite for the adoption of AI models in different fields. In healthcare, as it will be further discussed in Chap. 4 and 5, trust is essential to enable channels for open and frank communication between healthcare providers and patients and to develop relationships in which patients are willing to share private, but clinically relevant, information to their physicians.

However, trust in the general context of AI, and even more, ML models with their characteristic opacity, raise questions about who should trust whom *or what*? When we speak about the importance of trust for the adoption of ML in medical diagnosis, does it mean that efforts should be directed towards making patients trust the outputs of an automated model? Or does it mean that the clinicians should be able to trust that the clinical device using ML works as intended in the same manner they "trust" that an MRI machine is operating properly? This is a question of layered complexity and, although not the focus of this thesis, will be partly considered throughout the argumentation in the next chapters. For

[29] Lena Petersson et al., "Challenges to Implementing Artificial Intelligence in Healthcare: A Qualitative Interview Study with Healthcare Leaders in Sweden," *BMC Health Services Research* 22, no. 1 (July 1, 2022): 850, https://doi.org/10.1186/s12913-022-08215-8.

now, suffice it to say that trust in a morally relevant sense requires moral agency, which ML models do not have.

2.2 Ethical Approaches to Artificial Intelligence

2.2.1 The Principlist Approach

The first approach that will be discussed in this section is the principlist approach. It must be clarified that this is not a single methodology but the general practice of employing principles as the normative basis for an ethical evaluation. The appeal to principles is not a novel occurrence. They are considered a good starting point to address complex ethical challenges as they provide a general view or categorization of issues at hand. For instance, the principle of fairness in AI encompasses a series of ethical considerations like fair access to the decision-making processes, the importance of a non-discriminatory standpoint in designing AI systems and the allocation of responsibilities, just to name a few. Principles, by virtue of their generality, are useful if attempting to build a consensus or agreement between different parties. This is why so many companies agree to acknowledge certain high-level principles. However, there are also stark weaknesses in the use of principles as a tool for normative work. The major one is that principles are exceedingly difficult to translate into actionable practices.

There are several reasons for this: first, principles understood as general judgements to guide moral action may convey a variety of values, ideologies, and social perceptions. The principle of justice can be understood as instrumental in protecting the value of dignity, from the perspective of Rawlsian theory or as the social notion of fairness in the sense of people should receive what they deserve. This means that as a result of this broadness, principles can mask moral disagreement as well. For instance, principles that are used in international organizations and institutions often are unable to make sense of all the possible differences across populations regarding the meaning and implications of a principle and instead of embodying peoples' interests they risk imposing one of the views to the rest of the populations (i.e., moral paternalism). Whereas principles can be a useful tool to reach an initial agreement, this does not indicate that there is consensus about what the principles actually mean and what the implications of a particular definition could be.

Second, principles are not organized hierarchically. For example, the four principles of biomedical ethics (autonomy, beneficence, non-maleficence, justice) that

are conceived to be in equal standing with each other, perhaps supporting a pluralistic view of morality but with the critical disadvantage that conflicts between the values behind the principles are often not easily resolved. In the context of corporations, Lauer argues that while establishing principles might be an important first step, he warns that "frameworks that fail to account for system-wide complexities will struggle with relevance as the world shifts and changes, and as decisions are made in the face of scarce resources and competing incentives."[30]. Mittelstadt is concerned with what he calls 'virtue-signaling' from companies looking to delay regulation as much as possible by introducing principles which are purposely vague and value statements that fail to provide concrete recommendations about how to proceed in situations where ethical conflict occurs. He also argues that the way principles are conceived, they fail "to address fundamental normative and political tensions embedded in key concepts."[31]. Munn criticizes that ethical principles are meaningless the way they are being proposed because they tend to be isolated, i.e., formulated as single problems, lack enforceability and they often respond to the corporate interests of involved companies[32]. According to this argumentation, principles, by reason of their vagueness and lack of a widespread consensus of their implications, can be used to promote certain interests while undermining the rights of vulnerable groups or unsuspecting parties, as with LLMs like ChatGPT with the case of underpaid workers in African countries being tasked to flag unsafe content[33] or with text-to-image models like Midjourney whose training datasets were found to contain images of minors with links to identifiable information of them[34]. Principles can be defined or interpreted in ways that suit the corporation's interests prompting the practice of "box ticking", i.e., using principles as checklists, as explained previously.

[30] Dave Lauer, "You Cannot Have AI Ethics without Ethics," *AI and Ethics* 1, no. 1 (February 1, 2021): 23, https://doi.org/10.1007/s43681-020-00013-4.

[31] Brent Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence* 1, no. 11 (November 2019): 501, https://doi.org/10.1038/s42256-019-0114-4.

[32] Luke Munn, "The Uselessness of AI Ethics," *AI and Ethics* 3, no. 3 (August 1, 2023): 870–73, https://doi.org/10.1007/s43681-022-00209-w.

[33] Billy Perrigo, "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer," *TIME*, January 18, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/.

[34] Josh Taylor, "Photos of Australian Children Used in Dataset to Train AI, Human Rights Group Says," *The Guardian*, July 2, 2024, sec. Technology, https://www.theguardian.com/technology/article/2024/jul/03/australian-children-used-ai-data-stability-midjourney.

Furthermore, the work of Jobin et al.[35], Morley et al.[36] and most recently Prem[37] demonstrates the attention and resources put into contributing to the criticism of the approach of AI ethics based on principles. In 2019, Jobin and colleagues conducted a scoping review of the existing documents containing ethical principles or guidelines for AI. They mapped the principles identified in the guidelines and concluded that there was a nearly equivalent proportion of documents stemming from public entities and private companies, which suggested that both stakeholders were at the very least interested in the challenges posed by AI. Morley and colleagues elaborated a mapping review of the ethics of AI in healthcare in 2020 with the aim to identify the ethical issues discussed in the existing literature. They selected a corpus of 147 articles and concluded that ethical issues can be of three types: "(a) epistemic, related to misguided, inconclusive, or inscrutable evidence; (b) normative, related to unfair outcomes and transformative effects; or (c) related to traceability"[38]. The issue of traceability refers to the increased difficulty that algorithmic-driven models pose to identify and allocate moral responsibility and legal liability because of unclear workflows and diffused processes[39].

The interest of Prem in 2023 was to expand on the groundwork of Morley et al. to identify the issues in the approaches proposed by guidelines and frameworks to transition from development of AI models to successful implementation and adoption. However, according to Prem, a major issue these guidelines have is that they tend to have a primary focus on developing principles to adhere to, prompting once again the checklist approach. While several authors have identified a similarity of principles in diverse literature corpuses such as transparency, fairness, privacy, and responsibility, etc., they often remain at the conceptual level

[35] Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (September 2, 2019): 389–99, https://doi.org/10.1038/s42256-019-0088-2.

[36] Jessica Morley et al., "The Ethics of AI in Health Care: A Mapping Review," *Social Science & Medicine* 260 (September 2020): 113172, https://doi.org/10.1016/j.socscimed.2020.113172.

[37] Prem, Erich. "From Ethical AI Frameworks to Tools: A Review of Approaches." *AI and Ethics* 3, no. 3 (August 2023): 710–12. https://doi.org/10.1007/s43681-023-00258-9.

[38] Morley et al., "The Ethics of AI in Health Care," 1.

[39] In healthcare settings there are clear workflows that generally allow an investigator to trace the origin of a medical error and therefore, allocate the responsibility of the outcome. With AI models it is not clear as of yet who must bear the accountability and to what extent.

and are unable to provide actionable recommendations for the successful implementation of ML models[40]. A further related issue is that principlist approaches rarely offer a justificatory framework for choosing particular principles instead of others. This adds to the issue of attempting to tailor ethical work to suit specific interests. This must not be confused, however, with adapting or translating high-end concepts into actionable practices. Ultimately, although there are reasons why principles are preferred as cornerstone of normative work in applied contexts, the weaknesses exposed throughout this section open up the interest for alternative methodologies and approaches.

2.2.2 The Embedded Ethics Approach

The embedded ethics approach recognizes that the lack of regulatory frameworks leaves the task of determining what ethical development of technologies should be like to the companies and computer scientists. However, most of the times developers and computer scientists are not trained in ethics and as such they lack the tools, both theoretical and applied, to evaluate what should be done from an ethical perspective. Moreover, although there is an increasing attention towards the work of professional ethicists specialized in the field of artificial intelligence, it is not uncommon that those who work at tech companies developing models for clinical use do not have enough authority to make decisions that can contravene the corporate interests. The embedded approach thus takes two directions. First, it addresses the authority problem, and second, it focuses on the ethical training of computer scientists and developers. In the first direction, the embedded approach advocates for the meaningful integration of ethical, social, and legal considerations into the entire development process of ML models with a strategy based on collaboration and interdisciplinary work[41]. This means that the identification of issues should occur at every phase of the process. In practical terms, the approach proposes that companies should have an in-house ethicist or a team of them dedicated to working hand in hand with the engineering and developers' teams. If this is not possible, then companies should employ the services of an external ethics auditor that is given access to relevant information and provides ethical education to the technical teams. The approach makes emphasis on the role and

[40] Prem, "From Ethical AI Frameworks to Tools," 702.

[41] Stuart McLennan et al., "An Embedded Ethics Approach for AI Development," *Nature Machine Intelligence* 2, no. 9 (September 2020): 488, https://doi.org/10.1038/s42256-020-0214-1.

responsibilities of the ethicists and establishes that he or she should not only have competence in methods of applied ethics but also have domain expertise to provide accurate guidance and to foster interdisciplinary collaboration.

The second direction comes from a virtue ethics view. It proposes that ethical training should be integrated at the university and other higher education training levels, particularly for engineering and computer science careers, i.e., for those students who will be in charge of the development of ML models in the future. In practical terms, this means adding ethics modules in existing curricula or ethical reflection and deliberation exercises in existing courses[42]. The aim of this direction is to make the moral agents, i.e., the engineers and developers, be aware of the ethical challenges present in the design and development of these technologies and equip them with the necessary tools to tackle them. This is an approach with merit, because as Griffin has noted, there is a gap between the expectations placed on these professionals to code ethical considerations and the training they receive[43]. This discrepancy ignores that although the role of developers is indeed relevant, the responsibility of making AI systems ethical is not solely theirs.

However, this approach also has its set of weaknesses. First, the financial burden placed on the companies to hire ethicists or ethical auditors is not inconsequential. It is not clear if this should be a mandatory part of developing certain types of products, for instance, for critical fields like healthcare or warfare, or if this is left at the discretion of the companies which would end up highlighting the issue that this approach aims to tackle, i.e., that companies are left to decide how to interpret principles or develop their own ethical guidelines. Furthermore, ethical issues arise in all fields of application of AI, not only the most obvious ones. Second, receiving ethical training from the ethics experts does not necessarily mean that this will translate in ethical considerations being implemented in practice. Additional training might be deemed too time-consuming for companies or uninteresting for the engineers and developers -it can even discourage them from taking the matter of ethics seriously.

Third, it is unclear whether transparent reporting would be even possible if there were confidentiality clauses involved, and it is also vague how this approach would address the financial interests curtailing the ability of in-house professionals to act with independence. Closely related, the participation of external

[42] Hannah Bleher and Matthias Braun, "Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice," *Science and Engineering Ethics* 29, no. 3 (May 26, 2023): 4–5, https://doi.org/10.1007/s11948-023-00443-3.

[43] Tricia A. Griffin, Brian P. Green, and Jos V.M. Welie, "The Ethical Wisdom of AI Developers," *AI and Ethics*, March 20, 2024, 6–7, https://doi.org/10.1007/s43681-024-00458-x.

auditors would have to be considered against the rights of companies to protect their intellectual property and trade secrets. Fourth, there are no benchmarks to determine whether the ethical work being done throughout the development pipeline is being done properly and is actually guiding the process with appropriate normative aims. It has been suggested by McLennan there is a need for a standard of minimum methodological quality that would evolve as the field progresses, similar to what has happened in the field of bioethics.[44].

2.2.3 The Value Sensitive Design Methodology

The value sensitive design (VSD) methodology is one of the most prominent proposals to carry out normative work in the field of AI. Although it is not novel since it has been applied to different fields of technological innovation in the last three decades, it remains popular in engineering and other technical areas. In general terms, VSD is concerned with designing information and computational systems in accordance with human values. Similar to the embedded approach, VSD focuses on identifying and implementing value-driven decisions throughout the design process, so it also highlights the role of the engineers and computer scientists. However, it does not limit the participation to them alone, instead it looks to involve relevant stakeholders in a wider sense.

The notion of "value" as the basic component of this methodology is defined in a broad sense as what a human or group of humans consider important in life in the seminal work of Friedman and colleagues[45]. A key clarification, however, is that values must be considered from a contextual perspective, which means that they respond not only to an empirical account of the world but also acknowledge the particularities of a given context. A relevant point that the VSD methodology makes is that it recognizes that technology is inherently value-laden. This opposes a widespread conception that technology by virtue of being grounded in mathematics and logic is objective, and therefore morally neutral. VSD thus aims to identify which and whose values are already present in the design process. This reveals that the values encoded in it belong to the people or groups of people who are involved.

[44] McLennan et al., "An Embedded Ethics Approach for AI Development," 490.

[45] Batya Friedman et al., "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory*, ed. Neelke Doorn et al., vol. 16, Philosophy of Engineering and Technology (Dordrecht: Springer Netherlands, 2013), 57.

The methodology of VSD is composed of a tripartite approach. First, the conceptual investigation deals with questions on who the involved parties are, how are they potentially affected by the technologies and how should value tensions and trade-offs be addressed. Furthermore, the approach understands that conceptual clarity is necessary in these kinds of processes and, as such, it is also concerned with offering conceptualizations of specific values that allow to establish a starting point for the value assessment. Second, the empirical investigation looks to address questions about the context in which the technological artifact or system is going to be implemented and how to evaluate if they are being successful. For these purposes, the VSD methodology encourages the use of any quantitative or qualitative method in social sciences that is fitting to gather the relevant information to answer these questions, for instance, surveys, interviews, measurements of human behavior, etc. Finally, the technical investigation proposes the idea that certain technologies are more fitting for certain aims and purposes than others, not only in terms of the technical specifications but also regarding their suitability to support certain desirable values.

One of the weaknesses of VSD is that, while identifying the existing values encoded in the design pipeline is indeed an important task, it is not clear how to include values that do not belong to the groups of people that are *meaningfully* involved in the process. It is a known fact that in technology design and development, and particularly in traditional AI pipelines, the people with authority and decision-making power often belong to homogeneous socioeconomic and ethnic groups. As such, the values represented are those who are important for these stakeholders and they leave aside the values of other, underrepresented individuals, or communities, which are often part of groups of people already at a disadvantage. Given this context, it is unclear how VSD would solve this problem. Another challenge with VSD is that, since it is a methodology that does not ascribe to a particular ethical theory, there is a question about how to reconcile the tensions between the selected values, formulated as universal, and the practical implications of implementing them in particular contexts.

Although the tripartite methodology offers interesting practical steps that can be useful, the action-oriented ones in the empirical investigation seem to assume that stakeholders in positions of power and authority for making decisions would do the right thing for all the other stakeholders involved, even at the cost of some of their own interests. Given the conditions of financial incentive under which a company normally operates, this might be wishful thinking or plain naivete. This does not negate the valuable pathways that VSD opens. Instead, it is a call for considering further manners to secure enforceability, which is why this dissertation focuses on the established strength of rights as the basic normative components of the proposed approach.

Normative Foundations of the Rights-Based Approach

3

As shown in Chap. 1, since deep learning models appeared in 2010 in what is now called the AI spring (in contrast to the AI winters), the field of AI ethics, which is closely related to technology or digital ethics, has grown significantly. According to Borenstein et al., the first article on AI ethics was published in 1985 and in the next 10 years only 6 more articles on the topic were published[1]. This is a stark difference to the number of articles published since 2010 according to Google Scholar. A simple search with the queries "AI" OR "artificial intelligence" AND "ethics" OR "ethical" in the title of articles from 2010 to 2020 yielded over 1.000 results; similarly, a search with the same criteria performed in 2022 alone showed 546 published articles.

Likewise, numerous universities and scientific centers have created specialized programmes and research positions to address the challenges and opportunities of AI in and for society. The Institute for Ethics in AI at Oxford University, the Institute for Ethics in Artificial Intelligence at the TU Munich, the Centre for AI and Digital Ethics at Melbourne University, and the Ethics and Governance of AI research group at the Berkman Klein Center from Harvard University are a few examples of the shift academia has made towards the subject. This goes hand in hand with the establishment of two major academic conferences that serve as a platform for academics and industry researchers to present new findings and analysis regarding AI ethics issues: ACM FAccT (ACM Conference on Fairness,

[1] Jason Borenstein et al., "AI Ethics: A Long History and a Recent Burst of Attention," *Computer* 54, no. 1 (January 2021): 96–102, https://doi.org/10.1109/MC.2020.3034950.

L. D. Dias Duran, *Machine Learning in Medical Diagnosis*, Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics,
https://doi.org/10.1007/978-3-662-71357-0_3

Accountability, and Transparency) and AAAI/ACM AIES (AAAI/ACM Artificial Intelligence, Ethics, and Society)[2].

The technological and digital industry has also displayed increased interest in the contributions of AI ethics. Starting in 2020, major tech companies such as Meta (Facebook at the time), IBM, Apple, Twitter, Google, Amazon, and Microsoft formed dedicated ethics teams or proposed ethics frameworks and guidelines in response to growing public and governmental demand. According to a study made by Accenture, 80% of responding companies "plan to increase investment in Responsible AI, and 77% see regulation of AI as a priority"[3].

However, whether it was because of the economic uncertainty or because these dedicated AI ethics teams were a form of so-called "ethics washing"[4], in 2023 many of these companies fired their ethics teams en masse[5]. Briefly picking up the arguments developed in Chap. 2, there is a gap in the conceptualization and role of ethics as a discipline between stakeholders particular to specific fields of application. In this work, I include medical professionals, patients, and their families or proxies, healthcare officials, hospital, and clinic administrative staff, and AI developers.

This gap makes it difficult to effectively and promptly addressing ethical concerns, as a significant amount of time must be allocated to align the perspectives of relevant stakeholders and ensure a shared understanding. This task has two steps. First, it is an epistemological task, which involves translating normative concepts into a "language" that other experts can understand. Second, it is a methodological task that requires conceptualizing this shared understanding into a format that experts can integrate into the design of AI-driven models and their implementation in clinical contexts. In addition to this main task, there is the time-consuming effort of proving the importance of the role of ethical assessment to companies and AI developers, which are often skeptical and see ethics

[2] ACM stands for Association for Computing Machinery and AAAI stands for Association for the Advancement of Artificial Intelligence.

[3] Eitel-Porter, Ray and Grosskopf, Ulf, "From AI Compliance to Competitive Advantage" (Accenture, June 30, 2022), https://www.accenture.com/us-en/insights/artificial-intelligence/ai-compliance-competitive-advantage.

[4] A term coined after the related term 'greenwashing' in sustainability, which refers to companies that publicise their investment in ethical practices while failing to adhere to the measures proposed by their own internal teams.

[5] Cristina Criddle and Madhumita Murgia, "Big Tech Companies Cut AI Ethics Staff, Raising Safety Concerns," *Financial Times*, March 29, 2023, sec. Artificial Intelligence, https://www.ft.com/content/26372287-6fb3-457b-9e9c-f722027f36b3.

as an obstacle to progress or as impractical for the implementation and adoption of models[6].

Hand in hand with the technical developments in the field of AI, ethics as an applied discipline has also grown since 2010. As of 2023, over a 200 ethical frameworks, guidelines and protocols have been published worldwide by diverse actors in the field, including policymakers, companies, research institutes, universities, expert groups, among others[7] A frequent criticism stemming from the ethics discourse on the application of ML in different areas of society, including healthcare, disputes the idea that, despite AI's clear technical proficiency and potential, as society we should take these models as an inevitability to which we must adapt.

As will be reinforced in this chapter, ethics is a field of individual and collective human choice and the role of an ethical analysis is to first understand if these technologies could realistically benefit us, and if so, assess and advice how and why some paths can be harmful as well. As was examined in Chap. 2, there are serious concerns about the practical advantages of a normative approach based on principles to help resolve ethical conflicts that arise in the implementation of AI systems in healthcare. On that account, I propose that a rights-based approach offers an attractive alternative to addressing these challenges.

The chapter is structured as follows. In Sect. 3.1, I present an outline of the proposed rights-based approach supported in four basic pillars or assumptions about the content of rights: First, every person possesses an equal claim to the necessary conditions to be able to lead their lives. Second, rights are almost always *not* absolute, and they form a hierarchy. Third, rights are simultaneously negative and positive. Fourth, rights require effective protection. In Sect. 3.2, I propose the establishment of a transversal floor based on a broad notion of relationality that highlights the inherent relational nature of rights and the essential relevance to evaluate the challenges posed by implementing ML in diagnosis. Sect. 3.3 will be focused on the matter of the rights that are either affected by or that have emerged due to the interest of developing and implementing ML models in diagnosis and will develop four relevant rights: the right to healthcare, the right to an explanation, the right to a human decision and the right to privacy.

[6] Tricia A. Griffin, Brian Patrick Green, and Jos V. M. Welie, "The Ethical Agency of AI Developers," *AI and Ethics*, January 9, 2023, 7–9, https://doi.org/10.1007/s43681-022-00256-3.

[7] Nicholas Kluge Corrêa et al., "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance," *Patterns* 4, no. 10 (October 13, 2023), https://doi.org/10.1016/j.patter.2023.100857.

3.1 Outline of a Rights-Based Approach

Things being what they are in terms of the philosophical and political debate, it is commonly accepted in most countries that every person possesses fundamental rights and as such, rights have come to be the normative backbone of constitutions, international agreements, and numerous policy bills around the world. While a universally accepted notion about the origin of rights has not been determined, and seems unlikely that it ever will, we can at least be certain of the wide acceptance of the language of rights, i.e., the usage of rights to formulate complaints, concerns, and claims regarding people's fundamental needs.

From a theory based on interests, rights are, essentially, the entitlements of right-holders to certain objects that reflect the (essential) interests that people have and want to protect or promote. Henry Shue argues that "[a] right provides the rational basis for a justified demand."[8]. There are three elements in this phrase to consider, namely, that someone has a demand to something in particular, i.e., an interest; second, that said demand is justified. Finally, that the justification is grounded on a rational consideration. Going further, Wesley Hohfeld formulates the concept of claim-rights and introduces the notion of correlative duties, i.e., that the holder of a right has a claim against *someone else*[9]. For instance, the right to life of person *A* against all other people means a duty of everyone not to kill person *A*. In a more concrete example with the right to life of, say, a pedestrian, it means that every car driver that encounters the pedestrian crossing the street must stop his or her car to avoid crashing into the pedestrian as this would, in all likelihood, harm him or maybe even kill him. The pedestrian, then, has a *claim* against every car driver to refrain from certain actions that would lead to the pedestrian losing his life.

Shue's and Hohfeld's approaches supply us with elements for a first conceptualization of rights: rights as justified demands about some essential interests and correlative duties. Given that the practical appeal of rights has already been established, there will be no further justification for why rights are the normative basis from a formal standpoint. The primary concern of this section, however, is to provide an outline of the content of said rights and the correlative duties they establish, as is usually the main object of normative work in moral philosophy.

[8] Henry Shue, *Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy: 40th Anniversary Edition* (Princeton University Press, 1997), https://doi.org/10.1515/9780691200835, 13.

[9] Wesley Newcomb Hohfeld, "Some Fundamental Legal Conceptions as Applied in Judicial Reasoning," *The Yale Law Journal* 23, no. 1 (1913): 16, https://doi.org/10.2307/785533.

While this dissertation will not provide a comprehensive account of the definitions and discussions held throughout the history of philosophy about the nature of ethics, it seems relevant to begin, at the very least, offering a basic idea of what is meant here by the notion of ethics and what is meant here as a rights-based ethical approach and how it's suitable to examine the challenges posed by ML implementation in the medical domain.

Ethics, as an established academic discipline stemming from philosophy, deals with the analysis, conceptualization, and categorization of matters related to morality. Moral questions, norms, and judgements are directly or indirectly concerned with the matter of how agents ought to behave. The answer to these -and further- questions and concerns will differ depending on the particular approach of a moral theory. While a utilitarian theorist will appeal to a maximization of benefit, a virtue ethicist, on the other hand, will insist on the centrality of the agent acting according to a particular desired virtue or set of virtues.

Regarding the position of this dissertation, the proposed rights-based theory is concerned with determining what the interests of relevant agents are with respect to the implementation of ML in medical diagnosis, how these interests weigh against each other, what factors must be considered in making a trade-off, and what the potential ethical consequences of the trade-off might be. In other words, which and whose interests should be considered and prioritized, and to what extent. This endeavor is critical given that the interest in the application and development of ML models in clinical settings has created new ethical challenges and has added new levels of complexity to existing ethical challenges. While this will be discussed in detail in the next chapter, suffice it to say that to tackle the task of determining who are the relevant agents whose interests may conflict with the interests of other relevant agents, it is first necessary to establish a layout with pillars that provide normative justification. In the following sections, I follow the philosophical theories of Alan Gewirth and Klaus Steigleder on the subject of rights.

3.1.1 The First Pillar: The Necessary Conditions

The first pillar is grounded on the basic assumption that every person has an equal claim to the necessary conditions to be able to lead their life. This assertion is built on the argument that, fundamentally, all individuals possess equal moral relevance and therefore, by acknowledging this fact it follows that we

also acknowledge that other people's interests are as normatively relevant as our own[10].

The meaning of necessary conditions in this context can be understood as comparable to the concrete biological requirements for any complex carbon-based life form on earth to exist, survive and thrive, such as water, sunlight, and nourishment of some sort. In a similar fashion, every person has certain basic requirements to be able to lead their lives. Moral philosophers have considered this matter at length. For instance, Shue speaks of three basic rights as the absolute limits from which society cannot sink: Security, subsistence, and liberty[11]. Similarly, Gewirth speaks of two kinds of generic rights contained in his principle of generic consistency, namely, well-being and freedom. He conceptualizes them as generic insofar they contain the generic conditions necessary for any kind of agency, in other words, the conditions to be able to act and act successfully at all[12].

It seems, in principle, that it is straightforward and reasonable to start from the assumption that there are, or at least there must be, some necessary conditions to be able to lead one's life. The necessary conditions the first pillar alludes to are, first of all, and in this order: life, physical integrity, psychological integrity, and freedom. From a practical standpoint, this is also a reasonable statement, as these are contained in constitutions and human rights declarations. The first pillar, then, establishes that every person possesses equal moral relevance, and as such, we all must therefore have equal claim to those necessary conditions to be able to lead our lives.

From a conceptual perspective, the first necessary condition for anyone to be able to lead their life is, without a doubt, to be alive. Therefore, it logically follows that every person must have a right to life as a precondition to be able to pursue any activity. This has a primary direct implication: all people have a claim to non-interference against all other people in actions that may lead to losing their life. In other words, people must refrain from actions that would likely reprieve or restrict others from the ability to remain alive, for instance, driving recklessly, or building fires in unregulated sites.

This first implication is valid for the other necessary conditions. However, a first caveat to make here is that despite the fundamental quality of these rights, they are not always absolute. Even more, it shall be argued that they are *mostly*

[10] Klaus Steigleder, "On the Criteria of the Rightful Imposition of Otherwise Impermissible Risks," *Ethical Perspectives*, no. 3 (2018): 471–95, https://doi.org/10.2143/EP.25.3.3285426, 473.

[11] Shue, *Basic Rights,* 18–29.

[12] Gewirth, *Reason and Morality*, 63–75.

not absolute, even in the seemingly straightforward examples[13]. The notion of absolute rights implies that they cannot, under *any* circumstances, be interfered with. However, rights *can* and *have* come into conflict in many instances. This erodes this incontrovertible assumption, or at the very least, provides grounds for further reflection about absolutist postures about rights. In fact, one could argue that a fundamental justification of frameworks in normative moral philosophy is to provide guidance to resolve those conflicts. To give an example, in certain situations of self-defense it is morally permissible that a person retorts to killing someone who is severely threatening her life, or even the life of someone else who cannot defend himself, like a child. Even at the level of criminal law, such an action can be legally justified if it can be proven that there was enough reason to believe that there was imminent risk of death or severe harm.

An equally crucial second caveat is that, while the element of "necessity" in the first pillar means that the necessary conditions are indispensable to each and every person, this applies, as Steigleder puts it "insofar as they are all on the whole indispensable"[14]. This indicates that the necessary conditions, *situationally,* have degrees of "indispensability" relative to the impact they have on the overall ability of a person to lead his or her life, ranging from absolutely indispensable to temporarily dispensable at the other end of the spectrum. It must be emphasized that this dispensability is always provisional, and it should only be resorted to under circumstances where the normally indispensable rights are at conflict with each other and therefore a resolution is needed, for if a right could be dispensed easily then we would contradict a fundamental aspect of the first pillar as outlined in the previous subsection: that these conditions are indeed necessary for a person to be able to lead her life. The reasoning behind such a caveat is, thus, grounded in the argument that while each and every person has a claim to the necessary conditions of the first pillar, depending on situational considerations, rights may need to be weighed against each other, which logically leads to the conclusion that rights are mostly not absolute. In turn, this begs the question as to how to go about determining criteria for situational indispensability and to address it seems helpful to introduce the idea that rights form a hierarchy in situations of conflicting interest.

[13] Alan Gewirth, "Are There Any Absolute Rights?," *The Philosophical Quarterly (1950-)* 31, no. 122 (1981): 1–16, https://doi.org/10.2307/2218674.

[14] Steigleder, "On the Criteria of the Rightful Imposition of Otherwise Impermissible Risks," 474.

3.1.2 The Second Pillar: A Hierarchy of Rights

The matter of how to go about determining criteria for the application of such a hierarchy is complex. In order to tackle this task, I take Judith Thomson's distinction between the *infringement* and the *violation* of a right. Thomson starts establishing that an infringement of a right occurs when a person *B* performs an action that conflicts with a certain claim a person *A* has: "Suppose that someone has a right that such-and-such shall not be the case. I shall say that we infringe a right of his if and only if we bring about that it is the case"[15]. She offers the example of a person *A* that has certain medicine locked down in a box and a person *B* breaks it open without permission to access the medicine with the purpose of using it. In this case, the right to property of person *A* has been infringed. As a result, person *A* no longer has the medicine, and her box is broken.

However, Thomson clarifies that the infringement of a right does not mean that the right was *violated.* She continues "(…) I shall say that we violate a right of his if and only if *both* we bring about that it is the case, and we act wrongly in doing so."[16]. She introduces the idea that an action would only violate a right if the infringement is wrongful or unjust. In her example, Thomson expands that person *B* has broken the box and taken the medicine with the purpose of giving it to a dying child who needs the medicine to survive. She adds that the only accessible dose of medicine is the one person *A* possesses, and that there is no time to try to get it from somewhere else.

The introduction to the argument of the child in dire need of the medicine establishes a special consideration, namely, that there can be a sufficiently good reason to justify a person's otherwise impermissible action. Thomson suggests that an infringement occurs only when rights conflict with each other, as the right to life of the child conflicts with the right to property of person *A*. In this example, even though the right of person *A* was infringed, person *B* did not act wrongfully from a moral perspective, since he was acting to protect the *situationally* more fundamental right to life of the dying child. Now, let us take an alternative path on this example: if person *B* would have instead broken the box and stolen the medicine with the purpose of selling it to a third party because he wanted to get

[15] Judith Jarvis Thomson and William Parent, *Rights, Restitution, and Risk: Essays, in Moral Theory* (Cambridge, Mass: Harvard University Press, 1986), 49.

[16] Thomson and Parent, Rights, Restitution, and Risk, 51.

the money to buy a new car, the action is no longer permissible and we would speak of a violation of person's *A* right to property, as there is no compelling justification for person's *B* actions. These examples help to cement the argument that a hierarchy of rights is essential to the structure of a rights-based approach in normative terms.

However, there are some criticisms raised in the literature about the seeming incompatibility between a notion of a hierarchy of rights and fundamental rights as inalienable, equal to all persons, and indivisible as inscribed in, for instance, the Universal Declaration of Human Rights (UDHR). Although this thesis does not strictly use the UDHR's definition of rights as a basis, it seems relevant to clarify this apparent discrepancy to show the importance of the hierarchy for normative purposes.

The preamble of the UDHR states that "Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world"[17]. There are three aspects to highlight here: first, when we speak of inherent dignity, we refer to the metaphysical value of a human person that belongs to her by virtue of simply being a person. Second, the notion of inalienability explains that rights cannot be removed from persons nor surrendered or transferred. Finally, the equality of rights implies that the rights in the declaration are of equal status.

While such postures seek to assert the, on the whole correct, basic assumption that rights must be held equally by everyone without discrimination, which aligns with the first pillar, it does not account for the practical problems that arise when rights come into conflict. It is important to make the clarification that human rights as they appear in declarations, international treaties, covenants, and other agreements are principles that function as instruments of national and international laws to protect the fundamental interests of people under the authority of countries which sign them. Human rights are based on moral rights, which in turn, are grounded in moral reasons, values, or standards; however, not all moral considerations are codified as human rights and thus protected by law. Thus, when I speak of rights in general, I refer to moral rights and not international human rights unless otherwise indicated.

The introduction of a hierarchy aims to provide a normative solution for cases of conflict, and it does not contradict the idea that people must not suffer from

[17] United Nations, "Universal Declaration of Human Rights" (United Nations), accessed September 18, 2024, https://www.un.org/en/about-us/universal-declaration-of-human-rights.

discrimination and their rights must be effectively protected. As it is, the concept of hierarchy does not imply that rights are, from the start, organized according to their importance. The first pillar, i.e., the equal claim to the necessary conditions to be able to lead one's life, states clearly that these rights are indeed indispensable. However, two caveats were presented, namely, that rights are almost always not absolute and second, in case where there is a dilemma or conflict of rights, we must evaluate the *situational* indispensability of the rights at play.

The hierarchy of rights aims to ensure that the equal claim to the necessary conditions is fulfilled as it prevents an individualistic approach that puts the interests of one person above another. As put by Steigleder "It is the gist of a rights-based approach that focuses on the normative importance of each and any person that people must not be sacrificed for the well-being of other people and, insofar as it presupposes a hierarchy or ranking of rights, that rights can at best be outweighed by rights of equal or greater importance"[18]. As a simple example, even if non-essential property is considered a necessary condition as per the first pillar, when in conflict with the right to life, the indispensability of the right to property is adjusted to be of lower score than the right to life, as on the whole, a person can still situationally act with their right to property infringed but a person whose life is threatened cannot. Furthermore, a hierarchy of rights serves as an alternative to the weaknesses of the principlist approach explored in Sect. 2.2.1 because while principlism would use different methods to resolve a conflict of values like specification and balancing, it is often done in a casuistic manner and thus can be time-consuming and open to too much deliberation among interested parties. This poses a clear problem in practical settings, where translating the principles into actionable procedures that can guide morally critical decisions would become impractical and even ineffectual. The advantage offered by a hierarchy of rights is the existence of an established basis to settle most clear-cut cases of conflict, leaving only edge or outlier cases where a more in-depth analysis is needed.

[18] Klaus Steigleder, "Climate Risks, Climate Economics, and the Foundations of Rights-Based Risk Ethics," *Journal of Human Rights* 15, no. 2 (April 2, 2016): 259, https://doi.org/10.1080/14754835.2015.1083849.

3.1.3 The Third Pillar: Negative and Positive Rights

The third pillar of this justificatory structure of rights is proposed by several works of Steigleder[19] who builds on the theories of Gewirth[20], and is concerned with the dual nature of rights, namely, that rights are simultaneously negative *and* positive. The conception of rights that most moral theories propose is, generally, one of non-interference. This means, in broad terms, that a right claim imposes certain restrictions on the duty-bearers in their ability to act in a certain way. For example, the right to freedom in its negative form indicates that a person *B* must not, under normal circumstances, restrain a person *A* against her wishes. Similarly, the right to physical integrity means that a bicycle rider must not run his neighbor over while crossing the street because he is in a rush to get to a work meeting for which he was late.

Positive rights, on the other hand, require proactive action from the duty-bearer. Positive rights exist at two levels: individual and institutional. Positive rights at the individual level are observed in situations where a person is in acute need of help. Let us imagine a person, Alice, in a situation where her life is in danger and another person, Brett, is close by and bears witness to Alice's struggle. Brett, who acknowledges that Alice's life has equal moral importance as his own, rushes to save her. This positive action is considered a duty to the individual in the present proposal of rights-based theory, insofar two conditions are met. In the first place, that the right-holder, Alice, is unable to help herself, and in second place, that the assistance required comes at no comparable cost to Brett, the duty-bearer[21]. Both Alice and Brett have an equal claim to the necessary conditions to be able to lead their lives and in this situation, if Brett does not assist Alice, her life will be endangered, and as a consequence, she will not be able to lead her life.

Now, in a situation where the above conditions are not met, there is no individual positive duty. For example, if Alice was in a burning house and there was

[19] Steigleder, "Climate Risks, Climate Economics, and the Foundations of Rights-Based Risk Ethics.", Steigleder, "On the Criteria of the Rightful Imposition of Otherwise Impermissible Risks," and Klaus Steigleder and Johannes Graf Keyserlingk, "Public Tasks During Contagious Disease Pandemics: A Rights-Based Perspective," in *Ethical Public Health Policy Within Pandemics: Theory and Practice in Ethical Pandemic Administration*, ed. Michael Boylan, The International Library of Bioethics (Cham: Springer International Publishing, 2022), 149–66, https://doi.org/10.1007/978-3-030-99692-5_8.

[20] Gewirth, Reason and Morality and Gewirth, The Community of Rights.

[21] Gewirth, *Reason and Morality,* 218; Steigleder and Graf Keyserlingk, "Public Tasks During Contagious Disease Pandemics," 153.

no way for Brett to save her without possibly dying himself, then there is no positive duty to help Alice other than to call the fire brigade and perhaps try to put out the fire to the best of his ability. The "no comparable cost" condition is intended to argue that if the positive duty required Brett to try to save Alice regardless of personal danger, this would mean that Brett's right to life and physical integrity is less valuable relative to Alice's right. This contradicts the equality of rights articulated in the first pillar. This is similarly applicable to the second condition, namely that Alice would be able to help herself by, for example, using an emergency exit leading to a functional staircase outside. As argued earlier, the first pillar states that everyone has an equal claim to the conditions necessary to live their lives. So, if Brett is expected to take time to find a way to reach Alice to help her when an emergency staircase is available his rights are less weighty in comparison to Alice's. In other words, if a person is expected to restrict his ability to act in favor of someone who is not helpless, all things being equal, this would give greater value to the rights of the person being helped.

This conceptualization of rights as positive must go in accordance with the first two pillars, and it is only justifiable if one first accepts them as preconditions. Otherwise, the theory would fall into the indefensible position where some lives could be classed as normatively weightier than the lives of some others. It is of utmost importance to emphasize this point before moving forward, since it has a deep impact at the moment of making an analysis of the risk to the rights of, for instance, patients that are subjected to or are active users of technology driven by AI systems.

The introduction of new technologies, especially ones with a pervasive effect on society as AI has already proven to be, have created a set of ethical issues where we are confronted with the task of evaluating in which cases a restriction is necessary, but also, in which cases, we actually do need to put these technologies to use. As we will see in Chap. 4, there are situations where the rights of patients might be put at risk if we do not employ AI systems but also situations where the models, while seemingly highly accurate in prediction tasks for certain populations, are so drastically inaccurate for other groups that it is simply not permissible to deploy models in clinical settings as they are and they must be adjusted to ensure that no population group is disproportionally affected.

3.1.4 The Fourth Pillar: Institutional Protection

In the previous section, it was mentioned that positive duties derived from positive rights could be divided into individual and institutional. Regarding the institutional level, positive rights relate to the need for effective protection that is only

feasible through institutional action. While the definition and reach of so-called effective protection of rights is one that has been widely debated[22], it is assumed here that the first pillar demands that these necessary conditions be effectively secured, in other words, that people are in practice secured in their ability to be able to lead their own lives. It is the case that in many middle- and low-income countries in the world, but also in high-income nations, there are groups of individuals whose rights to the necessary conditions are not secured effectively. This means that while normatively these individuals *are* right-holders, in practice, for a variety of reasons, those rights are not actually protected. Since such a situation of vulnerability is systemic, it follows that no individual duty-bearer could be asked to be responsible for it, as any help would be first insufficient; and second, it would be a burden that would affect the duty-bearer's own ability to lead his life.

As such, it is necessary to establish institutions dedicated to ensuring the effective protection. The concrete meaning of how to go about this task and who should oversee it is a matter that requires further discussion and does not fit the scope of this project. However, it can be said that such endeavor requires robust collaboration between national institutions, and even in some particular cases, international intervention, and cooperation. Phrased differently, governments and even some supranational and international institutions have a positive duty to assist people who are endangered in their necessary conditions and cannot otherwise help themselves protect them or regain them. Moreover, while the effective protection of rights is summarily a direct duty of institutions, individuals have the duty to support them as well, for instance, by paying taxes.

It must be noted, however, that the institutionalized fulfillment of positive rights does not mean that the two conditions by which positive rights are established become invalid, namely, that assistance does not come at comparable cost and that assistance should only be a duty when the person cannot help herself. At the level of institutional duty to assistance, the content of the notion of comparable cost and its practical implications would need to be revised and adapted to the circumstances of the country or institution in particular[23].

---

[22] Linda Reif, "Building Democratic Institutions: The Role of National Human Rights Institutions in Good Governance and Human Rights Protection," Harvard H.R.J, 2000, 1–69; Bundesverfassungsgericht, "Zur Gewährleistung wirkungsvollen Grundrechtsschutzes bei der Übertragung von Hoheitsrechten an supranationale Organisationen," July 24, 2018, https://www.bverfg.de/e/rs20180724_2bvr196109.html.

[23] Steigleder, "On the Criteria of the Rightful Imposition of Otherwise Impermissible Risks," 474.

3.2 The Transversal Floor: Relationality

So far, this section has sought to establish the outline of a four-pillar, rights-based approach, following the work of Steigleder, who in turn builds on the moral theory of Alan Gewirth. Fundamentally, these pillars are grounded on the notion of purposive agency presented by Gewirth who argues that the ability to act successfully is the precondition for any discussion about morality[24]. However, although it is plausible to consider agency as an attribute of human beings, action itself does not occur as a single, isolated phenomenon. When a person acts, there are multiple factors at play, like intention, execution, direction, and consequences. If a person walks from a starting point-A to a destination point-B, this action requires that she first establishes the intention of covering this distance by using her legs, that she decides what is the destination, that the action is executed by moving the legs, and finally, that she faces the consequences derived of the action. She might have stepped on a piece of glass, causing herself injury, or she had muddy shoes, and she caused someone else's floor to become dirty.

Following this argument, within the framework of a rights-based theory, acting is only susceptible of normative relevance if the action is understood at the center of the interaction between, at least, two individuals, as only then it is possible to speak of a right-holder and a duty-bearer. This element of 'interaction' is central to the approach proposed here, although it goes beyond the correlative connection that arises from the right-claim owned by the right-holder and that enacts the obligations placed on the duty-bearer.

What I propose here is to consider a broad notion of relationality as a transversal floor where the four pillars rest on. This notion of relationality is derived from theories of relational ethics in the field of nursing. Bergum and Dossetor, for instance, propose an approach that explores in detail the relationships in which relevant moral considerations occur, paying particular consideration to not only the actions performed by the agents but also how their roles (of patient, physician, nurse, etc.) make connections that are important at a normative level[25]. On a more semantic note, the Oxford dictionary defines relationality as "Relating to or characterized by relation; that relates two or more things; expressing a relationship"[26]. This basic definition offers two relevant elements, one of them

[24] Gewirth explains that action in the sense that matters for morality, is characterized by two generic features: freedom and intentionality. (See: Gewirth, *Reason and Morality*, 27).

[25] Vangie Bergum and John Dossetor, *Relational Ethics: The Full Meaning of Respect* (Independently published, 2020).

[26] Oxford English Dictionary, "Relational, Adj., Sense 2" (Oxford University Press, September 2024), Oxford English Dictionary, https://doi.org/10.1093/OED/5956498667.

mentioned previously: that there are at least two individuals, or objects of some sort, involved, and that the primary interest is *how* they (agents and/or objects) are connected to each other.

Although for western philosophy the concept of relationality is relatively new, it has been studied at large in a range of disciplines such as physics and computer science, and interestingly, in other non-western philosophical traditions[27]. Recently, due to the interest of humanities and social sciences in decolonial perspectives to a variety of matters, a formal concept of relationality began to appear in multiple academic fields like psychology, sociology, and philosophy, often introduced from sources of indigenous scholarship grounded, for instance, in Ubuntu philosophy and Aboriginal traditions in Australia. In sociology, relationality is defined as "signify[ing] the ongoing, contextually specific processes through which cultures, bodies, practices, and subjectivities are (re) constituted and gain meaning through discursive-material encounters with other cultures, bodies, practices, and subjectivities"[28]. Similarly, Kan and Lejano, speak of relationality in the context of social ecology as "(…) a theory that underscores how social connectedness, through mechanisms of empathy, foster collective action in non-centralized modes of network governance."[29].

Although closely related, the concept of relationality is not identical to the notions of mutuality and reciprocity and this needs to be differentiated. Reciprocity is the concept farthest away from relationality and refers to a type of symmetrical relationship between two agents that implies a comparable exchange of benefits. Reciprocity is characterized by its temporal nature, which implies that there was a prior benefit that initiated the relationship, and the obligations are exclusive to the person who provided that benefit. Moreover, a reciprocal action is ruled by proportionality. The moral obligation to reciprocate extends

[27] Sabelo Mhlambi, "From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance," *Carr Center for Human Rights Policy*, Human Rights Policy Discussion, July 8, 2020, 1–27; Thaddeus Metz and Sarah Clark Miller, "Relational Ethics," in *The International Encyclopedia of Ethics*, ed. Hugh LaFollette (Blackwell, 2013), 1–10, https://philarchive.org/rec/METR-7; Craig K. Ihara, "Are Individual Rights Necessary? A Confucian Perspective," in *Confucian Ethics: A Comparative Study of Self, Autonomy, and Community*, ed. David B. Wong and Kwong-Loi Shun (Cambridge: Cambridge University Press, 2004), 11–30, https://doi.org/10.1017/CBO9780511606960.003.

[28] Carolyn Pedwell, *Feminism, Culture and Embodied Practice: The Rhetorics of Comparison*, First issued in paperback, Transformations (London: Routledge, 2012), 41.

[29] Wing Shan Kan and Raul P. Lejano, "Relationality: The Role of Connectedness in the Social Ecology of Resilience," *International Journal of Environmental Research and Public Health* 20, no. 5 (February 22, 2023): 1, https://doi.org/10.3390/ijerph20053865.

only in proportion to what it was received in the first place. An example of a reciprocal exchange occurs when a person A helps a neighbor B, for instance, cleaning the snow piled at the entrance of his house because the neighbor needs to be elsewhere. Person A does not necessarily have a duty to help his neighbor, but he recalls that, a few weeks ago, the neighbor took the trash out for both when he forgot to do it. Person A reciprocates the neighbor's good action with a comparable good action. Two elements stand out in this exchange: first, that the person's action is primarily motivated by the existence of a previous good action of the neighbor, and second, that the action performed in return is of comparable value. Mutuality, on the other hand, does not conform with the proportional principle of reciprocity, and does not require a preceding action for its existence. Instead, we speak here of a shared experience where all participants enjoy a certain equality and affect each other simultaneously. Gewirth explains that an exchange based on mutuality expands beyond the constraints of benefactor and beneficiary and is not concerned with proportionality as all persons carry the roles of right-holders and duty-bearers[30]. A general theory of relationality contemplates both types of relationships as existing and necessary in different scenarios. However, the theory of relationality as transversal floor for the proposed rights-based approach to the evaluation of benefits and risks of ML models in diagnosis only draws near to the notion of mutuality as it explains the type of normative relationship under the first pillar.

3.2.1 Criticism to the Atomistic Conceptualization of Rights

In a systematic manner, the four pillars formulate an alternative to the approach to rights that focuses solely on the individual right-holder and on the duties that are derived from the claim he or she makes. The common way to address infringements or violations of rights consists in establishing a unidirectional dynamic between the duty-bearer and the right-holder, in which their roles are static within the specific context of the right that is being discussed.

However, this dynamic is in fact a relationship that is formed from the moment a person is born. It connects normatively two individuals and establishes specific obligations. This is what Jean Thomas calls formal relationality: "In the case of rights, because they are directed obligations from particular persons toward particular persons, the action required or prohibited by the obligation constitutes

[30] Gewirth, The Community of Rights, 75–79.

the other pole of the constraint and thus the relationality implied by the form of rights."[31].

The criticism of this approach to rights has three facets. First, it points to the emphasis on rights having an individualistic or atomistic character. As normative elements, moral rights are conceived as justification for the protection of the basic requirements of any given person based on deontological considerations like personhood, dignity, autonomy, etc., or an analysis of benefits in a utilitarian fashion that considers the overall welfare obtained from the promotion of a particular right, and the analysis is frequently calculated as an aggregation of individual benefit. In both conceptions, rights are held by the person as an individual unit only and do not consider the normative importance of the relationships between agents.

Regarding this, Ho et al., allude to the divisions among society to find consensus in what values or interests should be considered rights, and which established rights should prevail in situations of conflicting claims[32]. An example of such situations occurred during the COVID-19 pandemic, where certain decisions were made prioritizing the interest of certain individuals to the detriment of other individuals but ignoring the normatively meaningful relationships between them, which might require consideration for other morally relevant factors. Ho et al. also make a point on the conception of rights only from their negative facet. They argue that in many instances, human rights allow certain parties to assert their interest against others, to limit their degree of freedom[33]. While it is true that rights require that we consider to which extent we must limit our freedoms in regard to justified essential interests of right-holders, this cannot be solely a unidirectional exercise. Both parties remain equally relevant right-holders and their equal claim to the necessary conditions must be taken into consideration. As such, rights should not be used as a protective bubble for the individual interests of person A, to which the interests of others are merely secondary.

In a similar fashion, Jonathan Herring criticizes that although one of the strengths of a rights-based approach is that a particular case can be observed from each of the involved parties' point of view, it often does so by isolating the

[31] Jean Thomas, *Public Rights, Private Relations*, 1. ed (Oxford: Oxford Univ. Press, 2015), 151.

[32] Emily Ho et al., "Relational Rights" (Relational Research, Discussion Paper, 2021,). ISBN 978-1-3999-0447-6. https://www.relationalresearch.org/product/relational-rights-book/.

[33] Ho et al., "Relational Rights", 32.

rights from each other and in doing so, it can overlook the harms that systematically affect already disadvantaged populations[34]. This is particularly relevant for assessing the risk of harm and actual harms derived from the application of ML in medical diagnosis. As will be discussed at length in Chap. 4, models that fail to adequately represent minorities in their predictions will perpetuate biases and result, for example, in incorrect or delayed diagnosis.

The second criticism asserts that an individualistic approach to rights often displays difficulties in offering guidance in situations where rights conflict and where some sort of weighing up or prioritization is required, particularly regarding the fulfillment of positive rights[35]. It was remarked before that the proposed rights approach does not consider rights to be absolute in most cases and it has been shown that conflict of rights is commonplace, especially when looking at large-scale social and moral challenges that affect individuals and groups of individuals differently, as is the case with AI technologies. The relational focus, in combination with the other four pillars, offers a reasonable alternative to the criticism of this atomistic approach to rights. Instead of treating conflicting right cases like a zero-sum analysis where the expectation of ethical resolution is the prevalence of the claim to non-interference against the duty bearer, the relational view first looks at the conflict as situated within an ecosystem of interdependent agents, who possess both claims to the necessary conditions and also responsibilities towards each other justified by the acceptance of the equal moral relevance of every person.

Lastly, the third problem of the individualistic approaches to rights is the failure to properly acknowledge the normative importance that relationships have and to include them as relevant factors to consider in rights-based normative frameworks. This might come across as counterintuitive, because it could be argued that any discussion about rights emerges from social and political contexts that demand that certain interests be recognized and protected. However, the traditional moral discourse often sets limits on an individual's pursuit of personal goals and interests but does not recognize the intrinsic value in relationships or collective goods. Jennifer Nedelsky considers this issue from a different perspective. According to her, a relational approach allows us to rediscover an already present quality of rights, i.e., that they build relationships, for instance, of power,

[34] Jonathan Herring, "Forging a Relational Approach: Best Interests or Human Rights?," *Medical Law International* 13, no. 1 (March 1, 2013): 32–54, https://doi.org/10.1177/0968533213486542.

[35] Christine Susienka, "Human Responsibilities: A Relational Account of Human Rights" (PhD diss., Columbia University, 2017), 16, https://doi.org/10.7916/D84J0SP8.

responsibility, and obligation[36]. This implies an understanding of rights not as the instruments from which relations between right-holders and duty bearers arise, but rather that rights are the product of relationality as an intrinsic property of persons. This point could be criticized on the grounds that relationality runs the risk of presenting rights as entirely relative to particular cultural norms, and thus superfluous as a normative basis for a universal moral theory. The relational approach proposed here, however, argues that rights derive *from* the kinds of relationships that are only possible among humans but that all humans are inherently capable of. In that sense, the relational aspect conserves a quasi-naturalistic note that points at a universalist intent, i.e., that is universally applicable to all humans. It does not, though, reduce the potential normative relevance of other types of agents or right-holders.

While these three critical points elaborated here offer a plausible justification for a relational view on rights that acts as a transversal floor of the proposed approach to analyze the risks of AI-driven models in medical diagnosis, there are three clarifications to be made before moving forward. Firstly, a relational approach does not negate that the individual is the starting point of all discussions about rights and duties. All relationships are based on individuals in connection with others and forming different levels of moral relationships that require consideration when evaluating the imposition of a risk or the protection of a right. Secondly, the criticism presented here does not exhaust the more severe disagreements towards rights approaches in general, like those of some postures in ethics of care and feminist ethics. However, the aim to add relationality as a common ground does not intend to dilute the strength of the claims of right-holders or to negate the proven benefits that rights have brought about in the protection of fundamental claims to the necessary conditions. Instead, the aim is to expand the understanding and meaning of what rights in context truly imply, require, and aid, in this scenario, different stakeholders in decision-making processes regarding the design and development of ML models, the usage of models in clinical and educational settings, and the elaboration of regulatory mechanisms.

Finally, the third clarification is that the notion of a hierarchy of rights as proposed in Sect. 3.1.2 does not dispute the relational approach. While at first it might seem that a hierarchy establishes a power dynamic that contradicts the fundaments of relationality, the way the hierarchy is conceptualized here actually implies a vertical order of relations. An interesting example of a similar conception can be seen in some instances in the tradition of Confucianism, particularly

[36] Jennifer Nedelsky, "Reconceiving Rights and Constitutionalism," *Journal of Human Rights* 7, no. 2 (June 17, 2008): 145, https://doi.org/10.1080/14754830802071950.

regarding the theory of the Three Bonds. This theory sets up the importance of the notion of the hierarchical roles of sovereign/subject, parents/children, husband/wife[37]. While the original texts and interpretations have problematic elements of sexism and social injustice that are justly criticized in our day and time, modern interpretations made by eastern ethicists show that what we can gain from these notions is that a hierarchy is not necessarily a matter of (imbalanced) power, but instead of distribution of responsibility and temporal or situational ability to fend for oneself[38].

In an example from the field of healthcare, the relationship between physicians and patients usually implies a distribution of authority and responsibility that could be considered hierarchical. That is why a fiduciary bond can, in principle, exist between them. As will be developed in more detail in Chap. 5, this hierarchical order does not mean that the physician has more power or is justified in imposing his will on the patient, but rather that the relationship between them establishes a distribution of responsibilities in which the physician is responsible for guiding and making decisions on behalf of the patient (to a certain extent) because of the specialized knowledge of the medical field, the training, and the experience the physician has. In short, elements that are crucial when diagnosing a patient and that he (the patient) does not possess under normal circumstances. Meanwhile, the patient places his or her trust in the physician and is responsible for providing him or her with truthful accounts of all relevant information related to the medical condition or disease to be diagnosed and treated.

A relational approach would put emphasis, not in the patient's rights as a single individual or the corresponding duties of the physician as another single individual, but in the relationship between them where rights and duties do play a role, but where they are considered stemming from both physician and patient, in other words, the relational approach requires the normative groundwork laid by the first pillar and also the establishment of a collective or shared goal, that in this example is to find a solution together where the patient is correctly diagnosed.

The relevance of introducing a relational component as cornerstone to the rights-based approach thus lies in the acknowledgment of the normative relevance of not only the agents that form any sort of relationship, but the relationship itself. Relationality in the context proposed here looks to make an emphasis on the specific relationships between agents and recipients of rights and the roles that form as a result of those contextual relationships. This acknowledgement

[37] Dau-Lin, Hsü. "THE MYTH OF THE 'FIVE HUMAN RELATIONS' OF CONFUCIUS." Monumenta Serica 29 (1970): 33–35. http://www.jstor.org/stable/40725916.

[38] Metz and Miller, "Relational Ethics," 3–4.

is compatible with the four pillars and provides further normative stability to the framework for evaluating the risks of ML in medical diagnosis that will be introduced in Chap. 5.

3.2.2 Conceptual Justification

It has already been affirmed in this section that I hold that relationships between human beings are normatively relevant, and as such must be considered for evaluating the violation or infringement of a right, as well as possible constraints or special obligations. There are at least five arguments to justify this assertion.

A first argument in favor of relationality having normative importance is that moral rights can only be understood, from a *practical* level, as situated within society. While perhaps a meta-ethical account of rights is concerned with questions like the epistemological or ontological feasibility of rights, this dissertation assumes a posture of rights as normative, i.e., regarding the question of how we *ought* to act. From this perspective, one cannot speak of right-claims if there is no "other" towards whom the claim is directed. Furthermore, if we go a step forward and consider the third pillar, i.e., positive rights, and the fourth pillar, i.e., the moral imperative of the effective protection of a right, this requires a third party or "observer" that validates the claim of the right-holder and the defensibility of the duty placed on the duty-bearer, e.g., the police, judges, mediators, etc.

The second argument is rooted in the fact that human beings are located within a network of diverse social connections and interactions from the moment we are born. Those connections have a social and moral component and create contexts with specific sets of conditions that reflect on the claims to rights they may have. Thus, rights as objects of morality can only be truly comprehended at the backdrop of a society, i.e., a collective of individuals connected through some form of relation, for example, familial, social, political. In other words, individuals can be understood as nodes connected to each other, and they are influenced or even determined by their placement within their networks, their attributes, and the connections to other nodes, i.e., individuals.

The relationships between individuals or nodes entail a set of social and moral rules, and in some cases, even entail special obligations and considerations. Furthermore, even in the legal space, rights as instruments of the law can only be practically enforced within specific contexts. While laws embody a general understanding of how certain matters of society should be regulated, the concrete application of the law is a subject of discussion and debate in every state

under the rule of law. That is why an essential part of a state is the judicial system, where judges, juries, attorneys, and lawyers, among others, are in charge of assessing, defending, and deciding how the laws apply in particular cases.

Third, relationships are contextual and, as such, the distribution, or the degree of "indispensability" of duties and rights is not necessarily equal in all instances -that is partly why a distinction between violation and infringement is necessary-, although there are always clear-cut cases that delineate the limits of the relationships. This assertion might be criticized as suggesting that, if we view relationships as entirely contextual, then there is no way to look beyond the specific case to make general assessments that properly support a normative approach. However, relationality functions in a dual way. It is an internal property in the specific context and an external property that establishes the general moral factors, such as, the special obligations in certain types of relations, as in the case of the fiduciary relationship between clinicians and patients or the familial bond between parents and their children.

A fourth argument for a relational approach is the potential expansion of the moral ecosystem. Human beings as moral agents form relationships of normative relevance with other agents, but additionally, this approach also acknowledges the possibility that other morally relevant agents or entities can enter or participate in relevant relationships with human beings. This would suggest that agents can possess right-claims against them or duties towards them. Here we find, for instance, future generations and, in some respects, sentient animals. An expansion of the moral ecosystem could also include considering normative questions about moral patiency and other forms of potential moral significance as is the case in some subfields of ethics about animals in general, the environment[39] and, recently, the discussion about morality and robots, especially those being used in healthcare and warfare[40]. Although this topic will not be touched upon in this thesis, it opens an interesting research path. An important caveat is that this conception of relationality acknowledges that are situations where a person might be in a situation or state of reduced agency. Such is the case for small children and infants, people in prolonged states of unconsciousness, as well as people suffering illnesses or impairments that affect the possibility of action and/or of holding accountability for it.

[39] For more on this matter, see: Brock Bastian et al., "The Moral Significance of Protecting Environmental and Cultural Objects," *PLOS ONE* 18, no. 2 (February 9, 2023): 1–22, https://doi.org/10.1371/journal.pone.0280393, and Martin Schönfeld, "Who or What Has Moral Standing?," *American Philosophical Quarterly* 29, no. 4 (1992): 353–62.

[40] John Danaher, "The Rise of the Robots and the Crisis of Moral Patiency," *AI & SOCIETY* 34, no. 1 (March 1, 2019): 129–36, https://doi.org/10.1007/s00146-017-0773-9.

Finally, the relational floor reinforces an argument that has been hinted at throughout this section and that follows Gewirth's explanation of mutuality as a quality of rights within the context of the community of rights. In a similar fashion, it is defended here that all agents are simultaneously right-holders and duty-bearers, partly grounded in the recognition of the equal claim to the necessary conditions but also due to the relational nature of rights developed in this section. The assessment of a violation or an infringement of a right is not done in the unidirectional way as is common in traditional theories of rights (agent A performs an action that violate/infringes the right of person B, who has a claim against agent A preforming said action). The following questions must be considered: How are A and B connected? Does this connection entail special obligations? What other rights-holders are associated with A or B? Does A have a claim against B to allow him to perform the action?

3.3 Rights at the Intersection of Machine Learning in Medical Diagnosis

The impact of new technologies, in particular ML models, in healthcare is a fact that must be acknowledged, for only this way it is possible to move ahead of unfounded postures of enthusiasm or skepticism and see the changes for what they are and evaluate the consequences either positive or harmful. As it can be surmised, the emergence of digital and information technology and the myriads of applications in all aspects of human society have, since decades back, created the necessity to consider a new way to formulate and delineate the reach and limits of rights. Even though, as I have endeavored to show throughout this chapter, rights are essential normative elements to protect the legitimate interests of people, the interest of governments and companies to employ AI-driven technologies for purposes of normative relevance like surveillance, warfare, and healthcare, has brought up a renewed discussion about how we should handle the effective protection of existing rights, and has prompted another one about the necessity to create new.

Considering this, the aim of this section is to examine if and how existing rights have been redefined and whether new rights have emerged as a consequence of these advancements, and if so, if they can be considered as justified. The broad concern of this examination is how these technologies impact on the complex structure of rights within the context of healthcare, and whether there are legitimate interests that have emerged because of the implementation of ML and how we ought to protect them. It must be clarified that this is not an attempt

at a systematic review of the literature about rights in medical diagnosis or to provide a deep analysis of all nuances of rights within the context of artificial intelligence. Instead, I intend to offer an analytical evaluation of the following rights: the right to healthcare, the right to an explanation, the right to a human decision and the right to privacy.

3.3.1 The Right to Healthcare

The notion of a right to health is acknowledged and generally accepted from a moral point of view, as well as by national and international laws and treaties with practical ramifications. The Article 25.1 of the Universal Declaration of Human Rights states that "Everyone has the right to a standard of living adequate for the health of himself and of his family, including food, clothing, housing, and medical care and necessary social services"[41]. Similarly, the International Covenant on Economic, Social, and Cultural Rights (ICESCR) adopted in 1966 offered a more comprehensive call for a right to health: "The States Parties to the present Covenant recognize the right of everyone to the enjoyment of the highest attainable standard of physical and mental health", and "(c) The prevention, treatment, and control of epidemic, endemic, occupational and other diseases; (d) The creation of conditions which would assure to all medical service and medical attention in the event of sickness."[42].

These treaties declare the fundamental importance of health for humans and there is no controversy there. However, a right to health does not imply a right of persons to *be healthy*, for someone to be considered healthy depends on numerous factors, some of which are not necessarily or at all within the powers of a state to ensure or fulfill, like genetic conditions, pre-existing illnesses, accidents, not man-made natural disasters, among others. This has been addressed by the treaties. The General comment Nº 14 of the ICESCR clarifies that: "the entitlements include the right to a system of health protection which provides equality of opportunity for people to enjoy the highest attainable level of health."[43]. The notion of "highest attainable level of health" affords states a certain margin of action to implement measures for the fulfillment of the treaty. As can be easily

[41] United Nations, "Universal Declaration of Human Rights" (United Nations), accessed September 18, 2024, https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[42] UN General Assembly, "International Covenant on Economic, Social and Cultural Rights," accessed September 18, 2024, https://www.refworld.org/legal/agreements/unga/1966/en/33423.

[43] "General comment No. 14: The right to the highest attainable", p. 3.

imagined, the highest attainable level of health can look quite different according to a country's resources and conditions, which grants states the chance to adapt the treaty requirements to their realistic possibilities. This has not been interpreted as simply abandoning the idea of setting a minimum standard of what the highest level attainable should be, on the contrary, it means that countries with stable healthcare infrastructures and means to do so assist countries with precarious conditions[44]. This concept of a highest level attainable is thus normatively useful as it not only considers the content of the right but the considerations regarding the burdens imposed on the duty-bearers as well.

However, a delimitation is needed here. A right to health as conceived by the treaties, even with the nuanced notion of highest level attainable, is too ambiguous for the normative scope of this dissertation. Instead, the focus here is on the justification of a right to healthcare, understood as one of the requirements of the right to health in the broader sense. A second reason for this delimitation is that the analysis of the provision of a right to health in full would require to contemplate ML models and systems that are extraneous to medical diagnosis or diagnostic procedures, for instance, models to manage patient data, performing robot surgery, augmented drug discovery and others. Such a spectrum of applications falls outside the scope of this dissertation. Now, regarding a possible moral right to healthcare, we must start with the question of what a right to health implies. This question can be divided into two more specific normative considerations. First, what interests a right to healthcare protects, and second, which duties would such a right impose on others?

To answer the first question, the four-pillar approach is useful. Within the proposed rights-based approach, a right to health is located undoubtedly within the necessary conditions to be able to lead one's life. An initial basic assumption to this respect is that certain physical or mental diseases can severely incapacitate a suffering person in their faculty to act, think, move, communicate, etc. In short, we can safely affirm that people require a degree of health to be able to lead their lives, i.e., a degree of physical and psychological integrity. This is contemplated in the first pillar as necessary conditions.

[44] For more information about concrete case studies that bring valuable insight on this matter, see: Paul Hunt and Gunilla Backman, "Health Systems and the Right to the Highest Attainable Standard of Health," *Health and Human Rights* 10, no. 1 (2008): 81–92, https://doi.org/10.2307/20460089; Paul Hunt, "The Human Right to the Highest Attainable Standard of Health: New Opportunities and Challenges," *Transactions of The Royal Society of Tropical Medicine and Hygiene* 100, no. 7 (July 1, 2006): 603–7, https://doi.org/10.1016/j.trstmh.2006.03.001.

It could be argued that not every sickness or disease prevents a person from leading his or her life. Someone with a common cold, while surely experiencing discomfort, still can perform most actions reasonably. That is, in fact, true. However, a person suffering from, for instance, fibromyalgia, a chronic disorder that causes widespread pain and fatigue and that has no cure, would see her life severely impaired. Establishing the threshold of health someone needs to be able to lead a life is a complex normative matter, and frequently, it might require situational evaluation. However, it might be reasonable to put forward some essential factors without which it is not possible for anyone to lead a life, such as being conscious, having a degree of awareness of the surroundings and the circumstances and having the ability to communicate their opinions, wishes, preferences and decisions.

This argument offers an initial justification in favor of a right to healthcare and tentatively establishes a duty of states and -perhaps- international institutions to provide persons suffering from a condition or disease that severely impairs their ability to lead a life the access to a system that can provide treatment or a cure so that they can regain this ability. However, a right only for persons that are impaired by a disease is not sufficient as it would imply that a healthy person would not have a right to healthcare, which is problematic. There are a couple of arguments to consider here. First, if a right were to apply exclusively to a certain group of people, then it would cease to be fundamental, in the sense proposed by the first pillar. However, even if a right pertains to certain circumstances which are a possibility for every person, the right is universal but only pertinent if these circumstances occur. In this sense, a right to healthcare is fundamental for all people and thus a right even if some persons perhaps never are in need of it, as has been argued so far.

Second, if a person suffers from a disease and requires medical assistance, this presupposes that she was healthy first, given that, by definition, a disease is defined as the loss or deterioration of an initial state of health. Considering that the process of deterioration of cells that leads to aging also leads to an increased chance of disease, any person has a chance to suffer from any type of disease with increased likelihood as she ages and as such, any person ought to have a right to healthcare regardless of their state of health, although, they would only be entitled to the positive protection of the right when the conditions where accessing the healthcare systems are met, namely, that the person *requires* care.

Third, a person might be born with a disease or a disorder that increases the likelihood of suffering from diseases throughout her life. In many cases, the disease or disorder will severely impair the person's ability to lead her life, e.g.,

a congenital anomaly of the nervous system like anencephaly[45]. In some other cases, however, the person will be able to lead a life with minor restrictions. Depending on the particular circumstances of each case, persons suffering from diseases or disorders that are present from birth would still have a right to health care, but the rights would be realized in different ways. For some patients, the right to health care would mean receiving the care necessary to lead a dignified life, while for others it would mean having access to palliative and hospice care, and/or psychological and bereavement counseling. In scenarios like these, the normative evaluation of the distribution of benefits and risks would have to consider factors like the material possibilities of the healthcare institution or the healthcare system, the prognosis of the patient and the chances that a medical intervention would in fact benefit the patients and not harm them[46].

Finally, a fourth point that needs to be addressed concerns the preventive aspect of health care. I argue that health care as a system has not only a duty to treat people who are or will be sick, but also a positive duty to implement strategies to prevent people from getting sick in the first place. This is in the best interest of people, since in general no person wants to be ill, and in the interest of the healthcare system, as it is a duty of the state to effectively protect the interests of its citizens. Furthermore, since intensive interventions and highly specialized treatments tend to be expensive, the use of effective preventive medicine can avoid a high volume of intensive interventions required to fulfill the duty of the system towards right-holders[47] and can act as a cost-effective measure. Thus, if people have a right to health care and the health care system has a duty to try

[45] Anencephaly is an example of a severe congenital birth anomaly characterized by the absence of a major portion of an infant's brain and skull. There is no cure or treatment, and the survival rate after a few weeks is under 5%. Given the extreme nature of the anomaly, the infant will not be able grow up and lead a life. In these cases, it is standard medical procedure to provide counseling, support and palliative care for the patient in the neonatal care unit. For more on this subject, see: P.A. Baird and A.D. Sadovnick, "Survival in Infants with Anencephaly," *Clinical Pediatrics* 23, no. 5 (May 1, 1984): 268–71, https://doi.org/10.1177/000992288402300505 and Shandeigh N. Berry, "Providing Palliative Care to Neonates With Anencephaly in the Home Setting," *Journal of Hospice & Palliative Nursing* 23, no. 4 (August 2021): 367, https://doi.org/10.1097/NJH.0000000000000770.

[46] Determining what constitutes benefit or harm with respect to medical interventions is a highly complex task that requires first of all addressing the question of the aims of medicine as a human endeavor. In some cases, prolonging artificially sustained life may be considered harmful, while in other scenarios it may be determined that doing otherwise, i.e., withdrawing life support, is unethical.

[47] Joshua Cohen and Peter Neumann, "Cost Savings and Cost-Effectiveness of Clinical Preventive Care," The Synthesis Project (The Robert Wood Johnson Foundation, September 1, 2009), 17–21.

to prevent the onset of disease, then every person, regardless of his or her health status, has a right to access the preventive resources of the health care system. In sum, every person has a right to the best minimum attainable level of healthcare available to her, which includes protection against the spread of disease.

The right to healthcare has negative and positive normative implications, according to the third pillar. On the one side, every person has a right *not* to be prevented from accessing the healthcare system, for example, as a result of any form of discrimination or as punishment (for instance, for people in jail or detention) or to be deliberately subjected to poor delivery of care. On the other side, the effective protection of this right, i.e., its positive implication, depends on the concrete requirement of the person, namely, what kind of care she needs and to what extent. This means that people do not necessarily have a right to every treatment available at every point in time. For example, a person who is in good health and whose physician determines she is, in fact, healthy, does not have a claim to costly medical imaging tests, like a brain MRI, even if she wishes to check that her brain is healthy. If this were the case, and the physician would comply with her wish, then this limited resource would be spent on someone who does not need it instead of being used on someone who does. The allocation of limited resources in healthcare implies a normative evaluation of what is the most justifiable way to spend them and doing so on a patient that has no need for them is not normatively permissible.

Regarding the content of the right, aside from the one encompassed in the justification above, there is a further factor to consider: that there are special obligations towards right-holders such as newborns, children and elderly persons, who are in phases of life where their vulnerability to exogenous and endogenous factors is higher and thus their ability to lead their life could be impaired if access to healthcare is not available or if the quality or availability of necessary conditions is not present or is not adequate. These special considerations would include their right to have the necessary means to achieve at least an average life expectancy.

This brings us to the matter of the duties that such a right would impose on others and the question of *how* this right would be protected. To this respect, I draw from the four basic factors required for the realization of a right to health proposed by the ICESCR in the General Comment Nº 14[48] and developed as a

[48] UN Committee on Economic and Social and Cultural Rights (22nd sess: 2000: Geneva), "General Comment No. 14 (2000), The Right to the Highest Attainable Standard of Health (Article 12 of the International Covenant on Economic, Social and Cultural Rights)," August 11, 2000, https://digitallibrary.un.org/record/425041.

toolbox by the Danish Institute for Human Rights[49]. This framework identifies four key elements for the realization and operationalization of social, economic, and cultural rights: Availability, accessibility, acceptability, and quality (AAAQ), Although the framework has been used until now to frame the right to water, it can be applied to the right to healthcare as well. First, the element of availability is concerned with the existence of facilities and resources such as hospitals or clinics, basic medical equipment, specialized medical equipment, qualified personnel, medications, sanitization elements. Accessibility refers to the material and non-material factors that make it possible that people to use and benefit from the facilities and resources. Material accessibility includes monetary affordability of medical expenses, physical infrastructure (ramps, elevators, sizeable doors, handrails, etc.) for patients with disabilities, injuries, reduced mobility, etc.; non-material accessibility refers to relevant medical or general information in a readable language and format and equality of access (non-discrimination).

Third, the element of acceptability requires that the protocols, workflows, design of facilities, and other administrative and logistic plans are made in consideration for the ethical norms, cultural contexts, societal perceptions belonging to the population where the facilities are, etc. Finally, quality involves considerations of safety, effectiveness, efficiency, adaptability, timely delivery of care and/or treatment, integrated procedures, and equitable provision of services. These elements must be understood from the concept of "highest attainable level of health", or, in other words, it must consider the material possibilities of the context from where the assessment is made. It might be said that such a provision weakens or even stripes the quality of universality off a right, however, a lack of sensitivity to the situation of a particular country or region would cripple the practical application of the framework and the effective protection of the right. The compliance with these elements is a positive duty of the state. International and supranational institutions, although not directly responsible for the practical implementations of these measures, could have duties related to providing supervision, guidance, and mediation to achieve this implementation. Finally, the duties of individuals, as stated in the fourth pillar, are to support the healthcare system by complying with the statutory health contributions and not exposing themselves to unnecessary risks that could cause accidents.

Now, a realistic approach of a right to healthcare must also recognize the challenges for its effective protection. The healthcare systems around the world

[49] Mads Holst Jensen, Marie Villumsen, and Thomas Døcker Petersen, The AAAQ Framework and the Right to Water: International Indicators for Availability, Accessibility, Acceptability and Quality; an Issue Paper of the AAAQ Toolbox (Copenhagen: Danish Institute for Human Rights, 2014).

have been facing crises of staffing, growing demand for care, and increasing healthcare worker's burnout rates for some time. We are also facing the challenges derived from a rapidly increasing ageing population that, historically, requires the most resource-intensive care in terms of intensive treatment and long-term care. Studies have shown that healthcare resource expenditure increases significantly in the last two years of life[50], peaking in the last year[51]. If we consider that the WHO has estimated that between 2015 and 2050 the proportion of the world's population over 60 years will nearly double from 12% to 22%[52], the challenges to maintain the quality of access and coverage are urgent[53].

While AI is not the silver bullet to address these challenges, the technical capabilities showed by certain models are seen as an alternative to assist and augment clinicians in making more accurate diagnoses and handling repetitive and time-consuming tasks such as note taking, charting, alert management, and other workflow optimization tools. Although these uses are not directly connected with the diagnostic process, they do contribute to the capacity of physicians to focus on essential aspects of diagnosis like clinical analysis and patient communication. Furthermore, certain models designed for automated diagnosis could be of considerable benefit for populations in countries with chronically deficient available personnel and poor access to quality care[54]. ML models approved by strict standards of use could help healthcare providers deliver a standard of care that complies with the AAAQ assessment and thus support the effective protection of people's right to healthcare. Naturally, the risks of harm must also be carefully assessed, as will be done in Chap. 4.

[50] Meredith MacKenzie Greenle et al., "End-of-Life Health-Care Utilization Patterns Among Chronically Ill Older Adults," *American Journal of Hospice and Palliative Medicine®* 36, no. 6 (June 2019): 507–12, https://doi.org/10.1177/1049909118824962.

[51] Ygal Plakht et al., "Healthcare Resources Utilization throughout the Last Year of Life after Acute Myocardial Infarction," *Journal of Clinical Medicine* 12, no. 8 (April 8, 2023): 2773, https://doi.org/10.3390/jcm12082773.

[52] World Health Organization, "Ageing and Health," *Newsroom* (blog), accessed March 13, 2024, https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[53] These challenges will be analyzed at length in Chap. 4.

[54] Carles Rubio Maturana et al., "Advances and Challenges in Automated Malaria Diagnosis Using Digital Microscopy Imaging with Artificial Intelligence Tools: A Review," *Frontiers in Microbiology* 13 (November 15, 2022): 1006659, https://doi.org/10.3389/fmicb.2022.1006659.

3.3.2 The Right to an Explanation

Although not explicitly established, the interest in a right to an explanation regarding digital technologies was sparked by the introduction of several Articles and Recitals[55] included in the General Data Protection Regulation (GDPR) that was enshrined into the European legislation in 2018[56]. The objective of these provisions is to safeguard the rights of natural persons regarding the collection and processing of personal data[57]. Perhaps the two most relevant Articles in the GDPR are Articles 12 and 22. Firstly, Article 12(1) establishes the right of the data subject, understood as an identifiable natural person[58], to transparent information about the processing of data in a "concise, transparent, intelligible and easily accessible form, using clear and plain language". Secondly, Article 22(1) speaks of the right not to be subjected to a decision "based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[59].

The debate over whether a right to explanation is established under the GDPR was then fueled by several academic analyzes published after the draft went public in 2016. While the debate will not be explained in depth here, it highlights several difficulties with the formulation of a right to explanation conceptually

[55] Articles 12, 13, 14 and Recitals 60–62, Article 15 and Recital 63, and Article 22 and Recital 71.

[56] Consolidated text: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).

[57] Article 4 (2) defines 'data processing' as: "(…) any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".

[58] General Data Protection Regulation (GDPR). "Art. 12 GDPR—Transparent Information, Communication and Modalities for the Exercise of the Rights of the Data Subject." Accessed March 13, 2024. https://gdpr-info.eu/art-12-gdpr/.

[59] Recital 71 defines profiling as "any form of automated processing of personal data which evaluates the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning his performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or similarly significantly affects him". It is interesting to note that it specifies that the subject should not be subjected to profiling if a legal effect or other type of harm occurs. This vague contextualization of harm makes it difficult to enforce this article, especially in relation to AI harms.

and in its practical implications[60]. First of all, there is the problem of technical unfeasibility. Many of the state-of-the-art ML models are opaque and are not easily scrutinized. Second, a right to explanation understood as disclosure of proprietary code or data can infringe the right to intellectual property and trade secrets, also observed in European law. Third, it has been criticized that a right to explanation is likely to stifle innovation as companies will be less incentivized to use ML technology, which could have an important effect on the European economy[61]. Finally, at a conceptual level, the right to explanation has only been vaguely delineated, and as a result, a problem of conceptual conflation has arisen with the emergence of similar concepts in corporate guidelines and ethical frameworks such as explainability, interpretability, transparency, accountability, among others.

The content of such a right, however, is not meaningless. When we are dealing with sophisticated predictive models that use citizens' private data to make decisions, often with little to no human oversight, in critical areas such as profiling, credit approval, hiring decisions, and health care, it seems reasonable to have a rights framework that allows the right holder to be informed about how their data is being used and to opt out of purely automated decisions. A further implication is the possibility to challenge the accuracy of that decision. Gryz and Rojszczak explain that a lack of understanding of what criteria led to the decision means that such an action cannot be exercised. They argue further that: "failure to provide a procedure to challenge the decision, including legal action, would deprive individuals of a key fundamental right—the right to a fair trial."[62]. Furthermore, as it will be developed later, explanations have a crucial role in diagnostic and treatment decisions. The incorporation of ML models in the decision-making process, especially of models with high opacity, require that we consider their impact on informed consent and patient autonomy, and whether a right to explanation may be a solution to address these challenges or if there is a better alternative to protect these particular interests.

---

[60] A full account of the debate can be found in: Bryan Casey, Ashkon Farhangi, and Roland Vogl, "Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise," *Berkeley Technology Law Journal* 31, no. 1 (2019): 1–143, https://doi.org/10.15779/Z38M32N986.

[61] Andy Crabtree, Lachlan Urquhart, and Jiahong Chen, "Right to an Explanation Considered Harmful," *SSRN Electronic Journal*, 2019, https://doi.org/10.2139/ssrn.3384790.

[62] Jarek Gryz and Marcin Rojszczak, "Black Box Algorithms and the Rights of Individuals: No Easy Solution to the 'Explainability' Problem," *Internet Policy Review* 10, no. 2 (June 30, 2021): 5, https://doi.org/DOI:10.14763/2021.2.1564.

The apparent importance of explanations stems from the interest of governments, civil society, academia, and other institutions to design regulatory frameworks and ethical guidelines to safeguard these interests. As shown in Chap. 2, the emergence of models able to run powerful ML algorithms propelled an array of literature focused on trying to determine the principles that ought to be the driving force behind the design and deployment of these systems. In this process various concepts were used to characterize the notion of "desirable" AI, such as explainability, interpretability, accountability, transparency, and trustworthiness, among others. However, the matter of conceptualizing these terms and determining their implications has not been a collaborative effort for the most part[63]. Instead, it has been mostly left to companies and individual governments to decide how to define key aspects of AI governance and has led, among other issues, to the pitfall of conceptual conflation. This is problematic because it confuses what an explanation should amount to and what legitimate interests of persons this right is meant to protect.

The matter of conceptual conflation as it is considered here has two problems relevant to the subject at hand: it makes interdisciplinary and converge research[64] difficult and time-consuming, and it can make policy making inaccurate, vague, and maybe even unenforceable. Conceptual conflation means the merging of two or more concepts into one, leading to epistemic error. For example, in the subfield of ethical and responsible AI, the concepts of explainability and interpretability are often used interchangeably even though an explanation and an interpretation are two distinct concepts and have distinct implications. For instance, Lipton shows that many authors have characterized interpretability as a means to engender trust in the user of ML models, but it is sometimes understood as a synonym of transparency as well.[65]. The problem observed here is that in the rush to make these guidelines and frameworks, these conceptualizations made by different actors forget the nuance and complexity of what they are attempting to grasp. In this confusion, relevant technical and normative questions and requirements

[63] This has changed recently with the passing of the AI Act, a comprehensive body of regulation to address the risks of powerful AI models in the EU. However, this remains an effort that still clashes with other attempts at regulation, for example, with the AI bill of rights in the US.

[64] *Convergence: Facilitating Transdisciplinary Integration of Life Sciences, Physical Sciences, Engineering, and Beyond* (Washington, D.C.: National Academies Press, 2014), https://doi.org/10.17226/18722.

[65] Zachary C. Lipton, "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery.," *Queue* 16, no. 3 (June 2018): 31–57, https://doi.org/10.1145/3236386.3241340.

might be misunderstood or misinterpreted, and the techniques or tools that end up being developed to address them are not appropriate or useful.

Another potentially harmful consequence of conceptual conflation regarding explanations has to do with the concept of accountability. In general terms, to be accountable is to have a moral and/or legal requirement to provide a justification or reasoning for the intentions and beliefs that guided an action or omission in a decision-making process to entitled actors; for instance, a physician is accountable for the decisions made regarding a particular treatment path to the patient or proxies[66]. However, research efforts are being directed towards making the models accountable or building accountability directly into the models. The notion of accountability requires the recognition that the person held accountable is a moral interlocutor, that is, someone who is receptive and responsive to moral reasons for or against the matter at hand. The focus of an analysis of accountability is the moral agent's ability to justify his actions and his disposition to regulate his own behavior[67], not to merely provide an explanation of how a specific output occurred. As such, since no ML model, despite how human-like it might appear, can meaningfully justify an output or prediction or to regulate its behavior voluntarily, it follows that a ML model cannot be held accountable for a determinate outcome but the people who develop it and apply it are.

To dispel the issues brought by conceptual conflation and move forward to understand the implications of explanations in this field, it is necessary to clarify what is understood here by an explanation of an ML model in medical diagnosis. From a broader scope, there are two ways to understand the meaning of an explanation at this intersection: First, the technical explanations derived from the work of computer science, and general explanations that link to causes that can be comprehended by people without a high-level of mathematical expertise[68]. The kind of explanation that is appropriate and necessary in a medical diagnosis setting will vary depending on factors like what is the aim of the explanation,

[66] Helen Smith, "Clinical AI: Opacity, Accountability, Responsibility and Liability," *AI & SOCIETY* 36, no. 2 (June 1, 2021): 535–45, https://doi.org/10.1007/s00146-020-01019-6.

[67] Marina Oshana, "Moral Accountability," *Philosophical Topics* 32, no. 1/2 (2004): 255–74.

[68] There is a substantial body of literature dedicated to this matter and even a subfield of technical AI research called Explainable AI (XAI) that emerged from the necessity of compliance to these standards in order to advance in the implementation of models in critical settings. This subfield has two main branches: a technical one that focuses on developing methods and techniques to produce more explainable models from a technical perspective, and a socio-technical one, that deals with the role of explanations for achieving standards of trustworthiness, accountability, and transparency.

who is the receptor of the explanation, and other aspects like the timing. I propose a distinction between two types of explanations, according to the role they serve: functional explanations and decision-oriented explanations. This distinction is relevant because it establishes specific possibilities, conditions, and aims of the explanations, and will help with the other concepts mentioned above.

Functional explanations are those associated with the functioning of a model. They can revolve around the data used to train the model, the type of algorithm, or the model understood as a system. Depending on these factors, the developers can provide different explanations, or no explanation at all. For example, in ML systems using decision tree algorithms, a technical explanation is usually simple to provide. A decision is the result of the answer to a question A, where the two options are B or C. The decision tree follows a form of conditional logic: IF A = TRUE, THEN C. ELSE B. In a different example, the output of a model using a DL algorithm is difficult to explain because the analysis relies on finding complex patterns and interconnected parameters in the data that makes the model opaque and difficult to explain in human terms. Since ML models are fundamentally non-static[69], meaningful functional explanations must address this characteristic. As such, functional explanations can be *ex ante* or *ex post*[70]. A functional explanation *ex ante* occurs at the point where the decisions regarding the model are made, i.e., which database was used, how it was curated, what algorithm was selected, what hyperparameters were chosen. On the other hand, a functional explanations *ex post* occurs after the model has been trained and validated. It should be noted that in both cases, functional explanations might not be possible at all, or they might be imperfect depending on the particular technical limitations of the specific model.

The second type of explanations are decision-oriented explanations. They are *ex post,* and such are possible only once the model has produced an output, i.e., reached a "decision". While functional explanations are of a general nature as they concern a ML model as a whole, decision-oriented explanations only concern a specific output. For instance, decision-oriented explanation would be the results of a mammogram showing suspicious lesions after being analyzed by a ML model. The radiologist in charge of the test would be the one obtaining the results and in charge of translating the technical explanation into one that

[69] Adding new data to ML models will change the predictions and even the parameters they had on previous "rounds".

[70] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 2 (May 2017): 76–99, https://doi.org/10.1093/idpl/ipx005.

the patient can understand. Then the clinician in charge of the diagnosis process would be the one responsible for communicating and further explaining the implications of the results and the steps forward. The aim of the explanation here is to help the patient understand the situation regarding her health and to enable her to make an informed decision.

Although it is true that decision-oriented explanations have a technical component derived from the information provided by the model, the explanation given to the patient must be contextualized, and ultimately, it is those explanations that matter. Thus, as far as the decision-making process of diagnosing a patient is concerned, decision-oriented explanations are relevant. This argument must be based on some basic assumptions. First, speaking concretely about the clinical encounter, we can assert that patients want to have the best possible clinical outcome, i.e., they want to regain their physical or mental health, or the best degree of physical and mental capabilities attainable. Normally, no one acting on their free will and in possession of a normal psychological state goes to the praxis or hospital, hoping for an adverse clinical outcome. Second, for the diagnostic and treatment processes to work and, ultimately, give the patient the best possible clinical outcome, the patient must provide the clinicians in charge of her care, i.e., physicians and nurses, with the relevant information about their health and must commit to adhering to the treatment plan.

Medical information can belong to the uttermost private sphere of a patient; for instance, it might relate to personal hygiene, sexual behavior, traumatic events, and other aspects that the patient considers as part of her intimacy, as such disclosing this information is not a trivial matter. Third, adhering to a treatment plan might require the patient to make minor or significant changes in her lifestyle, might cause the patient to experience mild or severe discomfort, or even acute pain -as it is the case with most chemotherapy options available-, or might require the patient to use financial resources that might impact on her general quality of life. In both cases, i.e., disclosing medical information and adhering to a treatment plan requires that the patient places her trust in that the clinicians in charge are medically competent, knowledgeable, professional and that they make decisions with the best interest of the patient in mind. Fourth, we can also assert that in the current manner of practicing medicine, the decision-making process is no longer solely a task of the physician but instead a joint effort where the autonomy of the patient is considered of high importance, as it enables the provision of informed consent for clinical treatments.

It has been established that explanations are normatively and empirically relevant *if* they are meaningful to the patient regarding the diagnostic process he is going through. He has a claim to the information necessary to make an informed

decision about this process, and technical explanations of the functioning of the model are not necessary for this. Thus, I argue patients do not have a right to such explanations and that they only have a claim to the content of decision-oriented explanations that are suitable for the diagnostic procedure and to provide informed consent.

Functional explanations are strictly technical and can contribute to the general goal of a model being explainable, but they require a level of technical expertise that most patients do not possess. Normally, patients are not computer science experts in an epistemic position to understand what this explanation would offer, but even if they were, the technical functioning of a ML model is not directly relevant to the diagnostic process where the patient is involved[71]. The goal of an explanation, as has been argued throughout this section, is to enhance the agency of the patient and enable them to make an informed decision regarding his care. This constitutes a first justification for decision-oriented explanations: information needs to be meaningful to the patient and a functional explanation falls outside this scope. This applies to models that make non-automated as well as automated diagnostic decisions, although in the latter case, the patient should have a right to *know* that the model is fully automated.

This argument remains a standard of practice, even in situations where no AI is concerned. Patients in most scenarios do not get exhaustive medical explanations, as they are not, in most cases, meaningful for the patient to make an informed decision. A physician could, in theory, explain how the formation of polyps in the colon occurs in a textbook manner, but having this information does not result in the patient being in a better position to decide about testing or posterior treatment options. Explained with a different example, one does not need to know how exactly the engine of a car works to be prepared to make the decision of slowing down in a curve or to veer to one side in order to not hit a passing animal. A patient does not have a right to have a full explanation of how the model produces a certain diagnostic output (a full explanation understood as both the functional *ex ante, ex post,* and the decision-oriented explanation). However, he does have a claim to the information necessary to being able to make an informed decision.

[71] A functional explanation will probably be relevant or even necessary in other scenarios. For instance, when seeking regulatory approval it might be required to provide the model's parameters, the results of the validation or testing phases, and other relevant metrics. Another scenario would be when training the clinical technologists or clinicians in charge of operating or directly using the model, as they would require to be fully informed of the working of the model, and they would be in an epistemic position where the explanation would be meaningful.

A second justification points to the potential unjustified burden a right to an explanation would impose on the clinicians and developers. Since explanations are an important aspect for patients to give their informed consent to diagnostic procedures and treatment decisions, theoretically they should receive an explanation of the symptoms, testing options available, the prognoses and the treatment's effectivity rate. However, physicians are usually in charge of dozens of patients that have an equal right to receive timely and care of quality. The time and capacity physicians have to see every patient is limited, and physicians' needs outside of the medical practice are also relevant for consideration. A right to a full decision-oriented explanation would mean that the physician has the duty to find the time and resources to make complicated technical terminology comprehensible to the patient. This would entail that the physician must dedicate significantly more time to each patient, potentially leading to a decrease in his capacity of attending to the same number of patients or, on the other hand, to the increase of workload that could potentially lead to burnout. In a world of healthcare systems operating at capacity, of overworked medical personnel and increasing medical needs, a right to explanation would infringe the rights of many others. Moreover, it could also potentially infringe the right to privacy or the right to intellectual property of the developing company, as in most cases the models are protected under privacy and copyright laws. As such, the positive duties derived from a right to an explanation could not be justified.

Given these arguments, I hold that a right to an explanation is unnecessary, and it cannot be justified, at least as an explanation in healthcare. The conceptualization, as it stands, is imprecise and risks being overblown. Most importantly, the content of this right in medical diagnosis already exists, at least in part, in another established right, namely, the right to informed consent. Therefore, I argue that instead of trying to create a new right, it would be more feasible to update or expand its content according to the new needs arising from the progress of new technologies. The right to informed consent already encompasses the importance of information for decision-making and has been pivotal to strengthen the notions of patient autonomy and enhanced agency. Patients have a claim to receive the information necessary to make an informed decision regarding their diagnosis and posterior treatment. Whether the information regarding the implementation of a ML model is relevant to this purpose would have to be determined according to the level of diagnostic autonomy of the model, i.e., how much of the diagnostic decision-making is left to the model, the complexity of the model, i.e., whether the model can be explained fully, partially, or not at all, the epistemic position and disposition of the patient and the resources available. It would be of equal

importance to make sure that the clinicians and other personnel are not overburdened by introducing a right to explanation or a strengthened right to informed consent. The constraints of property rights, data privacy of other patients in the model dataset and trade secret of companies also must be considered.

An updated and enhanced version of the right to informed consent would integrate the interests of patients in obtaining information about their diagnostic process where AI is concerned, to have access to tools and materials to become better informed about the benefits and risks of the implementation of models in the clinical setting where they are concerned and to object to fully automated decisions where legal and medical implications are at play, as delineated by the GDPR.

3.3.3 The Right to a Human Decision

The idea of a right to a human decision has also emerged as a response to the widespread adoption of digital technologies, including AI, and the ubiquitous presence of automation and its risks. The legislative hints to a right to a human decision appeared already in the Data Protection Directive of 1995[72] that conceded the importance of not being subjected to automated decisions. Furthermore, as seen before, the article 22(1) of the GDPR speaks of a right against "being subjected to a decision based solely in automated processing (and profiling)"[73]. Within the regulatory space regarding particularly decisions in healthcare and AI systems, the upcoming European Health Data Space (EHDS) regulation addresses this issue and gives patients a measure of control over how healthcare providers can use their digital health data, but it does not create a right against automated decisions in healthcare[74].

The first matter to clarify is what the concerns behind the idea of a right to a human decision are in general. Until relatively recently, decisions with an impact on people's rights were made only by other persons and the consequences

[72] European Parliament, "Directive 95/46/EC | European Data Protection Supervisor," September 20, 2024, https://www.edps.europa.eu/data-protection/our-work/publications/legislation/directive-9546ec_en.

[73] "Art. 21 GDPR—Right to Object," *General Data Protection Regulation (GDPR)* (blog), accessed September 23, 2024, https://gdpr-info.eu/art-21-gdpr/.

[74] European Commission and Directorate-General for Health and Food Safety, "Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space," Pub. L. No. 2022/0140/COD, 13.20.60.00, 15.30.00.00 (2022), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197.

of a mistake could be allocated to the person who had made the decision. Furthermore, there are established channels and procedures to appeal a decision, to request redress or to present legal charges against who made it. In the wake of AI technologies and their capacity to automate many processes at different levels of normative significance and in different areas of application, it is a concern that the decision-making authority (and hence the responsibility and accountability allocation) will shift from human beings to algorithmic models.

The concrete concerns can be identified according to three broad aspects. First, automated decisions made by AI systems could only consider the information that was included in the datasets used to train the systems. This is the general concern of algorithmic justice, which argues that it is highly problematic to assume that AI systems are suitable to take over decision-making processes because they are considered to be more objective or less prone to make unfair decisions than humans, since it is a computer program based on logic. This is not the case, as it will be further explained in Sect. 4.2.1, because of the biases that can be introduced on the datasets and that can generate discriminatory decisions. The second concern and closely connected with the matter of explanations is that for some types of state-of-the-art models, the decisions, i.e., outputs, cannot be explained satisfactorily with a decision-oriented explanation. This is problematic if an automated decision infringes or violates a person's right and there is no way to understand the reasoning behind the decision.

The third concern is that there is still a lack of clarity about how fully automated AI systems work, and regulatory frameworks are only now beginning to be adopted in many countries. This means that there are still no clear channels for redress and contestability, i.e. the ability to challenge a decision in a meaningful way so that the decision can be reversed or changed. Another concern is how these systems can and will be used. One of the major areas of debate is the use of technologies for surveillance purposes. Aside from the potential invasion or violation of citizens' right to privacy, these kinds of technologies could be used in ways that profile or discriminate against people. In general, these concerns encapsulate legitimate interests that citizens have, such as access to a fair assessment of the data that leads to fair decisions, and the ability to participate meaningfully in important decisions about themselves[75].

To address these concerns, one of the strategies that has gained more attention is the introduction of human intervention both during the technical development,

[75] Yuval Shany, "The Case for a New Right to a Human Decision Under International Human Rights Law," SSRN Scholarly Paper (Rochester, NY, October 4, 2023), https://doi.org/10.2139/ssrn.4592244.

implementation, and operational phases. The strategies are known as "Human-in-the-loop" (HITL)[76]. In the context of ML models in diagnosis, this means that an AI developer is present at different points of the developing process, and clinicians when the model is used. At the technical level, this approach integrates human judgment into machine learning processes, combining algorithmic automation with expert oversight. This interaction typically occurs in tasks like data labeling, model refinement, and decision-making, ensuring that the system adapts dynamically to human feedback. At the clinical level, clinicians contribute through data labeling, model validation, and decision-making, with the aim to ensure the decisions reflect clinical nuances. This collaboration allows AI to manage large data efficiently while relying on human oversight for complex or ambiguous cases, balancing machine efficiency with professional judgment.

However, this approach faces a series of issues. For instance, incorporating human oversight can slow down processes, especially when dealing with large amounts of data, making it less practical for real-time applications. Another issue is that the quality of the collaboration relies heavily on the expertise of humans involved, which might introduce bias in the process. In addition, even if there is a human being in charge of making these decisions, this strategy alone does not ensure that companies will act in a way that the rights of citizens are protected. As existing regulatory frameworks are still being adopted, there is a need to ensure, first, that companies are aware of these regulations; second, that compliance pathways are properly designed and do not put companies at risk of being rendered obsolete or bankrupt (at least not without time to adapt to the regulatory requirements); and third, that actors who may seek to exploit loopholes or gray areas in the regulations are prevented from doing so.

Another problem has to do with whether human intervention in the decision-making process is meaningful at all. In other words, even if a person is involved in the process, how much actual authority does that person have over the final decision? In other words, what is the distribution of decision-making control between the human and the automated system? For example, if the role of the HITL is merely to observe or annotate data, but the decision is made entirely by the system and the human cannot stop it even if there is a good reason to do so, then using this strategy to address concerns about unfair automated decisions is not useful. Similarly, it is important to consider how the human-model interaction occurs in practice and what the consequences for the agent's ability to make

[76] Eduardo Mosqueira-Rey et al., "Human-in-the-Loop Machine Learning: A State of the Art," *Artificial Intelligence Review* 56, no. 4 (April 1, 2023): 3005–54, https://doi.org/10.1007/s10462-022-10246-w.

decisions are. For example, if the interaction affects the skills of the clinician, then the role of HITL to prevent automated decisions for infringing the rights of patients could not be fulfilled as a consequence of deskilling (see Sect. 4.2.4).

However, there are good reasons to be skeptic about the need or the usefulness of a right to a human decision. First, some argue that the concerns brought by automation are already sufficiently addressed by existing legislation bodies and barring the usefulness of cohesion, a need of a new right is lacking[77]. Another criticism argues that although some state-of-the-art models are opaque, this is not an inherent trait. Huq argues that models that do not provide explanations are made to be that way. Therefore, this lack of explainability is a design choice and should not be seen taken as an inevitability resulting from the complexity of the technology[78, 79]. He also argues that, strictly speaking, there are no "fully automated models" since human intervention is always required at different points in the development and deployment of a ML model[80]. He agrees that while we should not be leaving decisions with an ethical or normative element to automated models, what we should aim for instead is a right to a well-calibrated machine decision instead of a right to a human decision[81].

The second general argument against such a right is that the opacity of a model is not a sufficient reason to demand a right to human decision, because human decision making can also be considered opaque. Because it is sometimes impossible to discern patterns of logical construction or adequate justification in human decisions, societies have learned to develop ways of dealing with situations of accountability through legal means and specialized methods, such as evaluating human testimony, regardless of the lack of adequate explainability.

---

[77] Elena Abrusci and Richard Mackenzie-Gray Scott, "The Questionable Necessity of a New Human Right against Being Subject to Automated Decision-Making," *International Journal of Law and Information Technology* 31, no. 2 (August 30, 2023): 132, https://doi.org/10.1093/ijlit/eaad013.

[78] Aziz Z Huq, "A Right to a Human Decision," *Virginia Law Review*, Technology, 106, no. 3 (2020): 611–88.

[79] Other bodies of literature have said the opposite many times. What Huq is saying is not that sophisticated models are made opaque on purpose, but that the design choice, i.e., the choice of which algorithm is used, is normally made based on higher performance and metric assessment. There are models that are much more scrutable and thus considered more transparent, but they are not as accurate or as efficient. A normative question that remains open is if, and when, we should prioritize one or the other.

[80] He gives the examples of model architectural decisions, choices related to which datasets and validation techniques are used in the training process and lastly that humans are in charge of moderation, maintenance and technical supervision of the model once deployed.

[81] Huq, "A Right to a Human Decision," 686.

This argument aims to justify that humans and machines should be held to the similar levels of moral and legal rigor in decision-making processes.

Although this might sound reasonable, given that opacity is a common characteristic, there are stronger reasons to dismiss this idea. In the first place, there are legal and moral instruments to deal with matters of justification regarding decision-making in administrative, electoral, political, medical, and civil settings. Agents can and have been held accountable for the consequences of erroneous decisions. It is true that legal systems and normative structures may have flaws, but it is usually not acceptable that, for instance, a judge deciding on the imprisonment of a felon without providing ample and sufficient justification and documentation for the decision. In the case of sophisticated ML models, this is simply not possible, and it can lead to highly problematic consequences. For example, in 2008, the state of Arkansas implemented an automated model to allocate provision of care and due to a bug, i.e., an error in the code, the system reduced the hours of care workers allotted to patients with cerebral palsy[82]. The decision was upheld for 10 years, and it was only after several legal battles that the decision was overturned and deemed a mistake. However, the system continues in operation and similar cases have ever since continued to occur. If the question is why concerns about automated systems seem to be at a different normative level, this example helps to illustrate why. Humans and ML models are epistemologically different. First, all ML models operate at scale, which means that a machine with a biased training set that makes biased predictions can affect multiple people in different settings. Second, models can be used to centralize power and make it difficult to assign responsibility and accountability. Third, ML models can also create complex and opaque feedback loops in which biased data continually produces more damagingly biased results, which are then fed into a new model, repeating the process.

Despite the reasons for rejecting the claim of a right to a human decision, the concerns raised at the beginning of this subsection are still relevant and thus merit further consideration. From a normative perspective, I argue that certain decisions, including those that have a direct bearing on a patient's clinical outcome during the diagnostic process, should be left to a human person because of the consequences for people's rights if an error is made, but more importantly, because in such scenarios only a human person can truly grasp all the elements of normative weight that the decisions require to be evaluated. The appeal to a

[82] Erin McCormick, "What Happened When a 'Wildly Irrational' Algorithm Made Crucial Healthcare Decisions," *The Guardian*, July 2, 2021, sec. US news, https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions.

right to a human decision should apply, but only in those scenarios of normative weight. In this sense, we would speak of a right with its universal quality, but only if certain circumstances are met. For instance, when the medical context of the disease supposes a significant risk for the patient, and when the decision is multifactorial and therefore, a human is better equipped to comprehend it than an algorithmic model.

In other contexts, however, patients would not generally have a claim to reject an automated decision or to demand that a physician make it. There are several reasons to justify this argument. First, it has been established that people have a right to the best minimum attainable standard of health, which implies a potential need for the medical infrastructure and resources of the health system. Medical personnel can be counted as part of a system's medical resources. In a scenario where ML models are successfully implemented and become the standard of care for some types of diagnostic procedures (which means they must meet strict standards of safety, performance, and accountability is well managed), an individual patient who wants a human decision instead of an automated decision from a model may not have a right to it. While this is not a debate about whether ML can or should replace physicians, the context in which this right to a human decision is to be implemented must recognize the realistic limitations and strains of a real health care system. This means, among other things, recurrent staff shortages, waiting times for specialist appointments, occasional unavailability of drugs, and limited hospital capacity, especially in ICUs and for intensive treatments. Accepting a generalized right to a human decision, under the four-pillar approach, entails that every patient would *have to* have a right to a human decision. However, increasing staff shortages and the demographic shift to a more aging population threaten the actual provision of care and will exacerbate current trends, resulting in poorer quality and efficiency, which would make unfeasible to protect such a right effectively.

Second, according to the analysis of the right to health care in sect. 3.3.1, the right to timely access to quality health care to the highest level attainable is undoubtedly among the conditions necessary to lead a life. This conflicts with the justification of a right to a human decision, if by accepting this right we curtail the chance of other people to even have access to healthcare because of the limited availability of medical personnel, and as such, it could not be accepted unless we speak of those scenarios where the human decision is the only one that can secure the rights of the patient.

Third, the pillar approach states that every person has a fundamental right to physical and psychological integrity. This applies to medical professionals. Considering the significant burnout rates (which will be explored in Chap. 4), it could

be argued that it is justifiable from a normative perspective to use ML models that make automated diagnostic decision for certain diseases if they, first, have reached the established standard of accepted medical performance, safety, and compliance with existing normative regulatory frameworks; and second, their implementation support medical professionals to decrease their workload and burnout ratios.

Of course, this initial assessment should be further considered against other normative aspects, like the condition or disease to diagnose, the accuracy of the model for this specific disease or condition, additional costs, level of user confidence, the type of recommendation patients would receive, among others. Furthermore, other ethical problems that arise in critical scenarios should be critically assessed as well. For instance, if the model provides an erroneous result, if it stops working and patients have no access, if it is hacked and the patient's data is compromised, etc.

3.3.4 The Right to Privacy

As with the right to explanation and the right to a human decision, the right to privacy in the context of AI in medical diagnosis is complex because it overlaps with many of the normative elements and scenarios already discussed. As will be shown, it is not possible to delineate the argumentation in isolation, and there will be some return to points already made that also play a role here. Nevertheless, this is another reason to reinforce the points made about the need for interdisciplinary work.

Although digital privacy as an object of public discourse is relatively new, we can think of an initial notion of privacy as the space we define as separate from the public sphere, where we keep some aspects of ourselves and our lives away from others who do not belong to that private sphere, and where a different set of rules of behavior may also emerge. The existence of a private sphere is important because it protects relevant moral interests, such as freedom of thought, expression, religion, and notions of self-determination and autonomy. This private sphere includes information about ourselves that we have an interest in being able to control and protect from unwarranted intrusion. This is the basis of a general right to privacy, and it has important normative implications, such as that states should not collect information about citizens without their consent and not without good reason, as has happened in dictatorial regimes where mass surveillance has been used as a means of controlling the population. It also requires that companies do not collect information about others that they wish to keep private without the explicit consent of those individuals. For example, an

employer should not place listening devices under employees' desks to find out what they think of their bosses or whether their political views align with those of an investor. This would normally be seen as a clear violation of employees' right to privacy, as well as a threat to their right to freedom of thought and political expression.

Despite this, however, there are contexts in which surveillance methods are nevertheless used under the justification of protecting other rights or interests, like safety. In some of these cases, AI systems have come to play an important role, as their technical capabilities are seen to be useful for surveillance purposes. For example, in a recent occurrence, schools in different states across the US began to use AI-driven tracking software to monitor the digital data associated or generated by students in students' school-issued devices such as tablets and laptops, and even in student email services. The data monitored included browser searches, texts sent through the school's devices and geolocation. The justification for conducting this monitoring is dual. First, it is meant to identify and block content deemed prohibited like pornography, violence, gambling, etc., and second, to help manage safety threats like school-shootings and other concerning behavior, like drug abuse, bullying and suicide[83]. The tradeoff between the privacy of students and their safety seemed reasonable to many school directors and boards.

However, there are problematic aspects to these measures. First, the companies do not provide independent evaluations of the results of the surveillance systems due to trade secret protections, which raises questions about whether incidents related to the aforementioned threats have actually been prevented. Second, in most cases, students were not informed about the surveillance software; instead, parents were asked to sign blanket consent forms for the use of technology in the classroom, which did not make explicit the scope and purpose of the technology being used. The rationale is that students might change their behavior if they knew they were being monitored, resulting in less accurate predictions.

It could be accepted that since parents and school officials are in fact responsible for the well-being of students, they have the authority to impose these measures on them if certain conditions are fulfilled. First, that the threat that they are attempting to prevent supposes a significantly serious risk to the lives and health of students, which would have priority over the right of privacy. And second, that there is evidence that the likelihood that the risks materialized in harm

[83] Danielle Keats Citron, "The Surveilled Student," SSRN Scholarly Paper (Rochester, NY, August 25, 2023), https://papers.ssrn.com/abstract=4552267, 4–5.

is also significantly high, which is the case of school shootings in the US[84]. However, the infringement of the general student population's right to privacy could become a violation for students who are nearing or have reached the age of majority. From a normative point of view, there are quite a few aspects that would have to be weighed to make a decision about using AI for surveillance. One, since there is not enough factual information about the benefits of these technologies, it is not clear whether their use can be justified. Two, surveilling legal adults might have other implications that should be considered; and three, there is a known danger that these technologies are used to the detriment of students belonging to minorities like Black, Latino and queer, thus exacerbating existing systemic inequalities and discrimination[85].

Taking a step aside from the discussion about AI and other types of digital technologies, the right to privacy is widely considered to be a fundamental right, and it is often invoked in political and social debate. In 1890, in an article in the Harvard Law Review, it was conceived as the right "to be let alone"[86]. The authors argued that no comprehensive legal protection of privacy had yet been established, and that existing legal protections for intellectual property rights and against defamation did not encompass what a right to privacy entailed. The interest of the authors in seeking a means of protecting the right to privacy was based on the increasing phenomenon of newspaper articles which, in their opinion, intruded into the personal matters of people and published them without justification as these matters being of any public or general interest.

Ever since the publication of this article, the subject of a right to privacy has been often discussed and is generally accepted that there are legitimate interests that ground the existence of such a right, although this is not explicitly enshrined in the UDHR. However, the existence of a right to privacy *has* been contested by several authors, including Judith Thomson and Ingmar Persson & Julian Savulescu. Thomson argues that the right to privacy is actually a cluster of rights that intersects with the rights to owning property and what she calls "rights over the person". She explains that the contents that make up the right to privacy are already secured by the protections belonging to those other of rights, and as

[84] Bellal Joseph et al., "Defining the Problem: 53 Years of Firearm Violence Afflicting America's Schools," *Journal of the American College of Surgeons* 238, no. 4 (April 2024): 671, https://doi.org/10.1097/XCS.0000000000000955.

[85] Claire Galligan et al., "Cameras in the Classroom: Facial Recognition Technology in Schools," Technology Assessment Project (Michigan: Gerald R. Ford School of Public Policy. University of Michigan, 2020), http://deepblue.lib.umich.edu/handle/2027.42/191755.

[86] Samuel D. Warren and Louis D. Brandeis, "The Right to Privacy," *Harvard Law Review* 4, no. 5 (1890): 193–220, https://doi.org/10.2307/1321160.

such, there is no need for a distinct right to privacy[87]. She makes the example of a person owning a photograph that he hides in a wall-safe to prevent others from seeing it. If a neighbor, she theorizes, builds an X-ray machine to look at the picture through the walls of the safe, he is violating the right of property in that the owner does not want that his property is looked at[88]. Someone could argue that the right to privacy is also violated, since the machine acquired information of the photograph without the owner's consent, but she argues that a right to privacy is unnecessary because what is wrong is a violation of the owner's right to property and rights over his person. She concludes that what is called a right to privacy is only a derivative right because there is no legitimate interest that it is not derived from other higher interest.

Similarly, Persson and Savulescu argue that a right to privacy cannot truly exist because a right of privacy implies that "outsiders do not acquire the beliefs or information about a person who she reserves for herself and a select group of others"[89]. They explain that while we can have rights against others to refrain from certain means and uses of information, we cannot have a right against the acquisition itself. The reasoning behind this argument is, in short, that the acquisition of information is the product of an internal state of an individual, and that this internal state is frequently involuntary. As a means of illustration of this argument, let us think of the times when we stood in a queue. The queue is made up of persons standing in front of another person. While standing there looking forward, one can perceive certain facts about the person in front, for instance, that the person is tall or short, has brown or blonde hair, that she is wearing a blue or brown T-shirt. This is all information about the person that we are acquiring without the express consent of the person. However, the person cannot reasonably demand that I stop perceiving these facts if they were acquired by the mere act of standing in a queue, as this is, according to Persson and Savulescu, an internal state. The point they make is that a right cannot be grounded in the fact of acquiring information only, as this is an involuntary result of having physical senses and brain functions that process what those senses perceive.

While the arguments presented by Thomson and Persson & Savulescu may have had a valid justification before the information age, I believe they are now incorrect. Moreover, I argue that a right to privacy is not only possible but also

[87] Judith Jarvis Thomson, "The Right to Privacy," *Philosophy & Public Affairs* 4, no. 4 (1975): 295–314.

[88] Thomson, "The Right to Privacy," 298–303.

[89] Ingmar Persson and Julian Savulescu, "The Impossibility of a Moral Right to Privacy," *Neuroethics* 15, no. 2 (June 28, 2022): 1–5, https://doi.org/10.1007/s12152-022-09500-3.

desirable. Privacy in the age of AI and big data has acquired a nuance that adds to the justification of such a right. The goal of advanced digital technologies and the acquisition of information about individuals is not necessarily only to have some control over the behavior of the population, as was the concern with government surveillance, but it has also become a business model. Information is data that can be sold for a variety of purposes and that can affect the rights of individuals. The acquisition of information is often done without informing the individual and in ways that make it difficult to understand.

Privacy concerns have shifted from instances such as the publication of defamatory information or private letters, and forms of visual media such as videos or photographs taken without a person's consent, to the new ways in which personal information can be obtained through the use of digital devices and the inferences that can be made about people based on that information. These occurrences resulting from the presence of new technologies are vastly different from what happens when we stand next to another person and, simply by virtue of our proximity, notice certain facts about them. In our current circumstances, there are three novel features that require us to revise the discussion of the right to privacy: First, the use of digital means to collect personal information is not only ubiquitous, but in many cases unavoidable. Second, they are deliberately placed or used to obtain the information, with or without consent, and third, personal information has become a commodity.

Take, for example, the privacy settings on smartphones. Prior to the implementation of GDPR in Europe, new phones came with a set of default privacy options that allowed software companies to access information about users without informing them. These companies were also not required to keep these settings disabled by default. In order to opt out of these forms of information collection, often disguised as features that help phone companies improve the performance of software, users had to be aware of these technical nuances (and the implications for their privacy) and have a degree of digital literacy to deactivate these settings. Of course, this was not designed with the average user in mind.

While the context introduced by AI renders Persson and Savulescu's argument inapplicable, the distinction offered in the article is still useful. The acquisition of information per se may not be a particular issue, but the fact that the information is acquired with a commercial purpose may be problematic. In this sense, the argument presented here is that it is reasonable to assume that any information about individuals acquired by private companies has a commercial purpose. In this case, we cannot separate the act from the intent. When we browse the Internet and a website places a tracker, also known as a "cookie," to identify us, there

are many uses for the information derived from this tracking. It can be used to improve the site or our experience on it by remembering language preferences or auto-fill options, for example, but it can also be used to create highly personalized advertisements, track how we interact with those advertisements, and, with enough aggregated information, help build a profile of our online behavior. All of this information, most importantly, can be sold to other companies. The business of online tracking is indeed a profitable one, and as such, our attention is usually focused on what the uses might be, and whether they are legitimate from an ethical and legal viewpoint.

The matter of privacy in healthcare has special aspects to consider. The importance of health data has warranted the establishment of special legal protections. The most famous framework for protecting health data is the Health Insurance Portability and Accountability Act (HIPAA) in the United States. Under this law, healthcare providers, and insurances are prohibited from sharing patient data with entities other than the patient and authorized proxies without the patient's consent. The aim of the law is to give a patient a degree of control and autonomy over her data and how it is used[90]. In the European Union the GDPR, although not specifically made to address health data, recognizes its value, and provides a definition for its protection[91]. Despite this acknowledgement, however, a first concern is that these regulations focus on health providers, biomedical companies, insurances, and similar defined entities that work with health data but do not provide instruction and protection for other forms of health-*related* data acquisition, such as browser searches, information from apps or devices not registered as medical, etc. In these cases, the individual is made to accept user agreements that are usually tedious and written in complicated legalese, and also important to highlight, the user is not a patient but a consumer or customer, which presupposes a different type of relationship than the one held between healthcare personnel and patients, and insurance companies and beneficiaries[92].

This can be exacerbated as commercial arrangements become more necessary for the dissemination of technologies for real world use and will place patient data

[90] 104th Congress of the United States, "Health Insurance Portability and Accountability Act," Pub. L. No. Public Law 104–191 (1996), https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.

[91] "Recital 35—Health Data," *General Data Protection Regulation (GDPR)* (blog), accessed September 23, 2024, https://gdpr-info.eu/recitals/no-35/.

[92] The normative implications of this argument will be discussed more in depth in Chap. 5.

under the control of private companies[93]. The existing accumulation of technical innovation in a few big tech companies, therefore, can create a power imbalance because public institutions may eventually become too reliant on private companies to produce the devices or software required to manage critical matters of public health. This is problematic in terms of privacy concerns because whilst public institutions with duties towards the citizens might have challenges where some interests compete against each other, these are mostly grounded in terms of prioritization for resource allocation or the protection of more urgent rights against the infringement of other less urgent rights (think of the management of the COVID-19 pandemic, that without going into an in-depth discussion about the normative justifiability of certain measures, essentially placed the rights of some vulnerable citizens against some, at the time considered so, less urgent rights of other citizens like the right of free association).

For companies, however, competing interests come instead in the form of profit and revenue against privacy. As such, it is not certain if companies, as the situation is nowadays, have enough incentives or deterrents to maintain the privacy of patient data if the expected profit is attractive enough[94] which would clash against the duties of public institutions and could place them in a difficult position if they have to balance the need to keep contracts with private companies in order to have efficient technology tools for the management of healthcare needs against the requirement to maintain the privacy of the citizens.

The normative questions that arise now are, in the first place, what are the interests we want to protect with a potential right to privacy in health and second, if there is a right to privacy, what are the implications for the use of patient data for medical purposes. I have explained the relevant role that health data plays for both medical development and nowadays, commercial use. As such, one of the primary interests considered under the right to privacy is the disclosure of identifiable patient information to third parties or for purposes other than those necessary for the diagnostic process, except with the explicit consent of the patient. This would include sharing the patient data with employers, insurance companies, and law enforcement institutions.

There are already provisions for such scenarios in the GDPR, and exceptions to this right may need to be considered for critical threats to national security, public health, and other comparable contexts. This would mean recognizing how

[93] Blake Murdoch, "Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era," *BMC Medical Ethics* 22, no. 1 (December 2021): 122, https://doi.org/10.1186/s12910-021-00687-3.

[94] Murdoch, "Privacy and Artificial Intelligence", 2–3.

this data is collected, how the data is used, and what protections we need to ensure that the use of the data does not lead to harmful consequences for patients, such as discrimination based on the sharing of health data, social stigma resulting from private information made public, and other subjective harms such as paranoia, anxiety, and embarrassment[95]. A second concern is the ability of patients to control with whom and how much information they share. This has to do with personal relationships, dignity, and the protection of certain personal freedoms.

However, a patient's right to privacy should not imply that there is no right to use this data for legitimate medical purposes, particularly when the data are used to improve the accuracy and performance of models that will provide significant diagnostic benefit to patients and potentially help to relieve the burden on healthcare systems, for example, through the use of automated diagnostic tools for specific, established diseases. This argument is similar to the justifiable use of health data to predict and manage an infectious disease outbreak. In this sense, I argue that a right to privacy can and should be upheld in the sense that patient data should not be shared with third parties for purely private commercial interests, unwarranted surveillance, or monitoring that is not for the explicit benefit of the patient. However, where the use of a model can be justified as the best standard of practice or to create necessary ground truth labels, as explained in Sect. 1.3, the obligation to use the data for the benefit of the vast majority of patients should prevail for other models. In such cases, the individual right of patients to privacy conflicts with the greater right of many patients to the highest attainable level of health care.

Going back to Thomson's argument, we can see that there is, in fact, an overlap of interests that can be seen as privacy concerns or belonging to other clusters of rights. However, the ubiquity of digital technologies designed and developed specifically with the purpose of acquiring as much information as possible from individuals, and the commodification of personal data, highly desired to, among other things, increase the performance of AI-driven models has created a context where a right to privacy is necessary. That being said, a right to privacy is not necessarily about the acquisition of information on itself. It has been argued that the mere collection of information does not have the normative force to justify the imposition of obligations. Instead, the focus here lies on three aspects: the means of equitable data collection, the justified uses of the data, and the consent

[95] W. Nicholson Price and I. Glenn Cohen, "Privacy in the Age of Medical Big Data," *Nature Medicine* 25, no. 1 (January 2019): 37–43, https://doi.org/10.1038/s41591-018-0272-7.

of data subjects to share their data with other entities for other justified uses, such as medical research or further model improvement. At the same time, a right to privacy would also require maintaining established measures, such as HIPAA, to protect patient data from being disclosed to third parties for commercial purposes or for uses that could harm patients if they were re-identified, i.e., if their personal information could be extracted from non-identifiable data (see Sect. 4.2.2).

The Matter of Risk

4

In Chap. 1, I presented and discussed the state-of-the-art ML applications in medical diagnosis and the opportunities they have to bring about beneficial outcomes to patients, clinicians and perhaps even to the healthcare systems at large. This was done from a medical and socio-technical perspective grounded in the actual (in the sense of both current *and* concrete) technical possibilities of the models and the headways made in terms of their development and deployment. Research evidence showed that some ML models in diagnosis meet and sometimes exceed benchmarks that rival those of human doctors and seemed to suggest that the models could, at the least, assist them in the diagnostic process. However, to have a balanced view of this context and therefore being able to judge if ML models are, in fact, beneficial, it is essential to also evaluate the risks that emerge from the development and, current and potential implementation of ML in diagnosis. This evaluation should also consider the risks that are already present during the diagnostic process and that have or could be exacerbated due to the particularities that ML models carry, such as the scale of data processing capacity or the lack of established methods to audit the more complex models, as well as other normative risks that have been hinted at throughout the chapters like the risk of patients erroneously diagnosed due to a discriminatory bias in the model's training data. In this chapter, I will focus on elaborating on these aspects.

However, the matter of risk is complex to address in general and more so in an interdisciplinary setting. The notion of risk, as it will be shown, can have different conceptualizations, implications, and methodological approaches to measure it. Furthermore, the promise of ML models in medical settings in general is to be able to predict the risk of disease, medical complications, and other negative

L. D. Dias Duran, *Machine Learning in Medical Diagnosis*, Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics,
https://doi.org/10.1007/978-3-662-71357-0_4

clinical outcomes. In a sense, when discussing risk in this context, we are confronted with the need to balance, on one side, the risks of using ML models and on the other, the existing risks in healthcare systems which could be reduced by using these models. Of course, it must be emphasized that this does not mean that all the issues of healthcare systems or in the practice of medicine *can* be solved or *should be* left to be solved by technologies like AI. As has been pointed out before, technological solutionism is a problematic approach that many times distracts the attention from the fact that ethical, social, and cultural challenges require far more reflection and action than a mere technical fix. An evaluation of the risks is a necessary step in order to determine whether certain applications could actually be beneficial to individuals or groups of individuals with vested interests without being prohibitive for other individuals or groups of individuals. Nevertheless, a first step when dealing with risk is to build on factual ground. An evaluation of the risks present in scenarios where there are multiple actors and potential conflicts of interests is only useful if it begins from those realistic possibilities, for if it does not, we are at the peril of falling into the extremes of the epistemic gap between excessive enthusiasm and unwarranted skepticism discussed in Chap. 2.

When dealing with risks, we must first start by identifying the relevant moral actors and the relevant normative questions. This means the person or persons to whom the risk is being imposed, and the person or persons who are imposing the risk. Some relevant normative questions are, for instance, if the risk imposition is permissible or impermissible, which criteria can be used to determine this, and if there might be contextual considerations that influence the evaluation. In the context at hand, the risk impositions created by the implementation of ML in medical diagnosis must be understood alongside the existing risks posed by the current circumstances of healthcare systems and population distribution, i.e., the shift from a bigger portion of the population in age groups of over 65 years old and the decrease of birthrates and the portion of the population in age groups under 65 years old. The underlined normative questions that arise are how risks should be addressed and prioritized, and who ought to be responsible for designing measures for this purpose and be responsible and/or be held accountable.

As will be shown throughout this chapter, the development, deployment, implementation, and adoption of ML models in diagnosis pose a set of normative challenges with high-stake consequences for patients, clinicians, and healthcare systems. Among these are the uncertainty to measure the risks of some models due to the lack of proper technical disclosure and transparency regarding the

usage of training data, the severity of the potential harms derived from potential erroneous diagnosis, the impact on the fiduciary quality in the relationship between clinicians and patients and the deterioration of skills relevant to the diagnostic procedure.

From a philosophical lens, the most pressing normative challenge of risk is the uncertainties present at different levels. It is not always possible to estimate the likelihood of the materialization of a risk, i.e., that the harm actually occurs, and also the severity of the harm. In AI, the risks are difficult to grasp and address in part due to the material characteristics of models but also because of the complexity of actors involved in the implementation and usage. In medical diagnosis, ML models can be used in various ways and each one may have its own set of risks to different involved actors. The risk of a model used for, let us say, mammography-screening offering a potentially biased output is not the same for the radiologist, who may be under pressure to identify the flaw and face a malpractice report if not, as for the patient, who may get the wrong diagnosis or treatment plan.

Normatively relevant uncertainties can be present in different areas. Steigleder distinguishes between three he applies to climate ethics but that can be of help for the subject at hand: First, regarding the normative aims, in this case, of the use of ML models in medical diagnosis. Second, regarding the effectiveness of measures to mitigate the harms, and third, regarding potential morally relevant side effects of the implementation[1]. In the first area of uncertainty, we are concerned with the questions of whether we should be implementing ML models at all, and if so, in which ways. While it could be said that the ultimate normative aim of implementing ML in this field is to provide the patient with a correct diagnosis so that there is a higher chance of a positive clinical outcome, and this is likely true for the patient involved, other actors may have different aims that have other normative implications. For healthcare institutions, the aim might be to simplify certain internal processes or to make them more efficient, while for clinicians, the aim is to be less burdened by the workload derived from the diagnostic process. From these different aims, a potential normative implication could be that healthcare institutions decide to increase the appointment availability according to the time saved by the use of the models, which would mean an increase in the workload for clinicians. The uncertainty regarding the aims, thus, could conflict and suppose risks for some moral actors involved.

[1] Steigleder, "Climate Risks, Climate Economics, and the Foundations of Rights-Based Risk Ethics," 252–253.

The second area of uncertainty deals with which efforts exist to mitigate the harms of the implementation of ML and how likely they are to be effective. At present, there are regulatory frameworks in the European Union with the recently approved AI Act and in the US with the AI Bill of Rights. However, until now, there has been no collaborative work of governments to establish international standards to ensure the protection of patients' rights and the interests of healthcare systems in terms of public health concerns. This includes the protection of patient data and guardrails to prevent malicious attacks on models that could produce erroneous outputs. There is also an urgent need for standards to ensure that models are not trained in ways that are discriminatory to patients or that exacerbate existing inequalities. The document of the AI act has been criticized for being incomplete and immature for the dual challenge of making sure the legitimate interests of citizens of the European Union regarding human rights are safeguarded while making the compliance demands comprehensible and not prohibitive for companies to adopt. Regulations that place an excessive burden on companies could lead to making the economic environment unappealing or even hostile for companies to come or remain in the European Union. Normative implications could be the stifling of innovation, fewer opportunities in the job market for workers and the decrease in the competitive capacity of the European Union in comparison to other economies like the US, China, or India with all the geopolitical and economic implications that follow.

Finally, the area of uncertainty concerning the morally relevant side effects. In the case of ML models this can go in both directions, namely, the morally relevant side effects of the adoption *or* the prohibition of these models in medical diagnosis. On the one hand, there are questions regarding the overall cost of the models if implemented that include, for example, the price of the license, the cost of hardware, the potential cost of technical operators or maintenance services, costs associated with the data used. Additionally, other financial uncertainties remain regarding established standards for reimbursement and payouts for insurance companies. Other normative side effects include those concerning the relationship between clinicians and patients, the impact on the moral and clinical skills of clinicians, the displacement or decrease of certain occupations in healthcare, and exacerbation of existing burnout and work overload[2]. On the other hand,

[2] This is particularly complex with the use of Large Language Models (LLMs) like GPT-4 and MedPalm for diagnostics. In many cases, the models have been known to produce false information and even produce medical literature citations that do not exist. So, while clinicians in several studies have been surprised by the sophistication of these models, they express concerns about the lack of reliability. See: Sophie Stoneham et al., "ChatGPT versus Clinician: Challenging the Diagnostic Capabilities of Artificial Intelligence in Dermatology,"

we have the uncertainty regarding the side effects of a prohibition or overly strict regulation of models in healthcare. As mentioned before, an economic side effect could be the discouragement towards robust research of advanced technologies in healthcare that could turn into a case of missing opportunity in economic and medical terms and affect the quality of care delivered to citizens. An analysis of other normatively relevant side effects would also need to include existing challenges of healthcare systems like diagnostic error rates, demographic transition, and workforce shortage.

The estimation of risks derived from the implementation of AI is complex for several reasons. First, it is not always clear who is imposing the risk. The implementation of AI models is a process where different actors participate at different points in the development timeline, starting from the stage of problem ideation to the monitoring phase of the deployed model, and also have varied levels of involvement. In this sense, it would be necessary to ascertain, for instance, who has more knowledge regarding the risks of the models, who has the decision-making authority or clearance to approve or halt the development process if a potential risk is identified, who is in charge to certify, inspect or authorize the use of a model in a clinical setting, among other aspects. Second, determining who are the recipients of risk is also not straightforward. Depending on the concrete application, the risk can signify different harms for different recipients. For example, a model erroneously diagnosing pediatric cases would pose a risk in three dimensions: A risk to the patient of a negative clinical outcome, a risk to the parents of the patient in terms of time and financial resources spent, and a risk to the physician in terms of liability and accountability. It would be thus important to define who the recipient or recipients of the risk are, and, from the perspective of a rights-based approach, which rights could potentially be at risk. Finally, a third aspect that makes the estimation of risks complex is where the risk originates because, depending on this, the normative elements to consider would change.

This chapter will address these challenges as follows. In Sect. 4.1. I will start by exploring the existing risks of healthcare systems and medical practice that ML models aim to solve or minimize. This includes the risks posed by diagnostic error, population growth, and distribution, and the shortage of the medical workforce. In Sect. 4.2. I will move on to examine the emerging risks brought by the

Clinical and Experimental Dermatology, November 19, 2023, llad402, https://doi.org/10.1093/ced/llad402; Ehsan Ullah et al., "Challenges and Barriers of Using Large Language Models (LLM) Such as ChatGPT for Diagnostic Medicine with a Focus on Digital Pathology—a Recent Scoping Review," *Diagnostic Pathology* 19, no. 1 (February 27, 2024): 43, https://doi.org/10.1186/s13000-024-01464-7.

introduction of ML in medical diagnosis. This examination will include risks of bias and discrimination, privacy, security, job displacement, deskilling, AI error, and environmental burden.

4.1 Existing Risks

In this section, I will discuss three phenomena that pose risks to individual patients, to patients as a group, to clinicians, and to health systems. These are the risks posed by diagnostic errors, the demographic transition characterized by increasing life expectancy and decreasing fertility rates, and the shortage of health professionals. These three challenges are global and affect individuals in different countries to varying degrees. Moreover, as will become clear throughout this section, these challenges are deeply intertwined and often have overlapping impacts and causes. The justification for choosing these three challenges is that the development of ML models in diagnostics aims to address them directly, in other words, the alleged benefits of ML models would at least help to mitigate these risks and, according to some, even become sustainable solutions for them[3].

4.1.1 Diagnostic Error

In the practice of medicine, the overarching aim is usually deemed to be the provision of high-quality health care to all patients. Such a task can be divided into specific objectives such as patient safety, effectiveness, efficiency, timeliness, equitability, and patient-centeredness[4]. However, due to its complexity and multifactorial nature there are instances in which errors or mistakes may occur that could compromise this aim of delivering high-quality care and affect the rights of patients. Diagnostic error is one of the most significant instances of this.

As it was explained in Chap. 1, the diagnostic process is complex in nature as it encompasses a variety of methods, approaches, actors, and media and as such there is not an agreed upon a general definition of what precisely diagnostic error means. However, since definitions are an essential starting point for any sort of

[3] Bertalan Meskó, Gergely Hetényi, and Zsuzsanna Győrffy, "Will Artificial Intelligence Solve the Human Resource Crisis in Healthcare?," *BMC Health Services Research* 18, no. 1 (December 2018): 545, https://doi.org/10.1186/s12913-018-3359-4.

[4] Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care*, ed. Erin P. Balogh, Bryan T. Miller, and John R. Ball (Washington, D.C.: National Academies Press, 2015), http://www.nap.edu/catalog/21794, 82.

problem evaluation and analysis, the Committee on Diagnostic Error proposed to define this type of medical error as: "the failure to (a) establish an accurate and timely explanation of the patient's health problem(s) or (b) communicate that explanation to the patient."[5]. From this definition, it is possible to extract several relevant elements to understand what diagnostic error *means*. First, there is the patient as recipient of the harm because of the error. Second, the notion of error is grounded on the failure to accurately determine what the set of symptoms displayed in the patient amount to and to do so in a timely manner, this means, within a time frame that allows the diagnosis to inform a treatment plan that ultimately leads to a possible, positive clinical outcome. It is important to establish that, while a diagnosis is correct if it explains the symptoms, if it occurs outside of a certain timeframe particular to each disease or condition, it might lead to an unwanted negative clinical outcome that may imply a deterioration of the patient's health, disability, or even death. That being said, a diagnosis that is successful does not guarantee a positive clinical outcome, which means that not all negative clinical outcomes are a direct or indirect result of a diagnostic error.

A third component observed in the definition is the communication aspect. A diagnosis might be achieved clinically, i.e., through laboratory testing, medical imaging techniques, clinician analysis, and other methods, but if it is not timely communicated to the patient, it becomes pointless. Failure to communicate a diagnosis can occur due to factors related to overburdened clinicians, faulty workflows, system errors, negligence, among others. Communication is a key aspect of the diagnostic process, and it is the responsibility of the clinician to see to it. However, communicating a diagnosis consists not only of merely providing a patient with test results. As mentioned in Chap. 3, for an explanation to be meaningful for the patient it requires to be adapted to his or her epistemic capacity, to be offered in a setting where it is likely that the patient will understand the implications of it and with an appropriate level of professionalism and empathy. For example, sending the results of a medical test that found a certain risk of Alzheimer's' in email form to an 85-year-old patient who lives alone and is not comfortable with computers would be a clear case of a failure to communicate a diagnosis correctly although the information was, technically, provided.

Diagnostic error is usually classified according to two criteria. The first categorizes the errors according to whether the diagnosis was incorrect, it was done

[5] Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care,* 85.

too late, or it was overlooked[6]. An incorrect diagnosis is the failure to establish the actual cause for the symptoms and it is instead mistaken for something similar. A late or delayed diagnosis occurs when it is not done within the appropriate timeframe and time is critical for the development of the disease. Finally, an overlooked diagnosis is one that did not take place at all and as a result the patient was harmed by the lack of action.

The second classification uses a five-dimension model of ambulatory diagnostic processes to help identify at which point the errors occur[7]. First, diagnostic error could occur at the patient- practitioner clinical encounter, if the information provided by the patient is not enough to reach a correct diagnosis or, for instance, in emergency settings, when the clinician does not have the time to review the medical history of the patient or perform a thorough examination. Second, at the point when the diagnostic testing is prescribed and/or interpreted. Error may occur here if, for example, the tests are unnecessary, mistaken or not reviewed correctly. Third, at the point where the clinician follow-ups with the patient. Error could arise if the results of the testing are not communicated appropriately or if, for instance, the management system fails to inform the clinician of abnormal results. Fourth, when consultation with another specialist is needed, and a referral is not given by the primary care clinician in time or when the referral is unnecessary. Finally, regarding patient-specific processes that include the ability of patients to communicate their symptoms accurately, the interest of the patient to seek care and willingness to adhere to the diagnostic process, for instance, attending appointments for testing, following-up with the clinicians, being willing to discuss the results and adhere to the available treatment options, etc.

In terms of the negative impact of diagnostic error a recent latest study on the population incidence of serious harm due to diagnostic error estimated nearly 795.000 cases combining deaths (~371.000) and permanent disabilities (~424.000) in the United States[8]. The data from this study shows that harm associated with diagnostic error is the largest source of death linked to medical error. The report of the Committee on Diagnostic Error confirmed that all people

[6] Mark L. Graber, Nancy Franklin, and Ruthanna Gordon, "Diagnostic Error in Internal Medicine," *Archives of Internal Medicine* 165, no. 13 (July 11, 2005): 1493–99, https://doi.org/10.1001/archinte.165.13.1493.

[7] Hardeep Singh and Saul N. Weingart, "Diagnostic Errors in Ambulatory Care: Dimensions and Preventive Strategies," *Advances in Health Sciences Education: Theory and Practice* 14 Suppl 1, no. 0 1 (September 2009): 57–61, https://doi.org/10.1007/s10459-009-9177-z.

[8] David E. Newman-Toker et al., "Burden of Serious Harms from Diagnostic Error in the USA," *BMJ Quality & Safety* 33, no. 2 (February 1, 2024): 109–20, https://doi.org/10.1136/bmjqs-2021-014130.

will experience at least one incident of diagnostic error, sometimes with serious consequences. This conclusion was based on the analysis of evidence on the epidemiology of diagnostic errors, supported by studies included in the report. However, they made it clear that a precise estimate was not possible due to the lack of access to systematic and up-to-date evidence[9]. Although there are no general standards of reporting of medical error in place that facilitate the study and comparison of this phenomenon, it seems evident that even the most cautious estimations are alarming enough to warrant seeing the problem of diagnostic error as a priority.

However, identifying the causes of diagnostic error has proven complex. Graber et al. have shown that diagnostic error is multifactorial and not, as previously believed, entirely a cognitive failure of physicians and other clinicians performing diagnoses. They demonstrated that diagnostic error can occur also due to systemic and technical factors, like inefficient procedures, coordination of care and even faulty or mis-calibrated equipment[10]. In the study, they observed that out of 100 cases where diagnostic error occurred, cognitive factors alone like faulty data gathering, data synthesis, and lack of sufficient knowledge corresponded to 28%, while purely systemic factors like organizational inefficiencies corresponded to 19%. Interestingly, the co-occurrence of system-related and cognitive factors amounted to 46%. This phenomenon was found to be happening when mistakes made at some early point in the diagnostic process would lead to further errors later on as a result of error propagation.

Cognitive factors are nevertheless essential to understanding the incidence of diagnostic error and to developing strategies to overcome it. One of the sources of cognitive error is actually the difficulties physicians have to recognize that they might commit errors themselves, partly due to the lack of feedback on their own performance and the resulting overconfidence in their diagnostic decisions. Another source is the overreliance on other's opinions, also called authority bias. This relates to the interconnected and complex nature of the diagnostic process that was developed in Sect. 1.3, i.e., that the diagnostic process is not just to the establishment of a differential diagnosis and therefore, the challenges that could lead to error also require contextual evaluation.

[9] Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care,* 355–356.

[10] Mark L. Graber et al, "Diagnostic Error in Internal Medicine," 5.

One of these challenges, and one that is inherent to the nature of the diagnostic process, is that diseases evolve over time. As illustrated by Zwaan and Singh[11], depending on the stage of progression of a disease, the challenges of diagnosing it are different and the outcomes can also vary. The stages of progression follow different courses of evolution and have different conceptualizations. For instance, in oncology, the progression of cancer is classified in two ways: The TNM (Tumor, nodes, and metastasis) classification and the stage classification. The TNM is based on testing results (imaging testing, biopsies, blood tests) and gives information about the size of the tumor, whether it has spread and if so, how much and how far in the body. The second classification divides cancer progression in 5 stages: stage 0, where there are no onset of cancer yet but there is the presence of abnormal cells that might lead to cancer later on; stage 1, where cancer cells are present but only in a localized place; stages 2 and 3, when the cancer cells have spread to nearby tissues or lymph nodes; and stage 4, when the cancer cells have spread to other regions of the body. This example of the staging of cancer is useful to understand that depending on what is occurring in the patient's organism, there are different types and amounts of information available to gather that is needed for the diagnosis. At stage 0 would not be useful to send the patient to chemotherapy since it is not clear if the cancer cells are going to spread, and a chemotherapy course would be actually disproportionately harmful instead. The treatment options are adapted according to the diagnosis, and, in this example, the patient would probably require targeted radiotherapy or targeted removal surgery.

A second challenge is the complex balancing between overdiagnosis and underdiagnosis. The general notion of diagnostic error as it has been defined in this section defines it in terms of underdiagnosis. The mistakes originate in the *lack of* timely, appropriate, and well-communicated decisions, in other words, the diagnosis was not made on time or at all. However, the practice of overdiagnosis can also bring consequences that could lead to harm and thus could also be considered a diagnostic error. Much in the same way that diagnostic error or underdiagnosis is complex to conceptualize, so is overdiagnosis. Brodersen et al. define it as "making people patients unnecessarily, by identifying problems that were never going to cause harm or by medicalizing ordinary life experiences through expanded definitions of diseases."[12]. Continuing with the example above,

[11] Laura Zwaan and Hardeep Singh, "The Challenges in Defining and Measuring Diagnostic Error.," *Diagnosis (Berlin, Germany)* 2, no. 2 (2015): 97–103, https://doi.org/10.1515/dx-2014-0069.

[12] John Brodersen et al., "Overdiagnosis: What It Is and What It Isn't," *BMJ Evidence-Based Medicine* 23, no. 1 (February 1, 2018): 1, https://doi.org/10.1136/ebmed-2017-110886.

clinical evidence has shown that detecting breast cancer in its early stages (1–2) is critical to patient survival,[13] but other significant evidence has also shown that detecting abnormalities very early (stage 0) has led to overdiagnosis and overtreatment[14], which is considered one of the most problematic consequences of this.

Normally, the action expected from a diagnostic result showing any kind of abnormality is to either move on with further testing to confirm the initial diagnosis, or to develop a treatment plan that deals with the disease with the aim of a positive clinical outcome. In the case of cancer, the goal is to eliminate all traces of cancerous cells in the patient's organism. However, a course of chemotherapy or complete breast removal would not necessarily be the most appropriate approach if the tumor is too small, for instance. Additionally, there are also the potential harmful side-effects of further unnecessary testing like being exposed to the radiation of X-rays and the emotional burden of being labeled with a diagnosis with a low chance of onset. Furthermore, the risks of overdiagnosis are increased due to the development of technology capable of identifying abnormalities in medical imaging with increased precision and a lower margin of error, like ML models. This matter will be analyzed in depth in Sect. 4.2.5.

Additional to the challenges between underdiagnosis and overdiagnosis, there are further decisions that clinicians need to make that may prompt one of these types of diagnostic error. Primary care physicians often need to balance the risk of missing serious illnesses with the need to use limited resources wisely[15]. In most contexts, a referral to imaging testing like an MRI or a CT scan is costly and limited by capacity, and physicians must be discerning of sending referral orders. For instance, a study on the cost of inappropriate diagnostic imaging showed that it accounted for over 20% of annual imaging costs for three common clinical conditions (lumbar spine, shoulder, and knee pain)[16]. Another study conducted in

[13] Marina Milosevic et al., "Early Diagnosis and Detection of Breast Cancer," *Technology and Health Care* 26, no. 4 (September 27, 2018): 729–59, https://doi.org/10.3233/THC-181277.

[14] Houssami Nehmat and Houssami Nehmat, "Overdiagnosis of Breast Cancer in Population Screening: Does It Make Breast Screening Worthless?," *Cancer Biology & Medicine* 14, no. 1 (2017): 1–8, https://doi.org/10.20892/j.issn.2095-3941.2016.0050.

[15] Hardeep Singh et al., "The Global Burden of Diagnostic Errors in Primary Care," *BMJ Quality & Safety* 26, no. 6 (June 2017): 484, https://doi.org/10.1136/bmjqs-2016-005401.

[16] Stephen Flaherty et al., "Magnitude and Financial Implications of Inappropriate Diagnostic Imaging for Three Common Clinical Conditions," *International Journal for Quality in Health Care*, January 23, 2019, https://doi.org/10.1093/intqhc/mzy248.

Canada with data between 2007 and 2012 showed that unwarranted routine imaging for early-stage cancer amounted to a cost between $4,418,139 to $6,865,856 Canadian dollars in this period of 5 years[17]. This factor warrants careful consideration as it presents a significant normative challenge. On the one hand, there is the uncertainty of whether the results of a routine testing or incidental testing (a result obtained from a testing made for a different purpose) could be indicative of a potential disease and on the other, spending resources that might serve no practical benefit to patients could be deemed to be a great financial burden on healthcare systems that are at least partially sustained by taxes.

4.1.2 Demographic Transition

A second challenge that must be considered is the phenomenon of demographic change. Although it is not generally included in the analysis of existing risks that have a direct impact on medical diagnosis, the consequences of this shift in the distribution of the population towards age groups that are more likely to require intensive and long-term care will inevitably affect health systems at various levels, including how internal workflows are affected by the increase in patients, the overall availability of human resources (which will be discussed in detail in the next section), the need to take into account clinical factors such as comorbidity (the simultaneous presence of two or more diseases or medical conditions in a patient) or polypharmacy (when the patient is taking several medications that may interact with each other), and so on.

This phenomenon can be explained by several driving factors. The first is the worldwide decline in fertility rates that began in the late 1960 s. Bongaarts explains that this can be understood as a consequence of two main factors: a) the decline in the previously desired family size due to the opportunity cost of raising children and the decline in infant mortality because of advances in sanitation and increased access to quality health care; b) the development and increased availability of family planning services, either voluntary or mandated, as in China, such as information on sexual health and contraceptive methods[18].

[17] K. Thavorn et al., "Cost Implications of Unwarranted Imaging for Distant Metastasis in Women with Early-Stage Breast Cancer in Ontario," *Current Oncology* 23, no. 11 (February 1, 2016): 52–55, https://doi.org/10.3747/co.23.2977.

[18] John Bongaarts, "Human Population Growth and the Demographic Transition," *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, no. 1532 (October 27, 2009): 2985–90, https://doi.org/10.1098/rstb.2009.0137.

The second factor is the continuous increase in life expectancy. Data from a recent report by the United Nations[19] on population growth trends show that, worldwide, people aged 65 and over will increase from 761 million in 2021 to 1.6 billion in 2050. Within this group, the cohort of people aged 80 and over is projected to grow from 155 million in 2021 to 459 million in 2050. The report also shows that the geography of the countries with the highest number of older population cohorts is shifting from Europe to East and South-East Asia.

At the European level, a report made by the European Commission in 2022 showed that while the population will not increase significantly in the next decades (it is expected an increase of the population of approximately 2.5 million by 2029 but then later decrease to 441.9 million by 2050), there is a significant change in the demographic distribution of cohorts. The population of 65 years old and older will go from 90 million in 2019 to 129.8 million in 2050. The percentage of persons aged 85 and over is estimated to double by 2050 while the population of less than 55 years old will decrease by 13.5%[20].

In Germany, a study made by Schoffer et al. to model the effects of demographic transition in the patient structure in German hospitals shows the following results[21]. The median of children and youths (aged 0–17) has decreased from 19.8% in 1955 to 17% in 2011, the people in the working age cohort (aged 18–65) went down from 64.4% in 1955 to 62.6% in 2011. Moreover, there is a continuous increase in the median share of the older cohorts: patients aged 65–75 have increased from 9.1% to 11.1% and 75 and older from 6.4% to 9.1% in the same years. The authors also found evidence that the demographic change correlates with the increase of diseases traditionally more prevalent in older patients like arthropathies (diseases of the joints like arthritis) and cerebrovascular disease[22].

These two factors combined result in a change in demographic composition of countries that brings forth challenges in terms of management of resources

[19] United Nations Department of Economic and Social Affairs, *World Social Report 2023: Leaving No One Behind in an Ageing World*, World Social Report (United Nations, 2023), https://doi.org/10.18356/9789210019682.

[20] European Commission: Eurostat, Louise Corselli-Nordblad, and Helene Strandell, *Ageing Europe: Looking at the Lives of Older People in the EU—2020 Edition.*, ed. Louise Corselli-Nordblad and Helene Strandell (LU: Publications Office, 2020), https://data.europa.eu/doi/10.2785/628105, 17.

[21] Olaf Schoffer et al., "Modelling the Effect of Demographic Change and Healthcare Infrastructure on the Patient Structure in German Hospitals—a Longitudinal National Study Based on Official Hospital Statistics," *BMC Health Services Research* 23, no. 1 (October 11, 2023): 1081, https://doi.org/10.1186/s12913-023-10056-y.

[22] Ibid, 6.

like healthcare expenditure, health workforce capacity and concerns related to public health. Declining fertility rates below the replacement threshold (the level at which each generation at least replaces the previous one) means that in the next generation there will be a decline of the number of people in the "economically productive" age group (15–64)[23] that constitutes the main bulk of the workforce and consequently, the tax contributions. Simultaneously, the increase of life expectancy rates means an increase in the dependency ratio[24] and could have a significant impact on health care expenditure that might not be sufficiently covered by the contributions of the cohorts in the productive age group.

According to data from the OECD, in terms of fiscal implications of this demographic transition, the expenditure in pensions in the OECD countries is estimated to increase by 3 to 4 points of the GDP, although with considerable cross-country variation. For instance, in Poland and the UK, it is expected that the expenditure will decline as private models become more attractive, while in Italy is expected that the expenditure will remain stable because of policy changes. However, in countries like Portugal is expected an increase of four points[25].

Regarding expenditure in healthcare, the matter is much more complex and varies significantly from country to country. While in the past some studies had estimated a dramatic increase in healthcare expenditure taking the age as the main decisive factor[26], at least two recognized theories have rejected this assumption as reductive and suggest that there are other factors that ought to be considered to evaluate the potential rising costs in healthcare expenditure: the red herring theory and the compression of morbidity hypothesis. Without going into too much detail, the red herring hypothesis argues that the focus on the factors of age and sex as drivers of healthcare expenditure is a red herring, i.e., a distraction from the real issue. The hypothesis explains that healthcare expenditure is instead correlated

[23] "Working Age Population." *OECD*. Accessed June 21, 2024. https://doi.org/10.1787/d339918b-en.

[24] An estimation that aims to measure how many "dependents" there are for each person in the "productive" age group. The ratio is estimated by equaling the ratio of population aged below 15 and over 65 to the population of age 15–64 at a given point in time.

[25] Thai-Thanh Dang, Pablo Antolin, and Howard Oxley, "Fiscal Implication of Ageing: Projections of Age-Related Spending," SSRN Scholarly Paper (Rochester, NY, September 1, 2001), https://doi.org/10.2139/ssrn.607122, 31.

[26] Ernest M. Gruenberg, "The Failures of Success," *The Milbank Quarterly* 83, no. 4 (December 2005): 779–800, https://doi.org/10.1111/j.1468-0009.2005.00400.x.

with the increasing morbidity present during the last years of life of a person[27]. The rationale behind this argument is that there is a tendency that in those last years (also called time-to-die TTD) is where the costliest medical interventions occur, and that people reach this timeframe of expenditure, naturally, at higher ages. In short, the red herring hypothesis does not dismiss that age is a factor that is correlated with healthcare expenditure but argues that it is *not* the decisive factor, and that is the TTD instead.

The compression of morbidity hypothesis proposed in 1980 by James Fries suggests that if the average onset age of chronic disease or disability in a person could be postponed faster than the increase of life expectancy from the same age, the period of morbidity could be compressed into a shorter span of time closer to the occurrence of death[28, 29]. This hypothesis has been demonstrated with scientific empirical evidence throughout the years although it remains a subject of open debate. The fiscal implications are that since most of the burden of illness, disability, and the costs of healthcare peak in these periods of morbidity, a compression of this period would lead to improvements of expenditure. Of course, this also opens the debate about societal preferences regarding accepting what we could call 'the natural course of life and death' or using medical resources to prolong life. In these scenarios, the role of medical technology and the costs of intensive treatments could cause the expenditure in healthcare to increase or decrease.

In both theories, the increase in cost still occurs but is much more moderate and there are scenarios where there is no increase but a trade-off between a greater number of gained years in good health vs increased morbidity in later years. Although there is not yet a definitive answer to this matter, studies have shown, with distinct methodological approaches, that while there is a certainty about increased healthcare costs in connection with the ageing rate of the population,

---

[27] Peter Zweifel, Stefan Felder, and Markus Meiers, "Ageing of Population and Health Care Expenditure: A Red Herring?," *Health Economics* 8, no. 6 (1999): 485–96, https://doi.org/10.1002/(SICI)1099-1050(199909)8:6<485::AID-HEC461>3.0.CO;2-4.

[28] James F. Fries, Bonnie Bruce, and Eliza Chakravarty, "Compression of Morbidity 1980–2011: A Focused Review of Paradigms and Progress," *Journal of Aging Research* 2011 (2011): 261702, https://doi.org/10.4061/2011/261702.

[29] Fries remarks that this hypothesis is only possible if we consider that the increase of life expectancy is limited. The theory is also based on the evidence of the incidence and prevalence of chronic diseases in the last decades that has replaced acute conditions as main causes of death globally. Further works of his and his colleagues on morbidity compression have admitted that real-life scenarios of morbidity are complex and usually do not follow a steady linear trajectory.

this is only one factor and not necessarily the determinant one[30]. Moreover, the impact of demographic transition on the use of healthcare resources requires us to examine the dynamic between life expectancy and health, broadly understood as low morbidity and disability rates. The metric of life expectancy alone is not sufficient to make a projection of costs or expenditure. Furthermore, the economic implications of healthy aging and longer life expectancy might differ from an individual level to a population level. What seems critical, however, is to look at the demographic composition and take an approach that consider multiple factors since there is no doubt that these changes will increase costs, especially if there is not an appropriate understanding of the circumstances and a lack of preparation.

4.1.3 Workforce Shortage

The workforce shortage is one of the most pressing challenges in healthcare in actuality. Qualified medical staff are undoubtedly one of the uttermost, if not the most, essential resources required for the delivery of quality care and to reach the goals of universal health coverage. The WHO made an estimation in 2016 of an overall shortage in a range of 16–19 million of workers by 2030 considering the evidence of countries under the threshold of a number of healthcare workers needed for appropriate delivery of care and the shortage of countries of the OECD[31]. This crisis is not new. Policy makers in the healthcare field as well as academics have warned about the impending shortage since the early 2000 s[32]. However, factors like increasing waves of migration tied to wars, socioeconomic precarity, climate change and religious and political persecution, and the COVID-19 pandemic have accelerated the development of the crisis and exacerbated weaknesses already present in some countries' healthcare systems.

---

[30] Daniel Howdon and Nigel Rice, "Health Care Expenditures, Age, Proximity to Death and Morbidity: Implications for an Ageing Population," *Journal of Health Economics* 57 (January 1, 2018): 60–74, https://doi.org/10.1016/j.jhealeco.2017.11.001; Fredrik Alexander Gregersen, "The Impact of Ageing on Health Care Expenditures: A Study of Steepening," *The European Journal of Health Economics* 15, no. 9 (December 2014): 979–89, https://doi.org/10.1007/s10198-013-0541-9.

[31] World Health Organization, *Global Strategy on Human Resources for Health: Workforce 2030* (Geneva: World Health Organization, 2016), https://iris.who.int/handle/10665/250368, 12.

[32] Christoph Aluttis, Tewabech Bishaw, and Martina W. Frank, "The Workforce for Health in a Globalized Context—Global Shortages and International Migration," *Global Health Action* 7, no. 1 (December 2014): 23611, https://doi.org/10.3402/gha.v7.23611.

The causes of the workforce shortage in healthcare are usually framed in simple terms of increased demand for care services and decreased availability of qualified personnel to cover that demand. In reality, however, the workforce shortage is a chronic and multifactorial problem that affects groups of people in different ways, and it is necessary to consider the perspectives of those affected groups to truly comprehend the phenomenon.

Increased demand is directly related to the challenges of demographic change discussed in the previous section. An overall increase in the world's population means an overall increase in demand, lower fertility rates are likely to affect the number of students in the health professions in the coming decades, and an increase in the number of people in the over-65 cohort means an increase in the demand for care services for these patients. Moreover, even if the compression of morbidity hypothesis continues to be reliable in the coming decades and the overall increase in health expenditure is moderate rather than exorbitant, the ratio of available personnel to the number of patients in need of care may still pose a number of risks for the parties involved.

Although the causes of this crisis are not homogeneous to every context where it is present, some of the main ones include: the implications of the demographic transition regarding the retirement of doctors over 55 years old, the problem of burnout, organizational and administrative hurdles, and the effects of the COVID-19 pandemic. First, the transition to retirement of the generation of baby boomers, born between 1946 and 1964. In 13 countries, 40% of doctors are over 55 which means that in a period of 10–15 years they will enter retirement. According to a micro census made by the German Federal Statistics Office (Destatis), it is estimated that by 2036 there will be 12.9 million of people belonging to the economically active cohort (18–64) that will have reached the retirement age (66–67 years). This represents almost 30% of the economically active population reported in 2021[33].

Second, burnout and high turnover. Due to the straining work conditions characterized by a demanding workload, long working hours, situational stress and other factors like lack of opportunities for further education, upskilling, and workplace issues. According to Looi, 70% of healthcare workers across Europe

[33] Statistisches Bundesamt (Destatis), "Press Release No. 330: 12.9 million Economically Active People Will Reach Statutory Retirement Age in the next 15 Years," Federal Statistical Office, accessed September 23, 2024, https://www.destatis.de/EN/Press/2022/08/PE22_330_13.html.

reported symptoms of poor mental health and 40% reported dealing with depression and anxiety[34]. Another study conducted by Azam et al. determined that the effects of burnout include intentions or decisions to leave the workplace, changing occupations, decreased professional efficacy, moral distress, emotional exhaustion and lower quality of life[35].

These effects are serious for the clinicians suffering from it, for patients that are indirectly affected by the decrease in the quality and availability of care, and for the health care institutions that are required to maintain certain standards of care and to remain financially viable. Furthermore, the study also showed that younger clinicians and women have a higher likelihood of suffering from burnout. Younger clinicians are unofficially required to work longer shifts in order to gain experience and confidence, and women, aside from working full-time jobs, are often also in charge of household and family care tasks. Added to these reasons, the COVID-19 pandemic is another factor that contributed to exacerbating the existing burnout problem. A survey study conducted in 2020 in 5 major hospitals in Singapore and India revealed that out of 905 participant healthcare workers, 34 (3.8%) were scored for moderate-to-severe levels of psychological distress, 79 (8.7%) showed moderate to extremely-severe anxiety, 48 (5.3%) screened positive for moderate to very-severe depression, and 20 (2.2%) for moderate to extremely-severe stress[36]. On a similar study conducted in 2020 with 580 healthcare professionals in Italy, 33.5% of respondents scored above the threshold for psychiatric morbidity[37, 38].

Another relevant aspect of the staff shortage crisis is that the burdens are not evenly distributed among countries and medical specialties. There are significant differences between primary and specialty care, between rural and urban areas, and between countries. First, primary care constitutes the first contact a

[34] Mun-Keat Looi, "The European Healthcare Workforce Crisis: How Bad Is It?," *BMJ* 384 (January 19, 2024): q8, https://doi.org/10.1136/bmj.q8.

[35] Kamran Azam, Anwar Khan, and Muhammad Toqeer Alam, "Causes and Adverse Impact of Physician Burnout: A Systematic Review" 27 (2017).

[36] Nicholas W.S. Chew et al., "A Multinational, Multicentre Study on the Psychological Outcomes and Associated Physical Symptoms amongst Healthcare Workers during COVID-19 Outbreak," *Brain, Behavior, and Immunity* 88 (August 2020): 560, https://doi.org/10.1016/j.bbi.2020.04.049.

[37] This score is taken from a psychometric assessment instrument called General Health Questionnaire used to evaluate the prevalence and severity of psychological and psychosomatic symptoms.

[38] Maria Laura Bettinsoli et al., "Mental Health Conditions of Italian Healthcare Professionals during the COVID-19 Disease Outbreak," *Applied Psychology. Health and Well-Being* 12, no. 4 (December 2020): 1063, https://doi.org/10.1111/aphw.12239.

patient has with the healthcare system and where the diagnostic process begins. To this type of care belong general practitioners, nurses, physical therapists, general pediatricians, among others. As a result, primary care clinicians generally have different responsibilities and an influx of patients that the other levels of care do not. Second, rural areas are generally less accessible, which is associated with fewer incentives for health workers to move or stay there, such as lower salaries and less infrastructure, which makes it difficult for people who want to settle with their families, as there are fewer options for schooling, other employment opportunities, and security.

Third, the inequity of distribution of healthcare personnel is clearly observed between high-income (HICs) and low-and-middle-income countries (LMICs). In a cross-sectional study conducted by the World Health Organization in Africa, it was revealed that there is an overall ratio of 1.55 health workers (physicians, nurses, and midwives) per 1000 people, while the threshold established by the WHO is of 4.45 per 1000 people in order to provide essential health care[39]. This difference is important because of the disproportionate impact of the burden on already fragile health systems and the fact that the share of the burden of disease affects these systems more. The Sub-Saharan Africa concentrates 25% of the global burden of disease and only 3% of the overall count of healthcare workers while the US and Canada, as comparison point, carry 10% of the burden of disease but have the 37% of the global healthcare staff[40]. The lack of access to qualified personnel capable of making timely diagnoses results in the population remaining ill longer, increasing the likelihood of complications, which in turn makes the system overburdened and financially unviable, causing more personnel to leave the areas, further exacerbating the situation.

Another aspect of the crisis is the migration of the healthcare workforce that once again affects LMICs more than other countries. HICs rely on qualified medical staff coming from LMICs. Boniol et al. assess the data about workforce reported by member states of the WHO in the National Health Workforce Accounts[41] and conclude that 22% of the world's population have close to half

[39] Adam Ahmat et al., "The Health Workforce Status in the WHO African Region: Findings of a Cross-Sectional Study," *BMJ Global Health* 7, no. Suppl 1 (May 1, 2022): 3, https://doi.org/10.1136/bmjgh-2021-008317.

[40] Yusuf Abdu Misau, Nabilla Al-Sadat, and Adamu Bakari Gerei, "Brain-Drain and Health Care Delivery in Developing Countries," *Journal of Public Health in Africa* 1, no. 1 (August 19, 2010): 20–21, https://doi.org/10.4081/jphia.2010.e6.

[41] "A system through which countries progressively improve the availability, quality, and use of data on their health workforce, and thus achieving universal health coverage, the United

(47%) of the global health workforce[42]. There are two main ethical implications of this phenomenon.

First, the country of origin is experiencing brain-drain and workforce-drain as a consequence of the migration of the health personnel and second, that countries spend significant economic resources to train their health care students (medical, nursing, and other specialties) and end up receiving no return of investment in the form of better quality of care and research for their own countries. Here, it is important to clarify that there is a difference between foreign-born medical staff and foreign-trained medical staff. Foreign-trained staff receive an education before arriving in the recipient country, which means that they benefit from the skills without having to invest in their education and training. Foreign-born medical staff only refers to the fact that a person may have a different birthplace than the country where they reside and work. A foreign-born healthcare worker may have been born in a LMIC, but this does not necessarily imply that they were trained there and does not infer on this analysis.

4.1.4 Normative Implications

These three overarching challenges discussed pose risks to individual patients and physicians, as well as to disadvantaged populations and already fragile systems. In this sense, to understand the implications of these risks, it is useful to take a panoramic view of the ways in which they are interrelated, overlap, and in some cases exacerbate each other. First, an overstretched health care workforce will struggle to meet the demand for care from a growing and, in some regions, aging population. Of particular concern are the dependency ratio and the potential increase in costs due to chronic diseases and the cost of technology. Second, as it was explained, the worsening of burnout rates among clinicians in primary care disciplines (general practitioners, pediatricians, nurses, etc.) has been shown to have an impact on their clinical capacity, effectiveness, motivation and commitment, which have been shown to be contributing factors to diagnostic errors.

From the perspective of the relational, rights-based normative approach proposed in this dissertation, it can be said that these three challenges pose, to a

Nations Sustainable Development Goals and other national and global health objectives" From: World Health Organization, *National Health Workforce Accounts*, 1.

[42] Mathieu Boniol et al., "The Global Health Workforce Stock and Distribution in 2020 and 2030: A Threat to Equity and 'Universal' Health Coverage?," *BMJ Global Health* 7, no. 6 (June 2022): 3, https://doi.org/10.1136/bmjgh-2022-009316.

greater or lesser extent, risks to the necessary conditions for people to live their lives. In the case of a missed diagnosis, the patient's health may deteriorate to the point where he or she is unable to communicate, move independently, or even be conscious. In some cases, it can lead to the loss of the patient's life, violating the most basic right, the right to life. The lack of availability of health workers will hinder the person's ability to access the health system when they need it. Constantly overworked clinical staff will have their physical and psychological integrity compromised.

The risks created by these challenges cannot be solved just with the implementation of ML models in healthcare. As will be shown and discussed at length in the next subsection, ML models bring forth their own set of uncertainties that require additional consideration. However, what *can* be argued is that we have a better understanding of these existing challenges as opposed to the uncertainties of ML models. We have reliable data that has been collected with multiple methodologies, by a variety of public and private actors, and that has been studied for decades. We have a fairly good understanding of the effects of these challenges to 10-, 15- and 20-year projections. This does not mean these modeling methods and predictions are flawless and provide absolute certainty, but they have been constantly peer-reviewed, debated, and confirmed and as such, what we are certain about is that we need alternatives to address the challenges and mitigate the risks as much as possible.

4.2 Emerging Risks

In the previous section, I delineated three existing challenges in healthcare that pose risks for individuals and groups of people at different levels. It was shown that these challenges are systemic and often overlapping in such a way that it requires both top-down and bottom-up approaches to understand their impact on the rights of people and take action to protect them. One of these approaches is the implementation of medical AI in various clinical and non-clinical settings. While this might have merit and could prove beneficial, as it was shown in Chap. 1, the fact remains that there are potential and unintended consequences derived from attempting to implement ML models that could represent a risk to people's rights.

In this section, I will focus on the risks emerging from the implementation of ML models in medical diagnosis or, more precisely, the potential risks from the incipient attempts at moving forward with this implementation. I make this caveat to acknowledge that, as analyzed in Chap. 2, the effective integration of

ML models in medical diagnosis has proved much more complex than originally anticipated, the risks that will be considered in this section are not risks from which there is a concrete idea of the extent and severity of the harm to patients and other involved actors as it was the case with the existing risks discussed in the previous section. Instead, the notion of risk used here must deal with an inherent uncertainty.

This aspect comes with benefits and disadvantages. On the one hand, the discussion about these risks will not need to justify its legitimacy with specific statistics or probabilities, as there is not enough of this data to support it. The justificatory source will be instead argumentative and, in some cases, drawn from similar contexts, e.g., the risks posed by EHR. On the other hand, the lack of these measures to back up the discussion might lead some people to debate their usefulness. To this potential criticism it can be argued that statistics are not necessarily the best way to measure risk and furthermore, that the matter of risk can be addressed from a variety of quite different perspectives such as risk identification, risk analysis, risk assessment and risk management, just to name a few. Each of these approaches has distinct methodologies and provides observations the rest do not. In this subsection, I will analyze and assess the risks of ML models in medical diagnosis from a normative perspective understood in its philosophical sense.

The matter of emerging risks derived from or strongly related to the implementation of ML models is complex. There are several factors that contribute to the complexity. In the first place, there are different levels at which risk may be present. For instance, at the technical level there are risks related to cyberattacks or algorithmic bias. At the clinical level, there are risks of medical error caused by ML models and derived risks associated with lack of clarity about responsibility and accountability allocation. There are also potential risks to interpersonal contexts like the relationship between physicians and patients and also risks to the environment associated with the resources used to train and run the AI-systems. Second, there is a multiplicity of actors with vested interests at these different levels. For instance, patients, clinicians, non-clinical medical professionals (technicians), data scientists or AI programmers, regulatory officers, companies, and health institutions, and it is sometimes not clear what roles and responsibilities they have or ought to have. Third, different types of ML models designed as health applications, for example, predictive models, diagnostic tools, or workflow tools, etc., may have distinct normative issues that would depend on their level of automation, and on a distinction between blanket risks, i.e., belonging to all models, and model-specific risks.

Some literature on the subject makes a distinction between risks as hard impacts to certain overarching values like health or safety, and broader ethical, social, and legal consequences as soft impacts[43]. Such a distinction is not useful at all as it misinterprets that risks can indeed be present at different levels and that can overlap, and that is why it is necessary to establish criteria to determine when a certain risk imposition is permissible or impermissible. Potential harm in terms of ethical consequences, like such of moral deskilling which will be discussed later on, are thus considered to be full-fledged risks and not merely soft impacts. In simple terms, a risk is the possibility of some type of harm befalling someone. Of course, we could speak on such terms of the risk of my ice cream falling to the ground if someone jumps in front of me. However, normatively relevant risks as understood in the subject at hand in this dissertation are those whose potential negative impact could affect the legitimate interests of individuals relating to the necessary conditions to be able to lead their lives. As such, the risks that will be discussed in this section are not restricted to the notion of risk as hard impacts. Instead, the risks included are evaluated according to the potential negative impact for different agents within the moral ecosystem that frames the implementation of ML models in medical diagnosis.

This section is structured thus: 4.2.1 will focus on the risks emerging from the different bias in the datasets, the algorithm and in society and how they can lead to discrimination in clinical settings. Sect. 4.2.2 will discuss the risks to the privacy of patients, in particular regarding health data; closely connected, Sect. 4.2.3 will examine the security risks to models in development and also when deployed. In Sect. 4.2.4, I will investigate risks particular to clinicians, such as deskilling and the potential deterioration of the patient-physician relationship. Sect. 4.2.5 will focus on emerging risks resulting from medical error directly derived from the implementation of ML models. Finally, in Sect. 4.2.6, I will consider some risks derived from the water and energy consumption required to store large amounts of data and to run the models.

4.2.1 Bias and Discrimination

In the early years of ML development, an often-repeated argument in favor of their potential benefits was that algorithmic-based systems could help solve some

[43] Tsjalling Swierstra and Hedwig te Molder, "Risk and Soft Impacts," in *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, ed. Sabine Roeser et al. (Dordrecht: Springer Netherlands, 2012), 1049–66, https://doi.org/10.1007/978-94-007-1433-5_42.

of the difficulties of diagnosis that were tied to human limitation. First, ML models were able to process and analyze data at a scale which humans could never be able to achieve, and second, since these models were fundamentally based on mathematical and statistical logic, they were purely objective, which is something people could not be assured to be. While the first assertion remains one of the strongest arguments, the second one has ever since been proven to be misleading. ML models are not subjective in the way people are, but they are nevertheless susceptible to bias present in the data used by data scientists or programmers to train them. As Crawford and Paglen point out "Datasets shape the epistemic boundaries governing how AI systems operate"[44]. That is why measures to ensure the quality and integrity of the data are a conditional first step towards harnessing the benefits of ML models for healthcare settings in a manner in which normative ideals like fairness and safety are prioritized.

The term *bias,* or *algorithmic bias,* usually describes a problem with the representativeness of the data used to train a model. There are several types of algorithmic bias that have been recognized and studied. First, bias as a consequence of missing data, a phenomenon called sampling bias or underestimation bias, where a biased dataset has more data points or samples of one feature or group of features than other equally important feature or group of features (an important feature is evaluated according to the context the model is meant to be representative of[45]). Second, measurement or classification bias, in which the data is mirroring an already biased context that may reflect on the biases of healthcare professionals. Third, label bias occurs when the labels (or outcomes) used to train a machine learning model reflect human biases or flawed data. This can lead to biased predictions, as the model learns from biased labels[46].

However, the phenomenon of bias is not restricted to AI or technology. As with measuring bias, there are existing social factors that play a determining role

[44] Kate Crawford and Trevor Paglen, "Correction to: Excavating AI: The Politics of Images in Machine Learning Training Sets," *AI & SOCIETY* 36, no. 4 (December 1, 2021): 1399–1399, https://doi.org/10.1007/s00146-021-01301-1.

[45] For example, let us think of a model that is meant to be used to predict risk of readmission of patients over 60 years old in a demographic context of 33.3% Caucasian, 33.3% Asian and 33.3% BIPoC (black and people of color) patients. If the model has a sample size of 70% Caucasians and 15% each for Asian and BIPoC patients, it will prioritize the Caucasian patients because it has more data on them, even if it has been established that BIPoC patients are actually more likely to need a readmission than Caucasian patients.

[46] Nathan Kallus and Angela Zhou, "Residual Unfairness in Fair Machine Learning from Prejudiced Data," in *Proceedings of the 35th International Conference on Machine Learning* (International Conference on Machine Learning, PMLR, 2018), 2439–48, https://proceedings.mlr.press/v80/kallus18a.html.

in how the data is collected, how the labeling process is conducted, i.e., how reliable are the annotations from human experts, in the case of medical images, which choices are considered when designing the algorithm architecture and the way the problem the model is trying to solve is formulated. In other words, the lack of data from certain populations or poor data modeling are not the only relevant sources of bias that may affect ML models. In social science, a broad conception of bias means an inclination towards one thing and not the other. One can say that a child shows a bias towards carrots over broccoli, for instance. However, the meaning acquires a normative importance when this inclination is considered to be unfair, which equates the notion of bias with harmful prejudice.

The nature and impact of bias have been researched extensively in cognitive science and social psychology, from which the studies of cognitive biases are grounded[47], and it is broadly agreed on that biases and prejudices cannot be completely avoided since they can be either conscious or unconscious[48]. It is part of human nature to be inclined towards certain options and this feature gets exacerbated by other factors like cultural codes, social norms, religion, education, etc. Human biases are also a target of actors interested in manipulating them for commercial and political gains. Propaganda, for example, is dedicated to nudging or creating certain biases in order to influence our behavior during the electoral period. Similarly, advertisement is designed to play on subconscious biases in ways that can impact our consumer behavior. Bias can thus be understood from different frames of reference.

As can be expected, biases also play an important role in healthcare and there are many studies that have shown the incidence of bias in medical professionals and how they may contribute to health disparities[49]. While explicit biases can usually be directly correlated with discriminatory actions, i.e., a human rights manager declining to hire a person because he believes that the ethnical group she belongs to is deemed to be lazy, implicit biases are much harder to pinpoint

[47] Martie G. Haselton, Daniel Nettle, and Paul W. Andrews, "The Evolution of Cognitive Bias," in *The Handbook of Evolutionary Psychology*, ed. David M. Buss, 1st ed. (Wiley, 2015), 724–46, https://doi.org/10.1002/9780470939376.ch25.

[48] Michael Brownstein, "Implicit Bias," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Fall 2019 (Metaphysics Research Lab, Stanford University, 2019), https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/.

[49] Ivy W. Maina et al., "A Decade of Studying Implicit Racial/Ethnic Bias in Healthcare Providers Using the Implicit Association Test," *Social Science & Medicine*, The role of Racism in Health Inequalities: Integrating Approaches from Across Disciplines, 199 (February 1, 2018): 219–29, https://doi.org/10.1016/j.socscimed.2017.05.009.

and can also influence a person's behavior subconsciously. In healthcare, a longitudinal study conducted with medical students to assess their racial bias in the span of four years concluded that the students that were exposed to discriminatory comments from supervisors or attending physicians, or that had reportedly negative interactions with patients of color displayed increased racial bias at the end of the study[50].

The matter of how bias affects AI and ML models cannot be restricted to the algorithmic bias observed in the datasets because the fact is that these systems reflect the complexities of our societies throughout the ideation and developing process and to the deployment phase and beyond. More concerning is that biases can be constantly reinforced by the models and the agents involved since both contribute different types of biases to the decision-making processes in healthcare.

An instance of this occurred in 2019 when scientists found entrenched racial biases in an algorithmic model used to predict the healthcare needs of more than 200 million patients in the US and to allocate resources for them in the form of assistance and health prevention programs[51]. The authors were interested in whether there were differences between white and black patients. First, they calculated an overall measure of health status and the correlation with race according to the algorithmic score, which showed that black patients had a higher burden of disease. However, when the algorithm was used to assign patients to the support and prevention programs, they found that black patients did not receive higher scores, even though they were sicker overall.

The measure of bias that the researchers found was that the algorithm used total medical expenditures per year as a label (see Sect. 1.2.2), which was found to be similar for both groups of patients. This showed a discrepancy between health status and health care costs, and the reason for this was not clear. The authors reasoned that because there is a disparity between needing and receiving care, the data used in the model reflected an existing socioeconomic bias and allocated fewer resources than would be fair, given the higher needs of black patients. The researchers concluded that the choice of label is a crucial aspect in ML development and required careful consideration at the technical level, but also at the point of the formulation of the problem because, in many cases, it can

[50] Michelle Van Ryn et al., "Medical School Experiences Associated with Change in Implicit Racial Bias Among 3547 Students: A Medical Student CHANGES Study Report," *Journal of General Internal Medicine* 30, no. 12 (December 2015): 1748–56, https://doi.org/10.1007/s11606-015-3447-7.

[51] Ziad Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (October 25, 2019): 447–53, https://doi.org/10.1126/science.aax2342.

be influenced by industry-wide approaches that bring along certain values and interests, like cost reduction and profitability.

Another problematic consequence revealed by this example is the potential emergence of a new phenomenon that could be named a process of continuous bias inheritance. It is clear that both people and ML models are contributing to decision-making processes in healthcare not only different capacities but also biases. ML models have the risk of adopting biases that are present in the socioeconomic contexts where the data is collected and left unchecked, can further exacerbate them, and end up generating discriminatory outputs. However, from this context the question arises of whether people can also inherit the biases of models and reproduce them even after the model is no longer being used.

A study conducted with psychologists offered the conclusion that there is indeed a high likelihood of these scenarios occurring since people in general, but also clinicians, can develop strong automation bias towards the models[52]. The lack of critical assessment of biases when employing ML models in healthcare settings can lead to their subconscious reinforcement by clinical professionals who may be more ready to believe the output of the seemingly logical model over their own analysis. A third reinforcement cycle would occur as more biased data would be created as a consequence of the automation bias and be left to be used by new models, completing a full circle of different types of bias constantly reinforcing each other.

It has been asserted that biased models are harmful because, without proper management, decisions based on their outputs left unchecked can lead to patients being discriminated against. This is the case for algorithmic as well as human biases. Discrimination in healthcare impacts on the right of patients to access medical care when they need it or can at least lower their chances of accessing the level of quality of care they require. This means that discrimination as a result of biased models or biased datasets may violate or infringe a person's claim to the highest attainable level of care as argued in Chap. 3. Furthermore, algorithmic bias comes with an additional challenge: it risks perpetuating the existing human and social biases through the model's outputs. Health data is collected from a variety of sources to which people in economically and socially precarious situations are not always privy to, especially in comparison with people living under better conditions. This includes access to telehealth apps and medical wearables but also includes the lack of access to healthcare facilities equipped to conduct

[52] Automation bias is defined as the belief that ML models by virtue of being mathematical processes would be more accurate, more reliable or more objective than the clinicians.

advanced testing, the lack of personnel with training and time to create comprehensive EHR and clinical notes and a solid strategy to manage the implicit biases of medical professionals[53]. Without participation in these contexts, the availability of data representing these groups of people is diminished, skewed or nonexistent which is reflected in the models and, as a result, individuals and groups of people that belong to this historically disadvantaged communities are affected the hardest.

It could be argued that a biased model is not necessarily to blame since the issue is the existing social, human, and systemic biases and the model is limited to reflecting these realities. However, this is careless at best and dangerous at worst. While it is a fact that a model's output is only as good as the data it is fed, this cannot be used as an argument to hand over the responsibility for the results or the choices that are taken based on the outputs. If biases in the data are a known risk of ML models that may negatively impact on the access to quality care or may risk physician harm due to an erroneous diagnosis or prediction, then there will be a normative duty to ensure that the data is adequately curated at every step necessary to prevent biased outputs as much as possible.

4.2.2 Privacy

It has been established that some of the most sophisticated ML models require great amounts of data. Although health data has always been considered valuable for the advancement of health technology, for public health monitoring and other purposes, its role in the development of medical AI has significantly increased the interest of the tech industry to acquire it. The added value of the data lies in that it can be used for many purposes other than what they are collected for. Some of these uses might be justified and could be even necessary or desired for relevant actors in healthcare. These include improving the performance of clinical ML models to deliver medical benefits to patients, such as improved predictive and diagnostic accuracy, and increasing the efficiency of clinical workflows to

[53] In many countries, the quality of the healthcare facilities is strongly determined by the socioeconomic circumstances of the place where they are located. Rural areas and poorer neighborhoods have lower tier facilities than in bigger cities and wealthier neighborhoods which affects their funding. Personnel is less inclined to work in these areas which impacts the quality of the delivery of care and their own well-being. Poorer areas statistically have more issues with violence which means a higher incidence of emergency and acute cases, and the facilities tend to have older and deficient equipment and supplies.

reduce bureaucratic and administrative tasks, such as report writing, scheduling, and billing, that are directly related to burnout and dissatisfaction.

However, other uses could violate the patients' right to privacy. The matter of privacy as a right as well as the potential risks of ML models to the privacy of patients regarding their private health data was discussed in Sect. 3.3.4. It was concluded that patients have a right to privacy, and this includes that they ought to have autonomy over their health data in terms of access to it, being informed about how data is being collected while in healthcare facilities and for which purposes is the data being used for. It was also concluded that since patient data is crucial to the functioning of the healthcare system and necessary to the provision of care, the healthcare systems could be allowed to use their data for clinical purposes pertaining to the patient and for the benefit of other patients as long as three minimum requirements are fulfilled: first, that the patient's data is not identifiable; second, that it is not shared with third parties without the express consent of the patient and third, that there are guardrails in place to make sure that any potential consent to share is given without constraints, for instance, patients should not be offered excessive incentives or being misled to share their data[54].

The risks to the right to privacy of patients are directly concerned with two main aspects: first, the collection and second, the use of health data, in particular, if this data can be identified. This means that the data can be recognized as belonging to a particular patient and be potentially connected with other types of personal data like purchases, online behavior, insurance data, etc. There are three types of privacy risk associated with data sharing: intentional or unintentional data leakage by authorized personnel, data theft, and virtual hacking[55].

First, data collection is the process of acquiring datasets that are suitable for the development of the ML model. As it was explained in Chap. 1, the data that are used in healthcare come from different sources and in varied formats, which means that it is often unstructured. Once acquired, the data then needs

[54] This argument draws from the logic behind regulations in place for the donation of blood and plasma. Blood banks and other organisations argue that better incentives, particularly economic, could boost the interest of non-donors to make the choice to donate. Blood donations are critical for many health institutions in poor areas and are life-saving in most scenarios. However, too high incentives could end up turning the structure of voluntary donation into a financial transaction. Similarly, incentives that become strong nudges to sell personal data would turn the patient into a data seller and could mean a shift in the duties of the healthcare institutions that, as buyers, no longer have the same responsibilities towards the person as patient. Furthermore, people in precarious situations could be targeted to become sellers.

[55] Barbara Hauer, "Data and Information Leakage Prevention Within the Scope of Information Security," *IEEE Access* 3 (2015): 2554–65, https://doi.org/10.1109/ACCESS.2015.2506185.

to be prepared so it can be useful for the training and validation phases. In order to protect the privacy of the patient, the acquisition of data generated by patients requires that the health institution or device collecting health data informs the patient about this fact, for instance, by providing an information sheet or displaying this information in the case of health apps. This becomes a more complex issue regarding data collected passively. While actively collected data requires awareness from the patient, data collected passively may be generated over periods of time without express participation of the user, for instance, by wearing a fitness tracker or leaving certain sensors activated in mobile phones. A systematic study conducted to appraise the literature on the subject showed that there are important ethical concerns regarding the lack of clarity of informed consent procedures, the disclosure about the amount of data collected and its expected use, and the lack of uncomplicated ways to opt-out from having data collected at any point of the user's interaction with the system[56].

Second, the main concern about the use of health data is that health care institutions, health data banks, or information brokers that collect and/or store patients' data may disclose the data to third parties, such as insurance companies, technology companies, or law enforcement agencies, which may then use the data for secondary purposes. As noted in Sect. 3.3.4, some types of health data, such as data related to mental health or substance abuse treatment, or sexual and reproductive data, are sensitive in nature. As such, there are concerns that law enforcement agencies could use them for purposes other than benefiting patients, such as for surveillance systems, which could lead to discriminatory policies, as with the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm[57]. Similarly, insurance companies might find it attractive to use health data for risk management or cost prediction. In addition to violating patients' right to privacy, such uses could violate other rights, such as fair access to health care or the right not to be discriminated against, because they could lead to increases in insurance premiums, denial of claims, or termination of contracts[58]. Furthermore, if the model is opaque or a black box, the results

[56] Nicole A. Maher et al., "Passive Data Collection and Use in Healthcare: A Systematic Review of Ethical Issues," *International Journal of Medical Informatics* 129 (September 1, 2019): 242–47, https://doi.org/10.1016/j.ijmedinf.2019.06.015.

[57] Julia Angwin et al., "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.," May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[58] Douglas B. Laney, "AI Ethics Essentials: Lawsuit Over AI Denial of Healthcare," *Forbes*, accessed May 23, 2024, https://www.forbes.com/sites/douglaslaney/2023/11/16/ai-ethics-essentials-lawsuit-over-ai-denial-of-healthcare/.

could not be explained, which could make it difficult to appeal or challenge the decision.

In this context we are confronted with a trade-off. The development of health technology relies heavily on the investment and expertise of private companies. This is not a new occurrence. However, it brings ethical concerns all the same. In the first place, there is a risk of enhancing the existing monopolies in AI developing companies. Partly due to the reliance on data collection, storage, and curation processes for the training of ML models but also because of the significant costs of computation capacity required to run the models in the first place and the cost of human talent to design them, the development of ML models is largely done by companies able to afford the costs and attract the right talent, which leaves smaller companies and independent providers with much less capacity to compete in the market. This creates a power imbalance that could have adverse effects for public institutions that become too dependent on the development of the technology. Such influence could be used by the developing companies as a means to lobby key actors in leading roles and public offices to further their private interests regarding policy and regulation against the duty of governments to safeguard people's rights. A cautionary example of such a situation could be observed from the recent debacle on the regulation of generative models like ChatGPT or Midjourney regarding the use of copyrighted texts and art, respectively[59].

Another issue in this context is the potential conflict between regulatory frameworks in different countries. It is not clear how companies that operate at the multinational level should be regulated and in particular, which jurisdiction should apply to the data they use for the models. Let us take the following example: a US-based tech company is developing a ML model using data collected through users of a certain health wearable around the world. The data is stored in the US and the development of the model is also done within this country. However, the model is deployed for its use not only within the US but reaches other countries. In such an example, there are relevant normative considerations distinct to processes of data collection, data storage, model development and model commercialization. At each point, a different geographical location or locations are concerned, and different regulations may apply, assuming the countries have enabled any sort of regulatory structures. First, the collection of the data should follow procedures to inform users about which data could be collected, in which manner, when, for which purposes, and ideally, the company collecting it should

[59] Mark Fenwick and Paulius Jurcys, "Originality and the Future of Copyright in an Age of Generative AI," *Computer Law & Security Review* 51 (November 2023): 105892, https://doi.org/10.1016/j.clsr.2023.105892.

allow users to opt-out of it. Second, the storage of the data should be placed under a regulatory framework that ensures that the data is appropriately secured. Third, the development of the model should also be secure against cyberattacks and other forms of hacking. Finally, the countries where the commercialization of the model occurs should make sure that its performance is safe for its use. The trade-off is then between the potential benefits brought by the development of models that in reality only powerful companies are capable of and the protection of individual rights and securing of public health concerns.

To minimize the risks to privacy described here there are both regulatory and technical approaches that have been developed. On the regulatory side, as it was mentioned in Sect. 3.3.4., there are frameworks like HIPAA in the US and the GDPR in the European Union who set strict rules to the distribution and sharing of data (in the case of HIPAA, health data in particular). A challenge of regulatory efforts, however, lies in the lack of uniform international standards that ensure that de-identification approaches are implemented efficiently and smoothly.

Technical approaches to data privacy include methods to de-identify and anonymize the data. De-identification is a process to remove or replace personal identifiable information (PII) like the patient's name or phone number, and protected health information (PHI) like test results or medical history from datasets[60]. This method was established and had been used long before the emergence of ML models to address the issues with the use of health data in clinical research and is divided into two approaches. First, data anonymization eliminates any data that could link back to the identity of the patient and second, pseudo-anonymization replaces the identifiable data with synthetic data in order to keep the dataset usable for purposes where PII and PHI might be needed. In fact, the use of synthetic data has gained the interest of clinical researchers in the AI field as a potential solution to privacy risks but also to address the issue of sampling bias (see Sect. 4.2.1). While de-identification is an established method and, particularly, anonymization is valued for its simplicity[61], it remains a challenge to make this process robust enough that can withstand potential attempts at re-identification. This is a growing matter of concern due to the development of algorithmic models based on ML that can re-identify previously anonymized

[60] Clete A. Kushida et al., "Strategies for De-Identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies," *Medical Care* 50 (July 2012): S82, https://doi.org/10.1097/MLR.0b013e3182585355.

[61] Mehmet Kayaalp, "Modes of De-Identification," *AMIA Annual Symposium Proceedings* 2017 (April 16, 2018): 1044–50.

data[62]. Another challenge is that the requirements to de-identify a dataset vary depending on the type of data used. This is the case with medical imaging data where identifiable features like the shape of a face can be linked back to a patient[63].

A second technical approach to mitigate privacy risks with special impact on healthcare is federated ML. Most of the techniques used in ML traditionally are based on centralized forms of training. This means that the data is stored in one centralized location and trained on a central server. While the data can come from various sources, the purpose of the centralized approach is to make the process as efficient and affordable as possible. The issues of this approach are related to privacy, as it is generally simpler to target a centralized place to steal or tamper with the data or the algorithmic architecture. On the other hand, federated learning is a type of non-centralized, also called distributed, ML approach that instead of centralizing the storage of data and training of the model, distributes the learning process to federated locations, also known as nodes. Each node relies on the principle of remote execution, which in simple terms means that copies of the algorithmic structure are distributed to the different nodes where they run the model with their own local data.

As its name suggests, the federated learning approach also has a central server that coordinates the model's learning process. Once the models generate some results of the computation, they are sent back to the central server to update the main algorithm and then the updated algorithm is distributed once again to the independent nodes. One advantage of this approach is that the data does not need to be shared among nodes, which minimizes the risk of leaks and attacks. This empowers the concepts of data governance and makes it an attractive approach to hospitals and other health institutions that require that models are specifically adapted to their own demographics or data availability while protecting the interests of patients in terms of privacy and the interests of health institutions in keeping the data secure in terms of ownership and liability[64].

---

[62] Kai Packhäuser et al., "Deep Learning-Based Patient Re-Identification Is Able to Exploit the Biometric Nature of Medical Chest X-Ray Data," *Scientific Reports* 12, no. 1 (September 1, 2022): 14851, https://doi.org/10.1038/s41598-022-19045-3.

[63] Gregory E. Simon et al., "Assessing and Minimizing Re-Identification Risk in Research Data Derived from Health Care Records," *eGEMs* 7, no. 1 (n.d.): 6, https://doi.org/10.5334/egems.270.

[64] Georgios A. Kaissis et al., "Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging," *Nature Machine Intelligence* 2, no. 6 (June 2020): 305–11, https://doi.org/10.1038/s42256-020-0186-1.

4.2.3 Security

The issue of security is closely related to the issue of privacy and discrimination, as it concerns the technical robustness of the systems. If the security guardrails of an AI system are breached, the model's data and algorithmic infrastructure are compromised and vulnerable to various risks that could affect the rights of patients as well as the public and private interests of healthcare institutions. It is important to distinguish between the concepts of security and safety, which, while concerned with negative outcomes, seek to address different types of threats. On the one hand, safety has a broader set of concerns related to the prevention of unintended harm resulting from the output or behavior of the system. Safety is concerned with the risks and harms that directly affect people as users or consumers, and in the context of ML model development and implementation, it includes issues such as fairness, transparency, value alignment, and so on. On the other hand, security focuses on ensuring that AI systems are protected against cyber-attacks, data breaches, unauthorized access, etc. The measures and procedures put in place are primarily aimed at the robustness and integrity of the systems themselves, and they are also designed to protect the interests of companies, investors, and other stakeholders in addition to users. This is the main reason why security in any IT development context is of paramount importance to companies.

There are different ways to approach security issues, and many are derived or directly drawn from the sphere of information technology. Given the complex nature of security risks and the multiplicity of ways in which a system might be compromised, approaches are used in combination with other approaches. One of them is threat modeling, a structured approach that aims to identify and prioritize security threats following queries such as who could be a potential beneficiary if the model's integrity was breached, what tools they have to carry on a potential attack, how could it be done, etc. There are multiple threat modeling frameworks and tools developed for specific needs like the MITRE, PASTA or STRIDE frameworks[65].

Another approach is the well-established Confidentiality, Integrity, and Availability (CIA) triad. Although it has been criticized for being narrow in its scope, the CIA triad remain core concepts for security in software development and now in AI development. The aspect of confidentiality requires that sensitive information or data are not available or disclosed to actors, companies, or other

[65] Lara Mauri and Ernesto Damiani, "Modeling Threats to AI-ML Systems Using STRIDE," *Sensors* 22, no. 17 (January 2022): 6662, https://doi.org/10.3390/s22176662.

processes without authorized access[66]. "Confidentiality" is particularly fitting to the healthcare context as it is already present and an essential aspect of the patient-physician relationship as well as a tenet of the Hippocratic Oath and it signals a responsibility placed upon the one(s) in charge of keeping a piece of information private. The aspect of "integrity" in this context refers to keeping the data consistent and complete throughout the model's lifecycle. If the data used to train a model is corrupted, it can impact on the model's output or on its performance, rendering unstable or unusable. Finally, the aspect of "availability" relates to its usability. A reliable model must provide authorized actors with timely access to the data and processes.

A recent literature review on the subject[67], proposed an analysis of security threats according to the component of an AI-system they target. The first type of attacks is oriented towards the system's data and the aim is twofold: to degrade the performance of the model or to manipulate the prediction outcome and there are three main types of attacks: triggerless data poisoning, the backdoor data poisoning, and adversarial attacks. A decreased model performance in healthcare means lower predictive accuracy which may lead to the model misclassifying the data, for instance, the pixels corresponding to a malignant tumor in a CT scan. The second type of attacks are model-oriented, and they seek to modify the training process or to manipulate the model once it has been deployed. Known types of threats include model poisoning, neural trojan attacks, model inversion, and model extraction, model stealing and adversarial attacks[68]. Since models can be the subject of attacks at the development stage and the phase of clinical use, it is imperative that the security guardrails remain robust at all times. This can pose a significant challenge due to the constraints placed on models that are being used for clinical purposes such as reliance on the model's performance, privacy concerns regarding the data being used with the model, and integrating the model within the security system of the context where it is being used (hospital, clinic, praxis, health center, research facility, etc.)

---

[66] Spyridon Samonas and David Coss, "The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security," *Journal of Information Systems Security* 10, no. 3 (2014): 21–45.

[67] Huaming Chen and M. Ali Babar, "Security for Machine Learning-Based Software Systems: A Survey of Threats, Practices, and Challenges," *ACM Computing Surveys* 56, no. 6 (February 23, 2024): 151:1–151:38, https://doi.org/10.1145/3638531.

[68] Mehran Mozaffari-Kermani et al., "Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare," *IEEE Journal of Biomedical and Health Informatics* 19, no. 6 (November 2015): 1893–1905, https://doi.org/10.1109/JBHI.2014.2344095.

4.2.4 Job Displacement and Deskilling

When assessing the risks of ML models in clinical settings, the focus is usually on patients, as they are the intended targets of medical procedures, as well as the moral agents on the receiving end of the potential harms or benefits of a diagnostic and treatment procedure. However, as will be analyzed in detail in Chap. 5, depending on the type of ML model, the patient is not necessarily the intended user. Instead, some of the most sophisticated models are designed as tools or aids for medical professionals in various medical subfields. This means that medical professionals, in their normative role as users of these technologies, tend to be recipients of risks that should also be evaluated. As such, I argue that medical professionals, apart from their normative status as duty bearers to patients and the public within medicine as a profession, remain rights holders as persons. This means that they have an equal claim to the necessary conditions to live their lives and should be protected from harm to their integrity in the same way as any other agent. The implementation of ML models in medical diagnosis could pose risks to them, and in this subsection, I want to highlight what those risks might be.

The first risk is the potential decrease in job opportunities. This one of the of the most discussed consequences is the impact of ML models in the healthcare job market. Although years ago, there was rising alarm about the potential of AI to replace jobs in healthcare at a speed at which the industry could not cope, the data in recent years has shown that although there are skills in healthcare that can be automated the hurdles for implementation have made this threat less pronounced although not altogether gone. Some experts argue that although it seems clear that AI will not replace medical professionals massively, it can generate acute disparities between those who use AI-based technologies and those who do not[69]. This could disproportionately affect medical professionals from geographical regions that have no access to such technologies and medical professionals that may not find feasible to acquire the skills necessary to understand and use AI-driven tools.

Closely related to job displacement, the second risk for health professionals is the risk of deskilling. In simple terms, deskilling refers to the loss or reduction of the skills necessary to perform a profession or job. This risk is unique in that the potential harm is directed at both the health professional and the patient. However, I classify deskilling as a risk for health professionals because it directly

[69] Curtis P. Langlotz, "Will Artificial Intelligence Replace Radiologists?," *Radiology: Artificial Intelligence* 1, no. 3 (May 2019): 1–3, https://doi.org/10.1148/ryai.2019190058.

and primarily affects them. The consequences of this phenomenon may also affect patients, as we will see later, but only secondarily.

Deskilling appeared academically for the first time in economic theory during the industrial revolution. Both Adam Smith and Karl Marx theorized about the origins and consequences of this phenomenon. Smith argued that deskilling was a consequence of the division of labor that came with the technical change introduced by the industrial revolution. According to him, division of labor was overall beneficial as it created more employment and affluence[70]. He was not overly concerned with how the loss of skills would affect ordinary workers, since it was expected that a greater division of labor would mean that some skills, though previously of great value, would become obsolete as machines took over. Workers were made to do simpler tasks that required less skill, and as the skills were lost, the workers became less skilled.

On the other hand, Marx perceived this phenomenon as mainly targeting workers who were at a disadvantage in comparison with factory owners and employers. According to Marxist scholar Braverman, Marx argued that skills were part of the bargaining power they had against their employers. Framed within his theory of the class struggle, he saw technical innovations as tools to deliberately reduce the skills of workers to replace them with less skilled workers[71]. Consequently, less skilled workers were less valuable and more easily replaced, which made them less expensive for employers.

The notion of deskilling as a process is then located at the center of the tensions between managerial decisions and the individual worker and has been widely criticized for several reasons. First, its seemingly deterministic nature as a "all-or-nothing" process; second, the disregard of the agency of the medical professionals as adaptable to new work conditions; and third, the lack of consideration for the ways in which individuals in managerial and worker roles can collaborate for their mutual benefit[72]. Deskilling, however, must not be simplified as merely the process of losing a skill in the manner one might lose a physical object, and the agency of a medical professional should not be dismissed as being passive. Hoff remarks that a clinician's way of adapting to the introduction of a

[70] Florian Brugger and Christian Gehrke, "Skilling and Deskilling: Technological Change in Classical Economic Theory and Its Empirical Evidence," Theory and Society 47, no. 5 (October 1, 2018): 667–69, https://doi.org/10.1007/s11186-018-9325-7.

[71] Harry Braverman, *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*, 25th anniversary ed (New York: Monthly Review Press, 1998), 56–58.

[72] Timothy Hoff, "Deskilling and Adaptation among Primary Care Physicians Using Two Work Innovations," *Health Care Management Review* 36, no. 4 (October 2011): 339–41, https://doi.org/10.1097/HMR.0b013e31821826a1.

new technology could be seen as integral in their own deskilling or upskilling process. He points out the case of EHR where clinicians developed skills to deal with the software to input patient data at the expense of dedicating time to talk to the patient. According to the study, the physicians were operating to fulfill the expectations placed on them by the health institution and insurance companies that required the completion of the records[73].

While there is no proven correlation between the usage of EHR and patient satisfaction due to the difficulties to measure it, there are studies that indicate to the effects of the pressure on the physicians to comply with managing health records and simultaneously engage with patients during consultations[74]. So, while the clinician's agency should not be dismissed, it is also important to consider how much those external expectations, that often follow a corporate business model aimed towards profitability, may influence the choices made by clinicians and as such, the responsibility for their potential deskilling should not be left to them alone. Moreover, this leads to the argument that the deskilling phenomenon, far from being a deterministic process, arises as a consequence of key clinical and organizational factors and the lack of proper analysis and management choices inside institutions.

A third view based on empirical evidence holds that technical change is the driving force behind the phenomenon of deskilling as well as the parallel phenomena of upskilling and reskilling[75]. These phenomena are shown to occur hand in hand with the introduction of different forms of technical advancement. Once the automobile started to being used, the skills related to carriage maneuvering and horse riding declined as they became less efficient modes of transportation. The invention of the electrical bulb displaced the skills required to make oil lamps and candles as they were no longer required to illuminate rooms in closed spaces. More recently, GPS technology replaced the map-reading skills that once were critical for many forms of navigation. This could perhaps lead to believe that the deskilling brought by the introduction of AI models in clinical settings would go through the same process and at the end the skills lost would be replaced with other emergent skills. However, as it was argued in Chap. 2, there are several reasons that indicate otherwise.

---

[73] Hoff, "Deskilling and Adaptation among Primary Care Physicians", 342.

[74] Tania Tajirian et al., "The Influence of Electronic Health Record Use on Physician Burnout: Cross-Sectional Survey," *Journal of Medical Internet Research* 22, no. 7 (July 15, 2020): 1–13, https://doi.org/10.2196/19274.

[75] While deskilling refers to the loss or deterioration of skills, upskilling refers to the acquisition of new skills while reskilling to the upgrade of previous skills into more advanced or desirable ones.

The phenomenon of skill-change as a constant tension between deskilling, upskilling and reskilling can be examined from different perspectives. On the one hand, the external perspective corresponds to the views of Smith, Marx, and Braverman. When certain production processes were replaced by machines during the industrial revolution like textile production, the jobs of many textile workers progressively disappeared as they could not compete with the faster and cheaper production of the same goods. However, other types of jobs appeared and along with them, new skills required to fulfill those jobs. There were low-skilled jobs, for example, to operate the steam machines, and jobs that demanded more specialized and qualified skills like those to repair and maintain the machines, to assess the quality of the goods, to administer the factories, etc. From a rights-based perspective, however, the fact is that in such a transition phase, individual workers are faced with the potentially existential consequences of impending job loss or decrease of income. Moreover, the disappearance of skills that could be considered to be inherently valuable, for instance, tied to artistic techniques, cultural tradition or of social value, could be of incalculable impact.

A skill can be defined as the ability to perform a task. As can be surmised, there can be many types of skills, some have to do with the ability to *make* something like build a table or paint a landscape, and some others can be the ability to *do* something like hold one's breath for long periods of time like in the case of professional free-divers or to convince someone to buy a product like in the case of salespeople. Skills are important because of different reasons. For instance, acquiring the skill of swimming could in high likelihood save the person from drowning; learning to ride a bike allows people to mobilize faster and could enable a person to have more freedom, to be healthier, to expand their social circle, even to access educational or job opportunities that were unreachable without this skill. In socioeconomic terms, skills are at the core of the structure of our development into sophisticated civilizations. Ever since humans started to group, skills were critical to ensure survival. We can think of primitive skills like hunting wildlife to consume, but there were other essential skills like finding optimal shelter away from potential threats or learning to tell apart edible from non-edible sources of nourishment. Some skills have been relevant at different stages of our development as a species. This is similarly the case in modern societies. Different skills are required depending on our social roles, our age, and they might even change drastically depending on choices we make as we live.

The level at which someone can perform said task relative to its maximum potential or result will determine the level of mastery of said skill. A person able to cook a meal would be considered someone with cooking skills, however, a chef who has practiced and perfected his cooking skills would be considered a

master in his occupation. Skills are thus abilities that are acquired and that can be improved. The improvement of a skill requires a certain amount of repetition over time. However, maintenance of certain skills also requires practice. Take, for instance, language skills. One can learn to speak German proficiently, but if one does not practice it regularly, after some time the skill would diminish and could be even forgotten completely.

In medicine, the notion of deskilling appeared around 1970 with the introduction of management models intended to reduce the cost of delivering care and the start of the implementation of electronic health records[76]. A study examined the effects of new technologies as a source of change potentially leading to deskilling in nursing between 1950 and 1990[77]. The author observed that while advancements in medicine have brought many benefits, it also posed new challenges to nurses that had to adapt to new demands which meant disregarding the practice of certain skills that until then characterized the field. For instance, the author noted that nurses were concerned about the loss of the skills to acquire knowledge about the patient and the community and thus their connection with them in the name of pursuing allegedly more sophisticated skills that were required by the introduction of biomedical machines[78].

There are two types of deskilling that could affect medical professionals: technical and moral deskilling. Technical deskilling refers to the loss or decrease of the quality of clinical skills as a result of not being practiced as often, not taught at all in medical schools, or not taught with the required rigor. This can be due to not enough time available to develop and practice them or the perception that they are no longer as important or even necessary at all. The technical deskilling of clinicians is an ethical problem because of the professional role they play as decision-makers, guides, and stewards regarding most aspects of the individual health of patients and also in matters of public health interest. Perhaps until now, it was not so acutely seen as a danger due to the progressive pace of the shift from discarding certain skills and acquiring new ones as technological developments appeared. However, the consequences of decisions made by medical professionals can have far-reaching implications and as such, this phenomenon warrants serious consideration.

Since the deployment of ML models in clinical settings has been occurring at a slow pace and also due to the complexity around measuring deskilling, or skill

[76] Hoff, "Deskilling and Adaptation among Primary Care Physicians" 339.

[77] Ruth G. Rinard, "Technology, Deskilling, and Nurses: The Impact of the Technologically Changing Environment." *Advances in Nursing Science* 18, no. 4 (June 1996): 60–69.

[78] Rinard, "Technology, Deskilling, and Nurses" 66.

shifts in general, there is not yet clarity about which concrete skills are at a greater risk of being impacted. However, in a mixed methods study conducted by Natali and colleagues[79], the authors identified seventeen concerns of the implementation of ML in clinical settings related to the deskilling of medical professionals. Twelve of these concerns were grouped according to the classification of clinical skills defined by the PACES-MRCPUK: physical examination, clinical judgement, clinical communication, differential diagnosis, managing patients' concerns and maintaining patient welfare[80]. The other five concerns were classified in two proposed items: operational and AI-specific .

Technical deskilling, according to this study, was observed as arising not just due to the introduction of ML models in clinical settings but to the attitudes of clinicians and institutions regarding their implementation and integration in existing clinical workflows. Most of the papers pointed out that the main risk associated with deskilling was of clinicians becoming over-reliant on them and not using their clinical skills as they would have without the models. This over-reliance would foster automation bias. The immediate ethical problem with this form of deskilling that comes to mind is that in the event of ML models failing, ceasing to function or making a mistake and clinicians not having sufficient skills to diagnose a patient correctly or make decisions regarding their treatment plan, then the rights of patients to receive appropriate care and even their right to life could be severely risked. Technological advancements in healthcare have brought immense benefits to patients, clinicians, and healthcare systems. However, it is the first time in modern medicine that we must consider leaving the judgement to something other than a clinician. Becoming reliant excessively on ML models to the point of dismissing clinical skills that have been proven essential for the practice of medicine would mean reducing it to mere pattern matching between data points when it is clear that diagnosing and treating patients requires much more than that.

The second type of deskilling that could affect medical professional is moral deskilling. It occurs when a moral agent loses the ability to make moral judgements and fails to develop necessary decision-making skills due to reliance on

[79] Chiara Natali et al., "AI-Induced Deskilling in Medicine: A Mixed Method Literature Review for Setting a New Research Agenda," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, March 5, 2025), https://doi.org/10.2139/ssrn.5166364.

[80] Practical Assessment of Clinical Examination Skills of the Federation of Royal Colleges of Physicians of the UK in: Andrew Elder et al., "What Skills Are Tested in the New PACES Examination?," *Annals of the Academy of Medicine, Singapore* 40, no. 3 (March 15, 2011): 119–25, https://doi.org/10.47102/annals-acadmedsg.V40N3p119.

automated models[81]. There is not a vast body of research dedicated to the conceptualization of what precisely constitutes a moral skill, and which normative implications are derived from such a conceptualization. There are much more studies on the subject of ethical decision-making or ethical competence[82]. However, Shannon Vallor considers moral deskilling from a virtue ethics approach grounded on a reconstruction of the Aristotelian theory of virtue[83]. In this conceptualization, moral skill are essentially Aristotelian virtues applied to concrete scenarios and they act as prerequisites for the effective development of practical wisdom and a virtuous character. According to her, a virtuous professional with honed moral skills would have the ability to "reliably discern the 'intermediate' or 'mean' course between an excessive and a deficient response, relative to circumstances"[84]. However, Vallor admits that moral skills are necessary but not sufficient conditions for genuine practical wisdom because an agent can fail at enacting the moral skill. In other words, one may possess a moral skill, but it might not be used in a way that exercises practical wisdom.

Vallor's notion of moral deskilling is helpful in this discussion because it expands the concept of skill beyond the traditional notion, which is often defined in terms of know-how and does not take into account the importance of moral choices that agents make in the course of their work. Moral skills in health care play a critical role in the delivery of care. They foster the fiduciary relationship between physicians and patients and ensure that patients' values, concerns, and interests are considered and respected. A conceptualization of moral skills from the rights-based approach proposed in this dissertation would be justified by two arguments: First, that skills are intrinsically action-oriented, meaning that, as with technical skills, a person must act purposefully to acquire and improve them. Second, moral skills are essentially skills that require moral judgment or have moral consequences. To put it more concretely, a moral skill would be to act in accordance with the normative assumption that every agent has a claim to the necessary conditions for leading his or her life, and to identify the best course of action in situations of conflict between the right claims of the moral agents

[81] Leslye Denisse Dias Duran. "Deskilling of medical professionals: an unintended consequence of AI implementation?" *Giornale di filosofia* 2, no. 2 (2021): 47–59.

[82] Jessica Hemberg and Håkan Hemberg, "Ethical Competence in a Profession: Healthcare Professionals' Views," *Nursing Open* 7, no. 4 (July 2020): 1249–59, https://doi.org/10.1002/nop2.501.

[83] Shannon Vallor, "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character," *Philosophy & Technology* 28, no. 1 (March 2015): 107–24, https://doi.org/10.1007/s13347-014-0156-9.

[84] Vallor, "Moral Deskilling and Upskilling in a New Machine Age," 110.

involved. While this may at first seem unfeasible, moral skills, like any other kind of skill, require practice and repetition, because even though we are moral agents by nature, our moral skills are not, in agreement with Vallor, innate traits. This seems obvious when we consider that agents everywhere fail to act in accordance with moral norms on many occasions.

Deskilling may pose a risk not only to the individual patient, for instance, in the case of technical deskilling that leads to medical error or to the individual medical professional if by losing certain skills his or her status as professional is lowered and he cannot perform his job with the same quality standards as before. It can also affect the interaction between medical professionals. For example, if a physician's clinical skills become unreliable, i.e., his or her judgement is erroneous, the examinations are not performed properly, she or he refers to the automated model by default, this might disrupt the clinical workflow intended as a collaborative effort to deliver care to patients. The threat of deskilling is multifaceted. It is not only a loss of technical skills but also an erosion of critical thinking, diagnostic acumen, and nuanced patient interaction that the ML model cannot replicate.

4.2.5 AI Error and Human Mistake

The risks associated with the use of ML models in the diagnostic process are most often assessed and evaluated from a purely techno-clinical standpoint. This means that risks are framed in terms of the sensitivity and specificity of the models, i.e., how likely is that the model returns true positives and true negatives (True positives are test results that correctly identify the presence of a disease while true negatives are tests that correctly identify the absence of a disease). In simple terms, risks are considered from the likelihood of a model returning an erroneous diagnosis, i.e., a false positive or false negative, and the consequences of such scenarios. Since the greatest value of the ML model, or at least the most marketed, is its predictive and classification accuracy, it seems logical that the technical efforts be directed towards making these metrics as close to 100 as possible. However, there are other normative considerations when dealing with mistake occurring in this context. If one asks: what happens when an AI model makes a mistake? There are several normative inquiries that follow: One, what kind of mistake are we discussing? Two, what are the potential or actual consequences of the mistake? Three, who is responsible and/or accountable for the mistake?

However, before tackling these questions, it is relevant to clarify between the notions of mistake and error. While they may be used interchangeably in general

speech, I hold that there is a significant difference for this section. From the Latin *errare*, to wander, an error is a deviation or departure from a standard, a correct value, an established "truth" or a in intended performance[85]. A mistake is an action, decision, or a judgement that produces an undesired or unintended result. Without delving too deeply into theories of action, I argue that making a mistake results from intentional human action, and that the algorithmic processes that characterize ML models cannot be called actions in this sense. The intentionality of an action denotes a series of normative qualities belonging to the moral agent that make possible the attribution of blame and the allocation of responsibility, among other considerations. Attributing a mistake to a model means assigning responsibility for the outcome. In previous sections (3.3.2) I have argued that the responsibility and accountability for an output cannot be ascribed to any form of AI or ML because, among other reasons, computer systems are not moral agents. They cannot make reparations to affected right-holders, request forgiveness or be held liable in court. Babushkina reflects that moral agents like physicians or nurses make certain commitments regarding potential risk of (medical) error to the patient, for example, to assess the risk beforehand and seek alternatives when possible, and acknowledge the harm if it comes to happen, particularly if the mistake is a result of their own actions[86]. Attempting to assign the ownership of a mistake to a model is problematic for two reasons: first, it removes the moral agent from the scenario and introduces what Babushkina calls a *moral substitute*, and second, offloads the blame from where it should be placed[87]. As such, from here on, I distinguish between an erroneous output or result as pertaining the model and making a mistake, to a person.

Types of model error and human mistake

The first thing that must be made plain is that error is virtually unavoidable as there is no model which has reached 100% accuracy rates. Furthermore, sensitivity and specificity rates may vary, in some cases significantly, from *in silico* development to real-world clinical scenarios due to poor data sampling

[85] *Oxford English Dictionary*, s.v. "error (*n.*), Etymology," September 2023, https://doi.org/10.1093/OED/3627921224.

[86] Dina Babushkina, "Are We Justified Attributing a Mistake in Diagnosis to an AI Diagnostic System?," *AI and Ethics* 3, no. 2 (May 1, 2023): 567–84, https://doi.org/10.1007/s43681-022-00189-x.

[87] Babushkina, "Are We Justified Attributing a Mistake in Diagnosis to an AI Diagnostic System?" 569.

(see Sect. 4.2.1) or to overfitting[88], among other technical reasons. Given these circumstances, it seems necessary to determine, in the first place, which accuracy threshold is acceptable for which models and in which applications[89], and then, what is the most appropriate way to deal with the risks of the remaining percentage of potential false negatives and false positives.

A second clarification following the distinction between error and mistake is that while a model might generate an erroneous output, the general clinical outcome of the diagnostic process ought not to be solely based on that erroneous result. The objective of a diagnostic procedure in broad terms is to achieve the best attainable clinical outcome for the patient (see Sect. 1.3.1) and a successful diagnostic process goes beyond assigning an accurate label to a set of symptoms, as explained in Chap. 1. As such, even if a mistaken diagnosis originated in the erroneous model's result, it is no longer just a matter of technical error but of a series of mistakes in judgement, for instance if the physician did not compare the model's output with their own clinical experience; or action, if the model's error is a result of incorrect use from the clinician.

We can then distinguish between three types of error: a purely technical error generated by the model, a model's error caused by human action, and a mistake made by a clinician based on the result of the model. Each of these scenarios brings forth particular normative implications. In the first place, technical error should be minimized as much as possible if the model is to be deployed for clinical usage and there should be a protocol in place based on state-of-the-art methods to detect errors in the model's performance as part of routine evaluations[90]. However, since eradicating error is not possible according to the possibilities of the models, then the clinicians in charge of the diagnostic process

[88] Data overfitting, a modeling error in which the algorithmic function of the model becomes too closely aligned with the data used in the training phase and that results in the model having a high performance for the training dataset but lower performance with any new data. Data overfitting is a significant problem because it ultimately makes a model useless as it is no longer able to achieve generalizability, i.e., the ability of the model to accurately predict an outcome based on the historical data used for the training (see Chap. 1).

[89] Establishing a clinical threshold or benchmark is a matter of deep complexity that does not belong within the scope of this thesis. As such, I stick to the literature's seeming consensus that the combined values should be between 1.5 and 2, 2 being a perfect test: Michael Power, Greg Fell, and Michael Wright, "Principles for High-Quality, High-Value Testing," *BMJ Evidence-Based Medicine* 18, no. 1 (February 1, 2013): 5–10, https://doi.org/10.1136/eb-2012-100645.

[90] See for example: T. V. Nguyen et al., "Efficient Automated Error Detection in Medical Data Using Deep-Learning and Label-Clustering," *Scientific Reports* 13, no. 1 (November 9, 2023): 1–19, https://doi.org/10.1038/s41598-023-45946-y.

must be aware of the positive and negative predictive values[91] and other metrics that show how many potential errors could the model make. Furthermore, at least at the beginning of the implementation phase, the golden standard used to diagnose the disease or condition that the ML model is being implemented for, should be kept in place and the model should be tested against it to ensure its validity. Even if the ML model becomes the best option to diagnose a disease or condition, the previous golden standard should not be discarded from clinical practice and fully eliminated from the curriculums of medical programs at universities. Since previous golden standards are not based on potentially opaque algorithmic processes in the way ML models are, they may provide important insight and help develop clinical skills that would remain critical in the scenario where the model is not available, for instance.

Second, an error as result of AI and clinician collaboration. In general, when speaking about human-in-the-loop strategies[92] or human-machine interaction it is necessary to consider two factors: a) that the system or model is designed with the particular human user in mind. This means that a model or system intended to be used in clinical scenarios must anticipate the needs of physicians, nurses, and other potential users. The user interface and user experience should be, for instance, simple to familiarize oneself with, intuitive and as little time-consuming as possible. A study conducted to evaluate unintended consequences of information technology in healthcare, found that human-computer interfaces were built under the assumption that when clinicians use information technology systems, they do so in isolation, with full-concentration and have time for data entry and retrieval tasks that may take time[93]. This unsuitable design ignores that the workflow of clinicians is full of constant interruptions, that they need to work with many other professionals on a daily basis and that they have limited time to complete routine tasks. An unsuitable design may lead to cognitive overload for the clinicians or to an increase in burnout, which might increase the incidence rate of mistakes occurring.

---

[91] Rajul Parikh et al., "Understanding and Using Sensitivity, Specificity and Predictive Values," *Indian Journal of Ophthalmology* 56, no. 1 (2008): 45–50.

[92] Samuel Budd, Emma C. Robinson, and Bernhard Kainz, "A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis," *Medical Image Analysis* 71 (July 1, 2021): 102062, https://doi.org/10.1016/j.media.2021.102062.

[93] J. S. Ash, M. Berg, and E. Coiera, "Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-Related Errors," *Journal of the American Medical Informatics Association* 11, no. 2 (April 2004): 104–12, https://doi.org/10.1197/jamia.M1471.

The second factor b) that the medical professionals must have the knowledge, experience, and motivation to properly interact with the ML models. If a clinician does not know how the model works in practice, how the results are displayed, and how to interpret them, this interaction can lead to errors and mistakes. Another source of error in the interaction of humans and models is labeling error. As explained in Chap. 1, supervised learning of ML models requires the establishment of a *ground truth* or standard from which the model learns the patterns, as an example, an area in an X-ray labeled as "tumor". The ground truth labels are annotated and validated by human experts. However, if the human experts make a mistake in annotating the labels used in the training data of the model, there will be a mistake in the training process leading to errors in the model's output. Labeling errors can be a result of poor image quality but also professional disagreement, for instance, in the case of outliers. In order to ensure that the human-model interaction leads to accurate outputs and correct decisions, it is necessary to work on the reskilling of medical professionals if they will be using ML models but also to consider carefully how to go about the integration of the model in the clinical workflows.

Finally, the mistakes made by the clinician derived from a model's error. As can be noticed, this third instance of something going wrong is referred to as a *mistake* which highlights the central role of the moral agent, in this case, the clinician. The most often mentioned risk in this scenario is automation bias, which has been briefly touched upon in Sect. 4.2.1. and 4.2.4. Although generally defined as overreliance on automated systems, the issue with this phenomenon is that it is difficult to determine if it is a consequence or a cause of further problems. In Sect. 4.2.1, it was described as a form of harmful bias affecting clinicians, and in Sect. 4.2.4 as a cause of other types of technical deskilling. As it is, automation bias is rooted in other similar phenomena, like that of authority bias mentioned in Sect. 4.1.1. At its core, it has to do with becoming overly confident in external sources for the decision-making process. While there might be a cognitive explanation as to why this occurs, for instance, because the agent wishes to find the moral substitute that Babushkina formulates or because there is a natural tendency to defer to those which are considered more competent, i.e., the experienced clinician or the seemingly objective and logical automated system, there are problematic implications that may affect the outcome of the diagnostic process.

It has been demonstrated from studies in different fields that an erroneous output has a high likelihood to lead the clinician's decision-making process astray[94]. While the formation of biases in people are still not very well understood and it does not seem feasible to eradicate them, a possible mitigation strategy to manage them in the medical field is to educate students, but also novice and experienced clinicians about the realistic capabilities of ML models. Much of the media and some part of academia's approach to it is overenthusiastic and careless about epistemic clarity. As it was emphasized in Chap. 2 these are the risks of overstating what AI can do, to set unrealistic expectations and pushing unproven estimates.

Consequences of the error

In Sect. 4.1.1 it was established that there are different types of diagnostic error that could be classified under two categories: underdiagnosis and overdiagnosis. Underdiagnosis is the phenomenon that has received most of the attention as it is the cause of an important number of medical errors in the practice of medicine and has severe consequences for the health of patients. Overdiagnosis, although not so intensely researched, also has the potential to harm patients and also clinicians. This is mainly due to the introduction of ML-driven technologies. Throughout this section I have been focused on the risks of AI error strictly defined as a deviation from the expected performance of the model. However, overdiagnosis is the resulting phenomenon of a different, perhaps more complex occurrence, namely, that the model is able to pick up on patterns of potential abnormalities at a scale and accuracy that was not possible before. This is called *overdetection.* This might sound unintuitive. As a general rule, having the tools and capabilities to accurately and efficiently diagnose a patient's disease or condition is highly desirable. Early detection methods are sought after in clinical research as they can drastically improve survival rates of afflicted patients[95]. ML models have been found to be useful tools to assist clinicians in this task, as was well shown in Chap. 1. Putting aside the exciting prospect of these technical possibilities for a moment, however, we must ask a fundamental question: is early detection useful in every case, for every abnormality or condition? At the heart

[94] Michael H. Bernstein et al., "Can Incorrect Artificial Intelligence (AI) Results Impact Radiologists, and If so, what can we do about it? A Multi-Reader Pilot Study of Lung Cancer Detection with Chest Radiography," *European Radiology* 33, no. 11 (June 2, 2023): 8263–69, https://doi.org/10.1007/s00330-023-09747-1.

[95] Crosby, David, Sangeeta Bhatia, Kevin M. Brindle, Lisa M. Coussens, Caroline Dive, Mark Emberton, Sadik Esener, et al. "Early Detection of Cancer." *Science* 375, no. 6586 (March 18, 2022): eaay9040. https://doi.org/10.1126/science.aay9040.

of this inquiry is a question about the goals of medicine in its quest for early detection, and how these goals relate to or overlap with the goals of patients.

Overdetection is the result of socio-technological factors like the refinement of testing machines able to produce high-resolution medical images, greater access to previously too-costly tests, self-testing technologies and much more engaged and autonomous patients. Another source of overdetection are so-called *incidentalomas* or incidental findings. This is the term to refer to abnormalities that are detected through testing for a different reason. One of the problematic aspects of overdetection is that it is not clear if detecting a condition or disease before there are symptoms actually improves clinical outcomes, and if so, how far ahead it is meaningful. There is a notable example of this occurring in South Korea in the early 2000 s where the incidence rate of thyroid cancer increased 15 times from 1993 to 2011, but the mortality rate remained stable[96]. The fact is that early detection alone is not necessarily always helpful because at the very stages of a disease, it is not possible to tell how, if at all, is going to progress. For instance, it is not uncommon that some small tumors never progress past formation and are dealt with by the person's organism on its own[97]. Another source of overdiagnosis is the overdefinition of the medical problem. Brodersen identifies two mechanisms: "lowering the threshold for a risk factor without evidence that doing so helps people feel better or live longer and by expanding disease definitions to include patients with ambiguous or very mild symptoms"[98]. Reasons for this range from the discovery of new aspects of a disease that were unknown before to the belief that a broader definition could help preventive strategies and avoid severe future complications[99], but they are also related to interests of private actors especially those connected with biomedical and pharmaceutical companies.

The impact of overdiagnosis is difficult to estimate clearly due to the underlined uncertainty about the likelihood of a disease's progression. However, there are clear potential harms from this phenomenon that fall into the category of iatrogenic harms, i.e., harms experienced by patients as a result of medical care

[96] Ahn Hyeong Sik, Kim Hyun Jung, and Welch H. Gilbert, "Korea's Thyroid-Cancer 'Epidemic' — Screening and Overdiagnosis," *New England Journal of Medicine* 371, no. 19 (2014): 1765–67, https://doi.org/10.1056/NEJMp1409841.

[97] Sante Basso Ricci and Ugo Cerchiari, "Spontaneous Regression of Malignant Tumors: Importance of the Immune System and Other Factors (Review)," *Oncology Letters* 1, no. 6 (November 2010): 941–45, https://doi.org/10.3892/ol.2010.176.

[98] Brodersen et al., "Overdiagnosis," 1–2.

[99] L. M. Schwartz and S. Woloshin, "Changing Disease Definitions: Implications for Disease Prevalence. Analysis of the Third National Health and Nutrition Examination Survey, 1988–1994," *Effective Clinical Practice: ECP* 2, no. 2 (1999): 76–85.

(It must be clarifies that iatrogenic harms are not necessarily harms derived from medical negligence although all negligence harms are iatrogenic harms). The first one is overtreatment, understood in this context as the treatment provided to the patient as a result of overdiagnosis and without sufficient evidence that it will provide a clinical benefit. It should be clarified that although overdiagnosis does not imply by definition overtreatment, it heightens the likelihood of it. Overtreatment can occur as surgery, prescription of drugs or with cancer, even adjuvant therapy[100]. The danger of overtreatment depends to some extent on the type of treatment to consider but in most cases as described above it can be invasive in nature. The detriment to the patient's health is twofold: being subjected to an unnecessary intervention and the possibility of her health being worse off as a consequence of the unnecessary therapy or surgery. For instance, a patient that undergoes radiotherapy for a small tumor will experience the residual effect of the radiation. If there is no sufficient justification for the risk undertaken, then the procedure will be harmful and should be prevented.

The second impact is the psychological effect of the diagnosis. For minor conditions or relatively harmless diseases it might be not seem to be a normative relevant consequence, but in the case of diseases like cancer that in some cases have low survival rates, a case of overdiagnosis might impact the psychological integrity of the patient severely, possibly impairing his or her ability to lead a life. Third, since overdetection is the by-product of improved testing technologies, like ML models for medical imaging, and an increase on testing being done, another impact that should be considered is the financial costs of the usage of the models associated with operationalization and maintenance.

Another not so acknowledged repercussion is the burden of managing the results of overdetection for clinical workflows and clinicians. Overdetection that leads to overdiagnosis is not a process occurring in a vacuum. When a model produces an output, it will generate an alarm that someone or some other system has to receive and manage. The result then needs to be placed in the clinical workflow in a series of actions that, at least for now, require at least some degree of human involvement, such as placing the result in the context of the individual patient, reviewing the entire case, scheduling a follow-up action (making a new appointment, contacting the patient, sending some form of communication, etc.), contacting other medical professionals, etc. Overdetection is not automatically

[100] "Adjuvant therapy is the additional cancer treatment given after the primary treatment to lower the risk that the cancer will come back. Adjuvant therapy may include chemotherapy, radiation therapy, hormone therapy, targeted therapy, or biological therapy" From: National Cancer Institute, "Adjuvant Therapy," February 2, 2011, https://www.cancer.gov/publications/dictionaries/cancer-terms/.

overdiagnosis, as this requires the involvement of a clinician. As such, overdetection can negatively contribute to already overburdened clinical workflow and clinician burnout.

Finally, a fifth impact identified is the so-called *diagnostic cascade* and *cascade testing* effects of overdiagnosis. The first term refers to the series of additional testing procedures that result from the initial finding usually required to verify the initial result or to determine more specific aspects of the potential disease or condition[101]. The second term refers to the process of informing the family or next of kin about a genetic condition found on the patient[102]. Since genetic conditions can be hereditary, the procedure of cascade testing exists to make the family aware in case they might need or want to be tested as well. The underlined issue in both cases is the normative imperative that requires clinicians to act after finding some evidence of the potential prevalence of a disease or condition. However, this imperative might be misguided. As mentioned at the beginning of this subsection, whether overdetection is beneficial is a question that needs careful consideration and in many situations, a casuistic approach.

The introduction of ML models able to detect patterns of actual and potential abnormalities in medical images has applications with demonstrated positive impact and potential positive benefits. However, the ability to detect patterns of disease must not be confused with a sufficient understanding of the nature of diseases. Even if we can generate this kind of information with outstanding accuracy, information alone does not lead to better decisions or improved clinical outcomes. This is yet another reason that cements the argument that clinicians cannot, should not, and will not be replaced by automated systems. We need to acknowledge detection and overdetection for what it is and judge its implementation in clinical scenarios critically. This is not meant to claim that we ought not to employ the best technology available for the benefit of the patients, rather, it is a call to reflect critically whether these approaches truly increase the likelihood of a positive clinical outcome and what this outcome should be according to the patient's goals, ethical, and legal considerations.

---

[101] Mor Saban et al., "Choosing Wisely in the ED: The Diagnostic Cascade of Needless Medical Testing in a Two-Level Study," *The American Journal of Emergency Medicine* 37, no. 9 (September 1, 2019): 1705–8, https://doi.org/10.1016/j.ajem.2018.12.017.

[102] Melissa K. Frey et al., "Cascade Testing for Hereditary Cancer Syndromes: Should We Move Toward Direct Relative Contact? A Systematic Review and Meta-Analysis," *Journal of Clinical Oncology* 40, no. 35 (December 10, 2022): 4129–43, https://doi.org/10.1200/JCO.22.00303.

4.2.6 Environmental

Throughout most of this dissertation, the discussion has focused primarily on the risks and opportunities of AI and ML models from their software component. However, these technologies could not exist and would not be as performatively powerful without their hardware component and other surrounding material infrastructure, as well as the resources they need to function, such as electricity and water. It has been shown in Sect. 1.1, as well as noted in several subsections, that one of the key factors enabling the development of DL architectures in the decade of 2010 has been the advances in the computational power of GPUs. Today, the sophistication and capacity of the hardware for training and running ML models has increased even further, and the manufacturing companies specialized in serving the needs of AI systems, such as Nvidia, Intel, and AMD, have acquired an almost unprecedented market value in the span of less than 10 years.[103]. However, on par with the leaps in terms of improvement are the environmental costs explicit and hidden generated by using state-of-the-art hardware solutions for ML applications.

The process of building hardware ought to be clearly typified as part of a complex network of resource providers, logistic chains, manufacturers and many hard to keep track of intermediaries. However, the complexity of these networks and the number of different actors involved at the different stages of the process to build chips, and computing servers bring forth many ethical issues and normative consequences, for instance, regarding human labor, new forms of colonialism and resource exploitation, that exceed the scope of this dissertation[104]. In this section I am to highlight only the most salient, known facts about the environmental costs, with a particular focus on energy and water consumption of training and using ML models in diagnosis and carbon emissions, and the normative implications of these costs for the rights of persons. An important caveat to make beforehand is that measuring the environmental impact of ML models is a task that until only recently began to be a matter of wide public and private concern and for reasons that will be mentioned throughout this subsection, the information about it can

[103] Dan Milmo and Phillip Inman, "Why Has Nvidia Driven Stock Markets to Record Highs?," *The Guardian*, February 23, 2024, sec. Technology, https://www.theguardian.com/technology/2024/feb/23/why-has-nvidia-driven-stock-markets-to-record-highs.

[104] For a deep dive into these matters see: Tamara Kneese, "Climate Justice and Labor Rights | Part I: AI Supply Chains and Workflows," *AI Now Institute* (blog), August 2, 2023, https://ainowinstitute.org/general/climate-justice-and-labor-rights-part-i-ai-supply-chains-and-workflows; Kate Crawford, *Atlas of AI—Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven London: Yale University Press, 2021).

be fragmented, incomplete and, in some instances, difficult to validate. However, many researchers have realized the importance of this task and the efforts to assess the costs of models and improve their efficiency and, as a result, there is a significant body of research on this topic.

Energy consumption

Every operation a computer performs has an energy cost. The physical building blocks of modern computers are transistors, a type of semiconductor that amplifies or switches electrical signals. For example, every time we type text into a computer, we are giving it specific instructions through the keyboard, which consumes energy. Typically, the more complex the set of instructions the computer has to process, the higher the energy cost, although the design of powerful computers prioritizes computational efficiency in terms of processing time and energy cost. When computers were first built, they could not perform very complex operations in a reasonable amount of time. Today, computers and devices such as cell phones can perform very complex computations, even in parallel, at a fraction of the cost. The operations performed by an ML algorithm are some of the most complex today, and although computational efficiency has improved dramatically over the past few decades, they are still very energy intensive.

The part of the development pipeline of a ML model that has been studied the most regarding energy consumption is the training phase, in which the model is fed the data and when the model must run the heaviest computations. According to information leaked in recent months, the training of GPT-4, OpenAI's most recent model available to the public required the use of approximately 25.000 Nvidia A100 GPUs over a period of 90–100 days[105]. An estimation based on data about the electricity consumption of the GPUs based on thermal design power (TDP), estimated that the model required between 51.772.500 kWh and 62.318.750 kWh PUE[106]. This is the equivalent of the yearly energy consumption of approximately 5000–6000 average households in the US[107]. However, with the development and recent deployment of generative AI models, especially

[105] Marc Bolitho, "Powering The Future: Meet The Scaling Energy Demands Of Generative AI," *Forbes*, April 22, 2024, https://www.forbes.com/councils/forbestechcouncil/2024/04/22/powering-the-future-meet-the-scaling-energy-demands-of-generative-ai/.

[106] Kasper Groes Albin Ludvigsen, "The Carbon Footprint of GPT-4," *Medium* (blog), July 18, 2023, https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae.

[107] It is estimated that the yearly average consumption is approximately 10.500 kWh. Source: U.S. Energy Information Administration (EIA), "Electricity Use in Homes," accessed September 20, 2024, https://www.eia.gov/energyexplained/use-of-energy/electricity-use-in-homes.php.

Large Language Models (see Sect. 1.2), the attention has also been set on the inference phase, i.e., when the model is generating content through queries and prompts sent by consumers on a regular basis, as a significant contributor in terms of CO2 emissions and energy consumption. This is believed to be the case because, although during the inference phase, the model has a significantly lower computation load, it is constantly running even if there is fluctuation in the usage and the inferences scale over time.

Water consumption

In a similar fashion, there are growing concerns regarding the usage of freshwater for processes like on-site cooling associated with the training and running of ML models in data-centers. Google reported that in 2022, the total water consumption at their data centers and offices was 5.6 billion gallons[108]. A seminal study on the hidden water footprint of data centers used to train and run AI models sought to estimate the operational water footprint of GPT-3[109], the service ChatGPT and GPT-4 are based on. The reason why the newer models are not often studied is because the key information necessary for the estimations and analyzes is not disclosed to the public under trade-secret and intellectual property arguments. As the information of which of Microsoft data centers was used for the training phase is not known, the authors made the estimation of each known center and according to known information about data centers and values of annualized average on-site PUE and water usage effectiveness (WUE) for each data center location. The data centers in Washington and Arizona were estimated the highest at 15.539 and 10.688 million liters a year respectively[110, 111].

The use of freshwater at the data centers has been criticized as lacking transparency and, in some respects, irresponsible, in the face of rising global

[108] Google, "2023 Environmental Report," 2023, https://sustainability.google/reports/google-2023-environmental-report/.

[109] The combination of on-site scope-1 water, and off-site scope-2 water. Following the terminology of GHG emissions, the authors define scope 1 according to on-site water for server cooling, scope-2 as off-site water for electricity generation and scope-3 as supply-chain water for server manufacturing.

[110] Pengfei Li et al., "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models" (arXiv, October 29, 2023), http://arxiv.org/abs/2304.03271.

[111] The authors used a methodology that accounts for particular conditions of the scope-1 consumption like the ratio of the on-site water consumption to server energy consumption, the variation depending on the outside temperature of the location the data centers, the type of cooling systems employed in the data centers (closed cycle, open cycle or hybrid) and for the scope-2 estimation, it is considered the energy fuel mixes and cooling techniques used by power plants as well.

temperatures, particularly in some areas where the data centers are located, where increasing amounts of freshwater are required to keep the servers from overheating. In 2023 tech journalist Karen Hao investigated the impact of Microsoft's data centers in Arizona for residents in the nearby towns and the responses from the company, which while providing modest disclosure about CO2 emissions and resource consumption, while keeping internal estimations and projections about the increase of costs of AI in upcoming years[112]. Companies like Microsoft, Google and Meta have made pledges about decarbonization, offsetting CO2 emissions and water-positive approaches for years, however, the question remains whether compensation and mitigation are sufficient measures in the wake of a demand to train and deploy more, and more powerful, AI systems. Furthermore, as Hao argues: "Carbon offsets and clean-energy power-purchase agreements may help Microsoft achieve carbon-negative and water-positive operations on paper, but they do not necessarily net out the effects on local communities"[113].

Normative implications

The data about the environmental impact of ML models discussed so far has centered on two aspects: the environmental cost or footprint derived from the training and deployment of models, and models that require significant amounts of computational resources which implies the use of large amounts of water and energy to function. The main normative problem in this context is the potential competing interests between companies investing in AI technologies that requires the use of resources needed to sustain life like freshwater or the emission of carbon dioxide (data centers do not operate solely or at all on renewal energy sources, at least at present) which threatens global warming, and the public health needs of people and their rights to the environmental conditions necessary to be able to lead their lives. While it is true that AI systems may contribute to more sustainable decisions and solutions, given their knack for making necessary processes more efficient and the unnecessary ones, void, the fact remains that big data is the fuel of models since the model is only as good as the data is fed.

Quality data in healthcare has its own environmental footprint. In a study conducted to measure the carbon footprint derived from the usage of the most used medical imaging modalities in an Australian public hospital[114], the authors

[112] Karen Hao. "AI Is Taking Water from the Desert." *The Atlantic*, March 1, 2024. https://www.theatlantic.com/technology/archive/2024/03/ai-water-climate-microsoft/677602/.

[113] Karen Hao. "AI Is Taking Water from the Desert."

[114] Scott McAlister et al., "The Carbon Footprint of Hospital Diagnostic Imaging in Australia," *The Lancet Regional Health—Western Pacific* 24 (July 1, 2022), https://doi.org/10.1016/j.lanwpc.2022.100459.

estimated that a single MRI scan emitted an average of 17.5 kg CO2e, and a CT scan an average of 9.2 kg CO2e; according to the contextualization made by the authors, this is the equivalent to the emissions of driving a new European car 145 km and 76 km, respectively. They argue that to reduce the carbon emissions associated with medical imaging, it would be necessary to reduce the overutilization of these modalities of medical testing. In other words, to reduce the unnecessary generation of images. However, since ML models are data intensive, there are strong financial and medical incentives to generate images as much as possible to train the models.

It could be argued that as medical AI develops, there is no longer such thing as unnecessary or low-value medical imaging scans because all quality scans are a desirable resource. However, this excessive interest in generating medical images could also contribute to the concern regarding medicalization and overdiagnosis exposed in the previous subsection, as well as bring risks to the privacy of patients. Furthermore, this also highlights the normative slippery-slope of attempting to transform every clinical challenge into a problem susceptible of being solved by AI and big data, a sort of "datafication" of medical issues.

Furthermore, in the scenario where the decision is made to make automated systems part of the critical infrastructure of countries, such as health care, mobility, and warfare, to name a few, we would need to consider further normative implications. Since most data centers are not currently powered by renewable energy sources, which remain unreliable, we would need to include the energy needs of the models that power this critical infrastructure as part of the consumption estimates in future projections. We would also need to demand that these companies, and not just the high-profile ones like Microsoft, Meta, and Google, adapt and comply with existing climate regulations and find climate-friendly production alternatives.

Finally, we must ask whether it is justifiable to overlook the environmental costs of generating, collecting, and storing medical data. As it stands, it is unclear how the increased demand for health data directly benefits patients, and instead seems to create an environmental burden that would be borne by the public[115]. It is necessary to clarify the diffuse distribution of benefits and the uneven distribution of burdens, so that in the end there is benefit for the people carrying the upcoming environmental burdens and not just a potential profit for data brokers, data center owners, and AI companies.

[115] This also brings the problems for public health associated and directly derived from global warming like the impact on incubation and spread of infectious diseases, exacerbation of respiratory diseases a result of poor air quality, malnutrition problems as a result of environmental catastrophes and lower availability of food and water, among many others.

The Ecosystem of Constellations

5

The aim of this chapter is to coalesce the technical, clinical, and normative elements presented throughout this dissertation and provide the blueprint for a normative framework to evaluate the potential risks and benefits of implementing ML models in diagnostic processes from a relational, rights-based approach.

In Chap. 3, I highlighted several criticisms of the atomistic conception of rights from proposals in feminist ethics and care ethics. However, I also argued that I do not reject the role of the individual agent as the starting point for evaluating potential benefits and risks understood as enhancing or hindering the realization of people's rights. Instead, my proposal for a relational turn on rights is to understand rights as mechanisms that emerge from the relationships that all human beings are exclusively capable of forming. These relationships are an inherent aspect of human nature and universal to all. However, the particular implications of how to evaluate which rights are justified, who are the duty bearers, and what is the content of those duties, as well as questions of special obligations and other considerations, must start from the assumption that persons are not merely isolated individuals, but exist as part of a network of normatively relevant connections and relationships with others. While this approach is based on recent developments in social science and moral theory and may attract criticism, I believe, as argued in Sect. 3.2.2, that it can provide important insights and guidance on the normative challenges of the use of new technologies in areas such as health care, where the quality of the relationship between moral agents has a direct impact on the enjoyment of fundamental rights, such as the right to adequate health care.

The concrete proposal for this blueprint is what I call an ecosystem of moral constellations. First, I use the element of "constellations" to describe groups of

L. D. Dias Duran, *Machine Learning in Medical Diagnosis*, Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics,
https://doi.org/10.1007/978-3-662-71357-0_5

normatively relevant relationships between moral agents at different levels and moments of implementing ML models in the diagnostic process. Second, the "ecosystem" element delineates the relational nature of the connections within the constellations and allows for a broader perspective to consider whether and how agents, institutions, and other actors influence each other. An ecosystem is defined as a complex and interconnected system of organisms that form a variety of relationships and need these relationships to survive. As such, each constellation aims to identify and explain the relationships that emerge at the various points of connection between agents and institutions involved in the process of implementing and using these technologies, how they might overlap, and what normative considerations emerge as a result.

It is important to note that this normative proposal is not meant to provide an exhaustive analysis of all aspects and nuances derived from the implementation of ML models in healthcare. Instead, I seek to establish the foundations for the normative evaluation of potential risks and benefits and justify why a relational, rights-based approach is suitable for this purpose. Moreover, ethical challenges should not be framed as if they are static problems that require one encompassing solution. In reality, it is often not possible to find an entirely conclusive fix to a conflict of interests in the context of rights. We are instead confronted with realistic scenarios in which we arrive at a tentative "best course of action" relative to the known-facts about the context at hand and the best ethical analysis possible depending on special or additional normative considerations.

While one aim of normative ethics will always be to provide reasons to justify certain actions and decisions, we must recognize that the contexts in which they are situated are subject to change, and therefore ethicists must be prepared to keep up with these changes and provide updated assessments and analyzes accordingly. Given the current state of development of AI systems and the speed at which they are being deployed in various domains, it is safe to assume that the ethical challenges arising from their implementation in critical domains such as health care are likely to continue to pose new challenges as technical capabilities increase and evolve. As such, normative work should always be seen as an ongoing task.

The chapter will be developed as follows: Sect. 5.1. will delineate the basic assumptions of the proposed relational, rights-based approach applied to the implementation of ML models in diagnostic settings. These assumptions are based on the arguments presented in Chap. 3, but are elaborated further. Sect. 5.2 will continue on to describe and characterize the four constellations formulated: clinical, operative, technical and regulatory. In Sect. 5.3, I will conduct an analysis of the normative aims identified at the different points where moral agents

come in contact with ML models. Finally, in Sect. 5.4, I will conduct an analysis of normative aims and will finish this dissertation with a theoretical exercise where I identify practical normative aspects of a hypothetical clinical encounter using the ecosystem of constellation framework and the relational, rights-based approach.

5.1 Basic Assumptions

There are basic assumptions about a relational rights-based approach that is necessary to establish for the analysis in the subsequent sections. The first assumption, and the one at the heart of the argument, is one that has already been established, namely that relationality is an inherent quality of human beings, from which rights arise. From this starting point derives a second assumption, that the contexts (social, economic, cultural) where these relationships form and develop are also underlined by the logic of relationality. In other words, these relevant normative relationships do not emerge in a vacuum, they are situated in socioeconomic and cultural contexts that are in part influenced by the relationships themselves but that also influence their formation and course of development. This applies to the different processes required to design and implement a ML model in a diagnostic process, i.e., from the beginning of the technical pipeline until the ongoing monitoring phase embedded within the clinical workflow of a particular clinical setting. There are several normative implications from this second assumption.

One, as argued in Chap. 1, ML models are socio-technical systems, and therefore the choices that people in leadership positions make about how best to address technical and ethical challenges are conditioned by factors present in their socio-economic environment. AI systems are not value-neutral, and whoever has the authority to make decisions about the technical development process also chooses whose values will be included and represented or prioritized in the outputs of the model[1]. For example, if only the decision-makers in a technology company developing a product for health care purposes are considered, there is a high probability that the interests and needs of people not represented in the decision-maker group will go unheard. This is particularly critical in a setting such as healthcare, where limited or no access to positions of decision-making authority has implications for users' rights.

[1] Abeba Birhane, "Algorithmic Injustice: A Relational Ethics Approach," *Patterns* 2, no. 2 (February 2021): 100205, https://doi.org/10.1016/j.patter.2021.100205.

Two, as socio-technical systems, ML models should be developed with the prior knowledge and understanding that their implementation implies the integration of an external, complex and potentially disruptive tool into complex existing clinical workflows, where medical professionals are the primary and direct users of the technology and whose needs must be at the forefront of the ideation process, together with those of the patients. Participatory design strategies can be useful here to address issues or needs that are known to clinicians, who are not only in close contact with patients, but are also well aware of the internal challenges within clinical workflows. For example, clinicians could provide valuable insight into the pitfalls of previous attempts to integrate similar models and clinicians' expectations for model performance. Such information could most likely help developers to better understand the challenges faced by clinicians and to design the models accordingly, and it could also prevent disappointment if a process of socialization with clinicians is undertaken before the model is deployed to set expectations in line with the actual technical capabilities and limitations of the models. Three, the moral agency of developers and computer scientists working at AI developing companies plays a critical role and should not be dismissed or downplayed. This includes the knowledge and awareness about their own moral responsibilities towards others and how to act upon them.

The third assumption is that the moral agency of medical professionals must also be recognized in a relational manner. Although in the medical profession the importance of the connection between patients and their physicians is broadly accepted, it is usually framed according to the duties of the medical professionals towards the patients. A posture like this can jeopardize the acceptance of the "self" as valuable as the "other," in this case the patient. Bergum criticizes that the traditional view of rights-based ethical principles "obliterate the difference between self and other, making us all objects, identical others, a world without selves, just others"[2]. Similarly, Shaw argues that ethical guidelines in medical practice rely in the ability of practitioners to separate the individual demands (of patients) from the context, promoting an "ethics of strangers"[3].

In rights-based moral theories, it is commonplace to identify both the duty-bearer and the right-holder. However, under the four pillars -particularly the first-, all moral agents, be it medical professionals, patients or developers, ought to recognize themselves as persons with human strengths and weaknesses and also

[2] Vangie Bergum, "Beyond Rights: The Ethical Challenge," *Phenomenology + Pedagogy*, January 1, 1992, 78, https://doi.org/10.29173/pandp14904.

[3] Elisabeth Shaw, "Ethical Decision-Making from a Relational Perspective," in *Ethics and Professional Issues in Couple and Family Therapy*, 2nd ed. (Routledge, 2016), 20.

acknowledge their simultaneous dual role as right-holders and duty-bearers. In essence, it means recognizing themselves at the center of their network of particular interconnections and relations with others. A practical example of this can be found in the correlation between clinician burnout and a decrease in the quality of care. If we first perceive physicians as human beings, it will allow us to see that they cannot strip off those weaknesses and strengths the moment they don the white coat. Interpersonal issues still affect them and their ability to make sound decisions. If a physician is in an emotionally vulnerable space of mind, this will influence their relationship with the patient during the consultation. This is similarly applicable to all moral agents involved. That is why this acknowledgment is necessary when evaluating the distribution of the burdens and benefits of ML models.

The fourth assumption is that the decision-making during the diagnostic process ought to be undertaken as a participative endeavor. Patients have unprecedented access to medical literature, health tools as mobile applications and wearables, and even to their own health data stored in the EHRs. This means that, in comparison to the models of patient-physician interaction, the medical professionals are no longer the only source of medical information. This corresponds with the shift from the generalized paternalism prevalent in medical practice until the middle of the 20th century to an approach that grants significant relevance to the autonomy of the patient. From a traditional paternalistic viewpoint, the patient has a mostly passive role in the clinical encounter and therefore, the physician carries the responsibility to be fully in charge of making all decisions regarding the patient's health. It was not uncommon that the physician would override the patient's decisions, values, or preferences in favor of his own opinion about what the best course of action was. The reasoning behind paternalism was the patient's well-being as the priority. However, as patients were generally considered lacking even a basic understanding of medical matters, and therefore incapable of making sensible choices about them, it was the physician's role to make the decisions on behalf of the patients[4].

The current circumstances are different. Instead of a passive role, the patient is encouraged and even expected to be active and engaged at the clinical encounter.

[4] While this assertion could be seen as paternalistic itself, the fact is: most people in Western organized societies did not have access to medical knowledge. In many societies, the role of the physician was one of great prestige and, as such, remained exclusive to persons of certain socioeconomic conditions. I am not considering societies in Eastern countries, tribal and other types of smaller communities where knowledge of medicine (understood as healing practices, herbalism, medical alchemy, and other alternative forms of medicine) could be more widespread and shared.

It is no longer acceptable, in terms of general societal perception but also of medical guidelines, to disregard or override the preferences, decisions, or values of patients. However, it must be noted that although we see the autonomy of the patient as essential for the practice of medicine, there is a potentially problematic effect of placing an excessive amount of emphasis on it. Responsibilisation is the recent phenomenon in which the responsibility for health harms is believed to belong mainly or solely to the patients. This posture is a result of self-management policies based on neoliberal postures that have emerged in the last decades[5]. Although this is a topic discussed mainly within the area of public policy and health prevention and promotion strategies, it can impact the relationship between physicians and patients if the guidelines of the institution or country where the clinical encounter occurs are strongly influenced by this notion.

As such, while a more active role in the diagnostic process should also entail a degree of responsibility and accountability for the patient, relationally this should be managed collaboratively along with the physician. For instance, there are diseases that are a result of or are significantly exacerbated by obesity. In some cases, the patients must manage their weight by changing their lifestyle to qualify for certain surgical interventions, like a bypass. This should not be framed *a priori* as the patient's "laziness" or "fault" because it ignores the potential coexistence of other health conditions, as well as socio-economic factors such as access to healthy food and availability of time to exercise, which play a direct role in managing weight.

The so-called "information age", coupled with the phenomenon of globalization, has enabled people with access to the internet to have more control over their health choices. Although perhaps an increased influx of information does not equate with people having improved critical judgment about how to make reasonable choices with it, the fact is that people are using it, regardless. Moreover, it is a concern of health organizations and systems around the world to improve the health literacy of the public as it has a direct impact on disease prevention and adequate management of health conditions[6]. Health literacy also plays a relevant role in the relationship between physicians and patients as it is a key aspect for effective and meaningful communication which can impact the

[5] Jacqueline Hutchison and Julia Holdsworth, "What Choice? Risk and Responsibilisation in Cardiovascular Health Policy," *Health* 25, no. 3 (May 1, 2021): 288–305, https://doi.org/10.1177/1363459319886106.

[6] Rabia Shahid et al., "Impact of Low Health Literacy on Patients' Health Outcomes: A Multicenter Cohort Study," *BMC Health Services Research* 22, no. 1 (September 12, 2022): 1148, https://doi.org/10.1186/s12913-022-08527-9.

achievement of positive health outcomes. Although it can also create scenarios in which it may be difficult for the physicians to navigate the clinical encounter with patients who are -or think they are- well-informed about their clinical condition. From a relational perspective, depending on the particular conditions of the diagnostic process, a participatory approach would mean engaging meaningfully with the patient, i.e., establishing channels of empathic communication and designing diagnostic plans together, involving close family members when necessary and appropriate, but also setting boundaries and expectations about what the patient should contribute to the relationship and what the physician can do for the patient.

The last assumption is that the use of ML models in the diagnostic process, from its broad and relational conception, cannot be reduced to a choice between two options. In other words, it does not merely entail that the user, i.e., the clinician, accepts or rejects the suggestions or outputs generated by the model. In Chap. 4, I discussed at length the topics of existing diagnostic error in healthcare and also the risk of AI error that raises important questions about responsibility and accountability. As seen in Chap. 1, AI models are achieving benchmarks and positive quantifiable results in a number of tasks that could help in clinical settings, but one of the most prominent barriers to the adoption of ML models is the factor of clinician acceptance. This is highly dependent on the knowledge and perception of the negative consequences of AI implementation. For example, research has shown that clinicians may be affected differently depending on factors such as how comfortable the clinician is with advanced technologies, how experienced he or she is as a practitioner, and what guidelines exist—and how clear they are—regarding the issue of medical liability.[7].

Acceptance also involves the question of trust both at the center of the fiduciary relationship between physicians and patients, but also regarding the ML models. I have argued that trusting in a strict moral sense is only possible among humans. Moreover, trust in healthcare has to do with vulnerability, empathy, responsiveness, and a mutual acknowledgement of one's moral obligations towards the other. One-sided trust is possible but is brittle and can ultimately hinder the diagnostic process, i.e., if the physician does not trust the veracity of the patient's experience or the patient does not trust the physician enough to disclose all the relevant information. Regarding ML models, the users cannot trust the machines as they are not something that is possible to trust in this strict sense.

[7] William Nicholson Price Ii and I. Glenn Cohen, "Locating Liability for Medical AI," *SSRN Electronic Journal*, 2023, https://doi.org/10.2139/ssrn.4517740.

However, if we conceptualize the notion of trust as confidence about the reliability of the technical capabilities of the models based on quantifiable benchmarks and standards, then it is reasonable to expect -and demand- that the models fulfill these standards so they can be employed in clinical settings. However, going back to trust in the strict moral sense, this notion is not truly directed at the model itself, but at the persons behind designing the regulatory standards, and at owners of the developing companies that have an interest to produce a reliable, successful product, be it for purely commercial reasons or also to fulfill regulatory demands and societal expectations.

While the interaction between clinicians and AI models is not one of trust[8], the clinicians employing these technologies still have to decide how to engage with the outputs. A novel study conducted in 2023 examined the behaviors of clinicians in an exercise where they engaged with AI-generated recommendations to treat hypothetical sepsis cases in an ICU. The authors included a methodology of "thinking-aloud" where the clinicians would voice their opinions, reasoning and decisions while going through the cases[9]. This type of clinical context (sepsis in an ICU) requires physicians to closely monitor the patient's progression and to decide carefully but quickly because of the danger posed by the co-occurrence of diseases, where a decision to treat one may affect the other(s), positively or negatively. Although the study aimed to understand the responses of physicians with experience in the ICU and with the management of sepsis, the results regarding the patterns of interaction are useful and likely applicable for models designed to provide recommendations regarding diagnoses. The study found that physicians had four attitudes in their interactions with the models regarding their recommendations about suggested treatments, dosage levels, and timing of administration: ignore, consider, trust, and negotiate. Out of the 24 participants in the study, 7 were not influenced by the AI recommendations, 3 considered it and decided conditionally, 2 accepted the AI recommendations in all decisions

[8] In this sense, the aim of building "trustworthy" AI systems should indicate that it does not consist in making the devices themselves trustworthy but that the focus lies instead in designing robust technical standards, the safety regulatory frameworks, legal pathways and other measurements that help users enhance people's rights and preferences and to prevent as much as possible the potential harms or undue risk impositions.

[9] Venkatesh Sivaraman et al., "Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI'23: CHI Conference on Human Factors in Computing Systems, Hamburg Germany: ACM, 2023), 1–18, https://doi.org/10.1145/3544548.3581075.

and 12 chose aspects of the recommendations[10]. For the present argument, the negotiation aspect is the most relevant because it suggests that doctors selectively assigned value or priority to different aspects of the model's recommendations and then evaluated them against their own intuition and reflections to find the best course of action.

These results highlight an important factor in the dynamic of physicians regarding the recommendation of computerized tools that, as it was explained in Chap. 1, are not new in healthcare contexts. It indicates that physicians have a broad range of ways to interact with these technologies that are not limited to a complete refusal or a complete reliance. It was shown in the study that even the physicians who ignored the AI recommendations were prone to develop their reasoning regarding their decision or criticize the model's recommendation, and the ones who accepted them after consideration first evaluated their own certainty. This ought to provide at the very least initial reasons to consider rejecting the most extreme worries about the ML models replacing the physicians' role.

5.2 Moral Constellations

The aim of using the notion of constellation to characterize the different relevant moral relations responds not only to similarities with the scientific concept that describes the perceived patterns that a group of stars forms during a period of planetary history. There is another methodological purpose, inspired by the basic aspects of modular programming. In simple terms, it is a software design technique that aims to build independent modules of code that fulfill one aspect of the program's functionality. It is usually a desirable technique when building a complex project and it ultimately seeks to collapse a complex problem into smaller, more manageable pieces without compromising the integrity of the program or sacrificing functionality. Similarly, the idea behind formulating each constellation is to make it easier to examine and analyze the relevant normative aspects of implementing ML models from the perspective of simpler relationships and of each relevant moral agent. A complete analysis would integrate the considerations derived from each constellation and evaluate them comparatively on a larger scale and with an external perspective. This would be an ecosystem level assessment. This is beyond the scope of this dissertation, and it will be left as future work.

[10] Venkatesh Sivaraman et al., "Ignore, Trust, or Negotiate", 12.

Focusing on each constellation would also facilitate addressing certain normative questions concerning the moral agents that constitute the relationship and the relationships themselves. There are four constellations that make up the ecosystem. First, the clinical constellation addresses the normative considerations that arise from the application of ML models at the clinical level and that affect relevant moral scenarios, such as the clinician-patient relationship, potential ethical conflicts during the diagnostic process, concerns about patient privacy and informed consent, and questions of attribution and accountability in situations of medical error. Second, the operational constellation, which includes normative considerations arising from the implementation of ML models in clinical workflows in hospitals and clinics, that have a relevant impact on the diagnostic process, such as electronic health record management and clinical workflow improvement. Ethical concerns at this level include the exacerbation of burnout among clinicians, clinician integrity and peer disagreement, the deskilling of medical professionals, and the cost-effectiveness of the models with respect to the allocation of scarce resources.

Third, the technical constellation, in which I consider normative questions and challenges resulting from the technical pipeline of design, development, and deployment process of ML models, regarding their impact on the rights of direct and indirect end users. The normative questions regarding direct-to-consumer diagnostic tools are also considered in this constellation and include questions regarding marketing strategies that oversell the benefits of the products, participatory design, machine bias, and transparency regarding explanations. Finally, the regulatory constellation contains the normative considerations that arise from the public health perspective in connection with medical diagnosis. At this level, decision-makers such as regulators, policy makers, public health officials and even civil society confront questions of ethical relevance like how to assess the distribution of burdens and benefits at a public level, for example, regarding the diagnosis and monitoring of infectious diseases or unnecessary interventions.

In the following subsections, each constellation in the ecosystem will be conceptualized following two fundamental questions: first, who are the moral agents that have a relevant role in each constellation and what prior moral relationships they already have that might be necessary to consider? And second, what type of relationships are formed among these moral agents? These questions are necessary to identify who the right holders and duty bearers are, which agents are able to make choices -and to what degree- about the diagnostic process and/or about the ML models that are implemented in the diagnostic process.

5.2.1 Clinical Constellation

The first constellation, as defined above, focuses on the relationships at the clinical level (see Fig. 5.1). There are two aspects of the relationships that exist in this constellation that characterize them: the level at which they form and the direction in which the moral duties are oriented. In the former aspect, there are three levels that can be observed: primary, secondary and tertiary, and in the latter, there are mutual and unilateral relationships.

First, the main relevant moral agents are the individual patient and his or her physician and, as such, the relationship between these agents is denominated as a relationship of *primary order.* These relationships are constitutive of each constellation, have a direct bearing on the diagnostic process and, unless under special considerations, should be prioritized. However, there are other relationships that also play a role in the clinical constellation, although they do not belong to the primary order relationship, under normal circumstances.

Secondary order describes the relationships that are derived from one of the primary order agents, but not from both, and that also play a role in the course of the diagnostic process and are connected to the main moral agents. On the patient's side, these include proxies and close family members who may play the role of caregiver or supporter. On the physician's side, secondary order relationships form with two groups of agents. First, with other physicians who have a role as members of the diagnostic team and who are indirectly needed for tasks tangential to the diagnostic process, or who are part of the profession and/or group of colleagues with whom the physician forms bonds of friendship and camaraderie. Second, with other patients for whom they are responsible and who, under the first pillar, have an equal right to the appropriate level of health care they require.

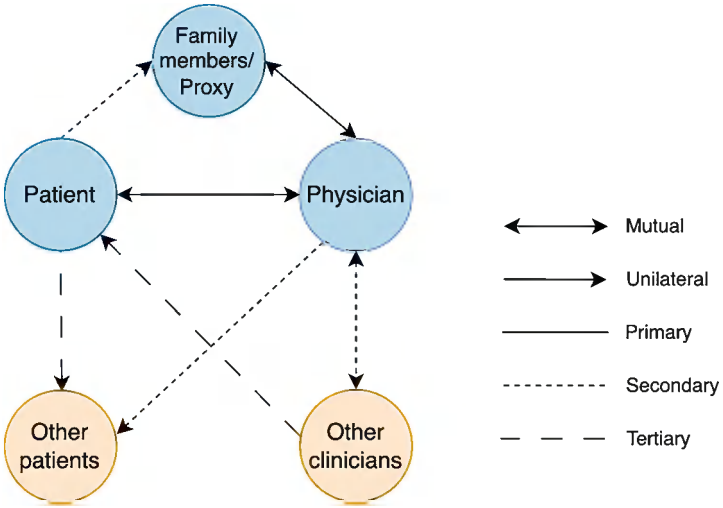

Fig. 5.1 Diagram of the relationships between moral agents in the clinical constellation

Finally, *tertiary order* relationships are those that arise situationally between agents of the primary or secondary order and other agents of the secondary order because of derivative moral considerations. For example, patients might form tertiary order relationships with the physician's other patients if the evaluation requires consideration of the burdens indirectly imposed on them. For example, when a patient arrives at an emergency room (ER), a triage process usually takes place to inform the physicians who needs to be prioritized. If the patient has a medical need that was deemed minor by the triage process, he or she must sit down and wait. The patient forms a tertiary relationship with the other patients in the ER and has an obligation to respect triage because it means there are other patients whose needs must be prioritized. Although the patient still has a right to medical care, in this case, the provision of care is situationally conditioned to occur after the most urgent patients have been diagnosed and treated.

Relationships are also observed according to the orientation of the duties of moral agents. In this sense, it is possible to identify two types of relationships within the clinical constellation: first, the relationships of mutual duties, which are characterized by the acknowledgement of the first pillar (see Table 5.1). It must be clarified that while this mutuality establishes that every agent has a set

of rights and duties that ought to be observed by the other(s) in the relationship, it does not mean that these are equal at all times. As it will be further developed in Sect. 5.4, the evaluation of risks and benefits will identify relational aspects of normative relevance that are necessary to adequately assess how the benefits and burdens are distributed under the second pillar.

Table 5.1 Definitions of the four pillars introduced in Chap. 3

First pillar	Second pillar	Third pillar	Fourth pillar
Every moral agent has an equal claim to the necessary conditions to be able to lead his or her life.	The necessary conditions (rights) form a hierarchy, necessary for evaluating conflicts and tensions.	Rights are simultaneously negative and positive (only if the person cannot help herself and, provided that assistance comes at no comparable cost).	The effective protection of a person's rights requires the intervention of public institutions, organizations and situationally, of private actors.

The second type of relationships, according to the direction of moral duties, are unilateral relationships, which in turn can be permanent or situational in nature. A relationship between the patient and his or her proxies or close family members may be mutual, unilateral, or situationally unilateral. Depending on factors such as the patient's level of dependence on the family members or proxy, the obligation may be temporarily unilateral. For example, if the patient is an infant, the parents are unilaterally responsible for making the decisions. However, once the child reaches the age of majority, the parents no longer have a unilateral duty but may have a mutual relationship when the patient requires medical care. In this example, while the patient is an infant or child, the physician has a direct unilateral moral obligation to the patient but must enter a relationship of secondary mutual obligation with the surrogate and/or guardian to make the diagnostic process and subsequent treatment process possible. A child or adolescent, in most scenarios, would not be able to follow diagnostic or treatment plans independently and would require the intervention of the proxy and/or guardian. Although these are mostly clear cases of how these two factors, level and direction of relationships, influence each other (see Table 5.2), this does not mean that the assessment of risks and benefits will be identical. As explained above, there are additional and special considerations that will also be taken into account

when evaluating each constellation, depending on the purpose and use of the ML model in question.

Table 5.2 Relationships between level and direction

	Primary	Secondary	Tertiary
Mutual	Patient–Physician	Physician–Proxy and/or caretaker (if the patient is situationally unable to join a mutual relationship)	Physician–Other clinicians
Unilateral	Patient (situationally unable to join a mutual relationship)–Physician	Physician–Other patients	Patient–Other patients

These normative considerations focus on the agents and relationships of primary order, in this case the patient and the physician, and the relationship they form (i.e., mutual, unilateral, or situationally unilateral). In the case of the patient, in addition to the capacity of the moral agent to make independent decisions during the diagnostic process (this includes the capacity to communicate preferences, values, and choices), other aspects that might contribute to this are, for example, whether and to what extent the patient wants to be involved in the process, and how well the patient is already informed about his or her medical condition. For example, the patient may be interested or willing to use a medical device to monitor a condition driven by ML or may prefer to go through more traditional means of testing in a scenario where both options are medically feasible and offer similar benefits.

On the physician side, such considerations might include how experienced the physician is and how comfortable he or she is with the ML device in question, how the diagnostic team is composed, i.e., what other medical professionals are involved in the diagnostic process as support. This is important because it would create a set of reciprocal obligations, such as sharing information in a timely manner and meeting certain standards, that could justify the use of a CDSS that uses ML to manage internal information in the EHR. Furthermore, it would also be of normative relevance to consider what level of care the physician provides to the patient (See Sec. 1.3) because this division of care into levels is intended to help patients receive appropriate and timely care according to their health needs, and to try to manage resources as efficiently and equitably as possible. However, depending on the health care system, patients may go directly to specialists

without a referral from their PCP. In the evaluation in Sect. 5.4, I will focus on primary and secondary care, as this is where the application of ML models to diagnostic procedures is more relevant.

5.2.2 Operative Constellation

The second constellation encompasses the normative considerations that arise from the implementation of ML models in workflows within clinical settings that have a significant impact on the diagnostic process (see Fig. 5.2)[11]. The use of ML in these scenarios is more commonplace than in direct clinical settings because the models are not meant to intervene directly in diagnostic decisions and therefore, do not face the same issues regarding clinician acceptance and regulatory constraints. However, as argued in Chap. 1, the diagnostic process understood from a broad relational lens is not limited to the cognitive action of assigning a label to a set of symptoms and takes into account the contextual and situated nature of the process for the clinicians as well as the patients.

Although the operational constellation includes the moral duties of health care institutions to clinicians and to patients, the constitutive relationship is between the physician and other clinicians, i.e., the moral agents within the working environment of a clinical setting[12]. While it is true that the medical profession is primarily patient-centered and, as such, the codes of conduct and professional deontological norms that clinicians follow are designed with the primary goal of providing appropriate medical care to the patient, the health care institution is also a structure composed of moral agents who are faced with decisions of moral significance on a daily basis. Thus, while most decisions have the patient as the ultimate beneficiary, there are decisions that should be made to maintain and promote the sustainability of the institution. Such decisions would have a direct impact on the clinicians (and other non-clinical staff) who are part of the institution. In addition, there are scenarios where moral choices need to be made that may involve a conflict between the rights of the individual patient and the rights of the public to whom the health care institution and clinicians also have obligations.

[11] The role of the patient and the relationships with healthcare institutions and the physician are shown in gray to indicate that they are not considered in this constellation.

[12] These are the reasons why the patient and the relationships derived from this role are shown in grey in the constellation's diagram (see Fig. 5.2).

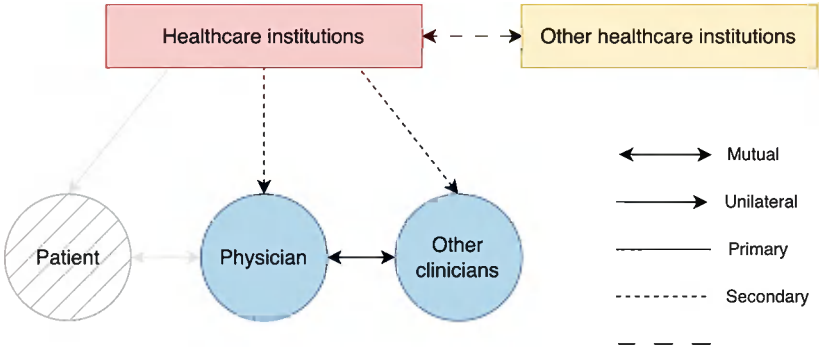

Fig. 5.2 Diagram of the relationships between moral agents and institutions with moral responsibility in the operative constellation. The role of the patient and the relationships with healthcare institutions and the physician are shown in light gray and with a diagonal hatch to indicate that they are not directly considered in this constellation

In the evaluation of the distribution of risks and benefits from implementing ML models in this constellation, I consider relationships of primary order those between clinicians, of secondary order the relationship between the healthcare institution and the clinicians, and of tertiary order the potential obligations between healthcare institutions, for instance, regarding establishing a system of interoperability and portability that allows the synchronization of EHR among institutions. In fig. 5.2, although the connection with the patients is included, it will not be the center of the conceptualization of this constellation but will be referred to instead in connection with the mutual obligation between clinicians and occasionally, with the duties of healthcare institutions towards them.

There are two overarching normative aspects that are of relevance in this constellation and that require ongoing consideration. First, there is the matter of what are the obligations healthcare institutions have towards the clinicians and patients. Particularly, it is of interest to explore how the introduction of ML models requires reevaluating existing protocols and institutional norms that are already in practice, and also if there are emerging obligations to ensure that the models provide benefits to the clinical workflow but also observing the potential long-term risks like the deskilling of the professionals using the models. Second, regarding the obligations between the medical professionals and how they balance against their obligations towards the patients they are treating and the obligations

as members of certain institutions, for instance, regarding choices that are costly for the institution but beneficial in the other senses.

One of the aims of the operative constellation is to highlight the complex role that clinicians have, not only as health practitioners with a fiduciary responsibility towards the patients but also in the position as health administrators within the institutions they work at, and as co-workers in settings where collaboration and teamwork are essential for the functioning of the clinical workflows, which have a direct impact on the performance of the institution but also on the delivery of care to the patients. The importance of this emphasis is what it has been hinted at through this chapter, namely, that physicians and clinicians in general are interdependent to each other and also are heavily reliant on the structure of the specific institution where they work -and to an extend on the healthcare system- to be able to deliver high quality of care to patients. If a physician is, for example, entrusted routinely with a high volume of patients without the support of nursing staff, other clinicians and a suitable administrative infrastructure, he or she will not be able to fulfill his or her moral obligations towards the patients.

To discuss the first aspect mentioned, i.e., the moral duties of healthcare institutions, it is necessary to first clarify which healthcare institutions are considered. By definition, healthcare institutions include any form of clinical facility that provides clinical services to patients. However, not all clinical facilities provide diagnostic services and therefore, I limit my analysis to institutions like hospitals[13], clinics and doctor's offices in which the ML models would be used for diagnostic purposes and exclude others like ambulatory surgery facilities, pharmacies and drug stores, medical laboratories and medical nursing homes. Furthermore, the healthcare institution is conceptualized here as an entity constituted not only by its legal personhood or its physical facilities, but including its organizational structure made up by individual moral agents connected at various levels and forming different relationships and governed by a set of guidelines, norms and protocols.

The matter of the moral status of the healthcare institution, however, begs the question whether it is the actual institution that has moral duties or is only the individual moral agents that comprise the institution. On the one hand, it has been argued that hospitals can act deliberately and thus can be assigned moral

[13] Although there are potentially relevant normative considerations regarding the types of hospitals and their moral status, for instance whether they are for profit or non-profit, funded by charity donations or religious institutions, public or private, etc., I will not add these considerations to this analysis.

responsibility[14], however, in practical terms, this occurs only nominally. If we consider that any sort of action performed by a healthcare institution, i.e., firing a medical professional unjustly, are decisions made by a governing board, then it seems more plausible to argue that the moral responsibility is held by the members of said board who voted in favor of firing the worker.

However, this argument would face difficulties if, let us say, all the board members resign and leave the hospital. In this case, if we argued that the former board members hold the responsibility and accountability for the unjust firing, then the responsibility would cease to exist the moment the board disbands. However, this is not the case. The new board, as the *de facto* new representative of the values and interests of the hospital as an institution, would have to take over the responsibility for the actions of the previous board, if, for instance, the worker sues the hospital. As such, I argue that healthcare institutions have moral status insofar we accept that there are actions that can be taken in representation of the institution and that are governed by the internal guidelines, norms and protocols. However, this status is nominal and in practical terms, it is necessary to ensure that those guidelines, norms and protocols are fair and that the moral agents that constitute the governing bodies with authority inside the institutions act in accordance with them.

Having clarified this matter, the duties of the institution towards the clinicians must be considered in two senses: first, the duties originated from the normative aims of the healthcare institution as such and of the physicians and other clinicians as practitioners of the medical profession, and second, the duties of the healthcare institution as the employer towards its employees. In the first conception, the duties of healthcare institutions derive from the duties that exist towards the patients to provide a level of health care in accordance with the standards setup by regulatory frameworks (which are guided by the notion of the highest attainable level of care explained in Chap. 4). These duties, as noted above, are normatively oriented toward the rights of patients and include the provision of adequate material resources, opportunities for continuing education, training and supervision of junior staff, and an infrastructure for safe practice.

In the second conception, the duties of healthcare institutions are instead oriented towards the rights of medical professionals as employees and as persons

[14] Richard T. De George, "The Moral Responsibility of the Hospital," *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 7, no. 1 (February 1, 1982): 87–100, https://doi.org/10.1093/jmp/7.1.87.

with an equal claim to the necessary conditions, including healthy work-life balance conditions that have a direct impact on their physical and psychological integrity, and the freedom to act in accordance with personal values. In this sense, implementing ML models in diagnostic processes would have to be evaluated in terms of their utility in fulfilling these moral duties (e.g., does the model reliably transcribe referral orders and reduce the time spent on this task by a certain amount of time per week so that the professional has more time to foster meaningful communication with the patient), or does it promote the well-being of clinicians as employees or persons (e.g., the time saved by the transcription model is used to give clinicians an extra break during a shift).

The moral duties among clinicians are usually inscribed in codes of medical practice, also called codes of ethical conduct. They usually refer to duties to other physicians, health professionals and other staff regarding three main components: a) general rules of behavior that include being respectful, communicating effectively, cordial and not discriminating colleagues, b) rules about proper manners while being consulted by a colleague or assisting him or her with the treatment of a patient[15], and c) rules about reporting circumstances that could impede the clinicians from providing a high standard of care. This includes reporting a colleague if they are suspect of carrying a communicable disease that might infect the patients or other colleagues, or if they are in any sense impaired from providing a satisfactory level of care[16].

While these norms encompass most of the classic scenarios of interaction among clinicians in theory, they cannot account for other important aspects of the relationship between them that also have an impact in, perhaps, more subtle ways like collegiality, friendship and camaraderie. Although this is an under-researched topic in healthcare, there is some initial evidence that positive relationships in clinical workflows have an influence on clinician burnout rates[17] and intentions

[15] These norms require that physicians not interfere with the relationship between the patient and the attending physician, except to protect the patient, including not making comments or suggestions that may cause the attending physician to lose authority in the eyes of the patient, surrogate, or family member.

[16] World Medical Association, "WMA International Code of Medical Ethics" (World Medical Association, April 14, 2023), https://www.wma.net/policies-post/wma-international-code-of-medical-ethics/.

[17] Anthony C. Waddimba et al., "Predictors of Burnout among Physicians and Advanced-Practice Clinicians in Central New York," *Journal of Hospital Administration* 4, no. 6 (August 9, 2015): 26, https://doi.org/10.5430/jha.v4n6p21.

to withdraw from their jobs[18]. Moreover, there is also evidence that healthy interprofessional collaboration among clinicians may improve patient outcomes[19]. The introduction of ML models in clinical practice and workflows must be assessed having these aspects in mind because it is not yet clear how different types of models can disrupt existing interactions and how the benefits of their use are distributed. A lack of understanding of this area could prove detrimental to the adoption of ML models since, as has been remarked at various points throughout this dissertation, some of the challenges of implementing ML models in clinical practice come from the lack of acceptance of clinicians.

The matter of how ML models may change the workplace dynamics in clinical settings is connected to four main questions: first, who are the professionals that are part of the setting in which the ML model will be implemented? For instance, primary or secondary care physicians, allied health professionals (lab technicians, medical assistant, radiographers, medical technologist, etc.), non-clinical staff (secretaries, healthcare managers, etc.), or would it require interprofessional or even intra-professional collaboration?[20]. This has practical implications because, for instance, it could be necessary to establish a new relationship with the technologist or in-house developer in charge of operating and/or maintaining the model (and that would likely be in charge of serving as the liaison between the developing company and developers, and the clinical institution and the clinicians), setup the time and resources to train the staff to use and understand how the model operates (which should also include managing the expectations and concerns of clinicians) and make a plan to introduce the model to the existing workflows.

---

[18] Leah E. Masselink, Shoou-Yih D. Lee, and Thomas R. Konrad, "Workplace Relational Factors and Physicians' Intention to Withdraw from Practice," *Health Care Management Review* 33, no. 2 (June 2008): 178, https://doi.org/10.1097/01.HMR.0000304507.50674.28.

[19] Jacqueline S. Martin et al., "Interprofessional Collaboration among Nurses and Physicians: Making a Difference in Patient Outcome," *Swiss Medical Weekly* 140, no. 1718 (May 8, 2010): 1–12, https://doi.org/10.4414/smw.2010.12648.

[20] Interprofessional collaboration occurs when professionals from different medical professions work together, for instance, a nurse and a physician. Intraprofessional collaboration, on the other side, refers to collaboration among professionals from different disciplines, for example, a cardiologist and an oncologist. The latter type of collaboration is usually more common to diagnose and treat patients with multimorbidity while the former is a normal occurrence in most hospital settings. See: Simon T. de Gans et al., "Effect of Interprofessional and Intraprofessional Clinical Collaboration on Patient Related Outcomes in Multimorbid Older Patients—a Retrospective Cohort Study on the Intensive Collaboration Ward," *BMC Geriatrics* 23, no. 1 (August 26, 2023): 519, https://doi.org/10.1186/s12877-023-04232-2.

Second, what are the expectations of clinicians and health care institutions regarding the implementation of the model? This question includes aspects such as whether there is an institutional requirement to use a particular model or whether clinicians are free to use traditional techniques, how the model is expected to perform, and what to do if there is a disagreement between the results offered by the model and the clinician's medical opinion, i.e., a question of clinician autonomy and liability. Third, what is the structure of the existing clinical workflow (e.g., what are the steps a radiologist follows to review an x-ray and make a diagnosis) and at what points will the model be introduced? For example, at the beginning of the process, before the radiologist looks at the scan, or instead at the end of the review process, after the radiologist has done his or her own review. This is closely related to the fourth question, which asks what role the model should play in the clinical setting. This question concretizes several points made in previous chapters, such as the displacement of jobs in the medical profession, or the potential for ML models to enhance the skills of clinicians, freeing them from repetitive tasks.

Pee et al. conducted a study on the impact of using ML-driven robots in various healthcare settings in a Chinese hospital[21]. They were interested in how they affected people and the relationships between them. Although robotics is not the focus of this paper, the study is valuable for the insights it provides into the practical implications of introducing ML models for medical purposes. The authors identified four ways in which the various ML systems used in this hospital affected the work of clinicians: a) augmentation, the model enhanced the work of clinicians by integrating their knowledge and experience; b) automation, the model reduced the physical effort of a particular task; c) assistance, the model reduced the cognitive effort of a particular task; and d) actuation, the model reduced the workload of clinicians by performing certain tasks autonomously. Each of these potential roles would need to be evaluated against the previously formulated considerations.

5.2.3 Technical Constellation

The third constellation considers the normative questions that arise from the distribution of burdens and benefits resulting from the relationships between the

[21] L. G. Pee, Shan L. Pan, and Lili Cui, "Artificial Intelligence in Healthcare Robots: A Social Informatics Study of Knowledge Embodiment," *Journal of the Association for Information Science and Technology* 70, no. 4 (2019): 351–69, https://doi.org/10.1002/asi.24145.

moral agents associated with the developing companies and corporations that build ML models for diagnostic purposes and the end users of the technologies, who are either direct or indirect users.

On the one hand, direct users are defined as agents who interact directly with the ML devices or software applications and employ their recommendations or results like the clinicians at the healthcare institutions that acquire the ML models, or as consumers like patients using direct-to-consumer applications (see Sect. 1.3) such as telehealth apps or wearable health devices. On the other hand, indirect users are those for whom the devices or software applications are used for. In this case, the indirect users or end-beneficiaries of ML models in diagnostic settings are the patients at healthcare institutions for whom the clinicians use the models. It must be stressed that in this constellation, the models that are considered are exclusively the ones developed by tech companies to offer direct diagnostic

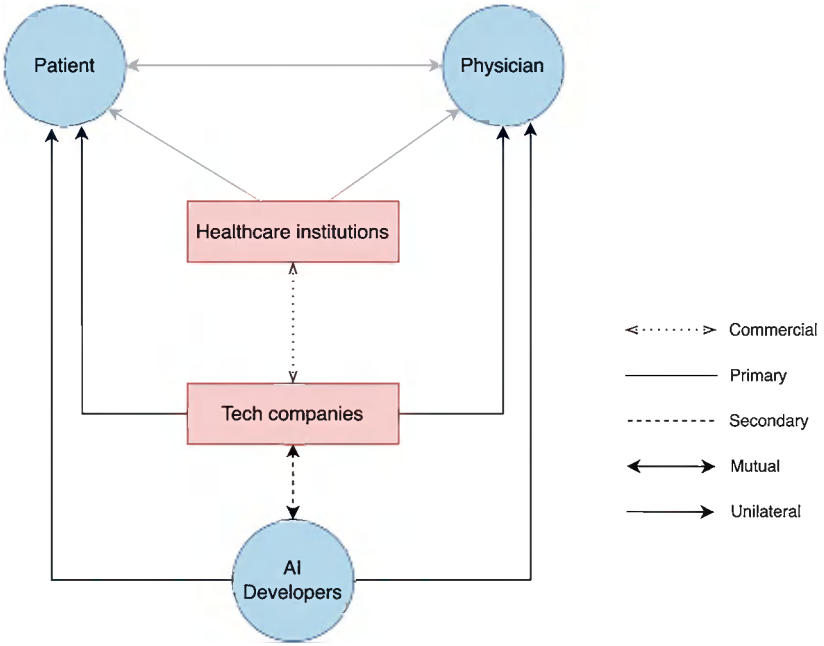

Fig. 5.3 Diagram of the relationships in the technical constellation. The relationships between healthcare institutions, physician and patient are shown in light gray to indicate thar they are not considered in this constellation

support. Aside from the patients and physicians as end users, the other type of relevant moral agents are the AI developers that are employed at tech companies. Furthermore, as seen in Fig. 5.3[22], the technical constellation also takes into account the role that healthcare institutions and technology companies and corporations as healthcare vendors have.

The first relationship considered in this constellation does not follow the conceptualization proposed for the first two, in that it cannot be evaluated in terms of its moral obligations to other moral agents but nevertheless requires analysis because of the normative consequences that flow from the decisions made within it. It is the commercial relationship between the technology company that develops the ML model and the healthcare institutions that purchase it. It is driven primarily by the pursuit of mutual self-interest, where the benefit to the technology company is generally monetary compensation and the benefit to the healthcare institution can be estimated in terms of improved care delivery (e.g., reduced patient waiting time, increased patient retention, improved patient outcomes) and cost reduction (e.g., reduced clinician turnover, improved team coordination, more efficient clinical workflows and resource allocation, etc.)[23]. Although this relationship is not normative, it is conceptualized as the constitutive relationship of the constellation because the other relationships originate from it. Moreover, even if we cannot strictly speak of moral duties on the part of health care institutions and technology companies toward each other, there are still certain duties that have moral significance and should be considered. For example, the technology company has a duty to ensure that the model meets technical standards to ensure the security and privacy of patient data.

The second relationship identified in this constellation is between the technology company and the AI developers as employees. Beyond the standard employer-employee duties governed by workplace regulations, there are special duties related to the value-laden nature of ML model design and development. Companies have a duty not to interfere with the freedom of AI developers to make good moral decisions at the various points in the ML development pipeline

[22] The relationships between healthcare institutions, physician and patient are shown in gray to indicate thar they are not considered in this constellation.

[23] It should be noted that the benefits for healthcare institutions will probably not be immediate and moreover, they would be contingent on the performance of the model in the actual clinical setting and on how successful its integration in the existing clinical practices and workflows goes, as it has been argued throughout this section. This is an important clarification because without clinical evidence obtained from randomized clinical trials or longitudinal studies, it is difficult to have sufficient certainty that the models *will* provide the expected benefit for the healthcare institutions.

where it is necessary (e.g., when cleaning the dataset before the training phase to ensure that it does not contain negative discriminatory biases), and not to punish them for taking action to call out potential or actual unethical practices that may lead to harmful consequences for users. Companies also have positive obligations to provide AI developers with the necessary tools (ethical training, relevant literature, access to experts in ethical issues, etc.) to understand how personal and societal values and interests can be surreptitiously introduced into technical decisions, and how to take action to address relevant concerns.

Similarly, AI developers have special obligations to technology companies. First, they should follow existing internal protocols for raising concerns about potentially unethical practices before resorting to other legitimate avenues such as whistleblowing, and second, where appropriate, they should follow the company's internal intellectual property policies to protect trade secrets, code, or copyrighted content. Of course, illegal activities or sufficiently serious concerns about matters of public interest (e.g., a leak of patient data or a cyber-attack that compromises the integrity of the model) should be reported and would not necessarily constitute a violation of the company's intellectual property rights but would be a permissible infringement of them. Thus, this relationship is considered reciprocal and is of secondary order in the classification of constellations, being subordinate to the previous one.

The conceptualization of the next relationships requires several normative clarifications. As proposed, the primary relationship in this constellation argues that technology companies have a position of some moral significance towards the users of the models they develop. In addition, two types of rights holders have been identified as a result of the commercial relationship: patients and clinicians as end users or beneficiaries. Through the contractual agreement with the healthcare institutions and the regulatory frameworks stemming from regulatory entities, tech companies are obliged to ensure the protection of rights like the right of patients to the privacy of their health data and the safety of the users that manipulate the ML models. However, regarding other types of moral duties towards the end users and beneficiaries, it is unclear if it is the tech company or the AI developers who should carry the responsibility.

On the one hand, tech companies do not have fiduciary duties to patients or clinicians because, in their role as health care providers, their relationship is with the health care institution, not with clinicians or patients directly, and is commercial in nature. In the case of patients as consumers of health applications or wearable devices, technology companies cannot owe fiduciary duties because they are not required to act in the consumers' best interests. When consumers purchase a device or software service, they accept a set of terms and conditions that

typically place the responsibility for the use of the product on them and release the company from potential liability (there are cases where liability claims can be accepted if the product fails to comply with existing regulatory requirements or if it can be shown that the product was intentionally designed to be misleading, coercive or false, but this is generally not the case).

On the other hand, the AI developers are already in a mutually binding relationship with the tech companies as employees, and it could be argued that they cannot be responsible for the potential violation or infringement of the rights of users or consumers, since they are first governed by the norms and guidelines of the company with which they have signed a contract. However, the products are actually designed and developed by the various technical professionals working in the companies. They have the domain expertise, and they make decisions with moral significance at different stages of the ML pipeline. The problem is that, unlike clinicians who are educated throughout their careers about the ethical challenges of their profession and are given the tools to guide their actions in scenarios of ethical uncertainty, there is very little structured training for computer scientists about the ethical and societal implications of their decisions in developing AI models.

There are two main reasons that may help explain this problem. First, the codes of ethics that most technology companies have follow an approach based on the operationalization of high-end principles. However, as discussed in detail in Sect. 2.2.1, these approaches are ill-suited to providing AI developers with actionable solutions due to the inherent technical complexity of coding principles and the general lack of consensus about what to do in scenarios where principles and interests conflict. Second, there is a phenomenon characterized by a disconnect between AI developers and the consequences of their technical choices in ML development. In 1990, building on Milgram's work on obedience to authority figures, Batya Friedman reflected on the distancing effect of computer-mediated actions on the distribution of responsibility in digital environments. She argued that there is a characteristic distance to actions performed through a computer that affects the agent's ability to understand his responsibility and accountability to the recipients of the action[24]. Similarly, Widder and Nafus theorize that the modular nature of ML development, in which large projects are broken down into more manageable tasks, has an isolating effect on AI developers, causing them

[24] Batya Friedman, "Moral Responsibility and Computer Technology," April 1990, https://eric.ed.gov/?id=ED321737, 3–4.

to perceive themselves as parts of the development pipeline rather than as agents with moral relevance[25].

It seems that there are conceptual and practical difficulties in trying to place all the responsibility for the potential risks of ML models on either the technology companies or the AI developers. However, they should not distract from the fact that these potential risks are real and need to be addressed. There are two assumptions that can be made in this situation, and that help find a way to conceptualize the next relevant normative relationships in this constellation. First, that there must be *some* sort of moral obligation regarding the distribution of burdens and benefits that can be placed on the tech companies and AI developers that produced the ML model. In their mutual relationship, tech companies and AI developers work together to bring together material and non-material resources and domain expertise to create a product of high complexity, which in most cases is understandable only to them (i.e., even in the case of black-box algorithms, they understand the architectural design of an ML model). In other words, they are in an epistemic position about the functioning of the model that is unique to them. Even if regulators had their own in-house AI experts, intellectual property barriers may prevent them from knowing all the technical nuances of the code that makes up the model. Given this level of knowledge about their product and the fact that no one else has it, they ought to have some responsibility for its performance in clinical settings.

Second, it can also be argued that any company or agent developing a product that is intended to be used in a clinical setting or for a medical purpose must know that this implies that the model is dealing with people's health, and in any scenario, positive or negative, it will have an impact on it. The consequences of failure are therefore not simply a matter of customer satisfaction, measuring metrics or benchmarks, or customer retention and attrition. In the scenarios we are discussing with the implementation of ML models that provide direct diagnostic recommendations, we are dealing with the potential impact on a person's fundamental right to health or even life. If these facts are not already clear, they need to be made clear at the outset of the ideation process. There is no compelling reason why AI developers or involved non-technical staff should not be made aware of this key information and all that it implies.

[25] David Gray Widder and Dawn Nafus, "Dislocated Accountabilities in the 'AI Supply Chain': Modularity and Developers' Notions of Responsibility," *Big Data & Society* 10, no. 1 (January 1, 2023): 20539517231177620, https://doi.org/10.1177/20539517231177620.

A final assumption is that technology companies and AI developers have already benefited from patients and clinicians and therefore have corresponding obligations to them. The argument is based on the fact that ML models have been made possible in part by the availability of large clinical datasets. The data in these datasets come from patients who have already been diagnosed and treated, and from the efforts of clinicians and non-clinical workers to transcribe their health data into digital form. As explained in Sect. 1.1, this process took decades to complete and perfect. Therefore, it could be argued that since the data used to develop the model required the contributions of these actors, companies and AI developers have a corresponding obligation to use it responsibly, so that it benefits the actual patients and clinicians who will use or be affected by the model.

This argument could also be expanded to consider that if patients and clinicians do indeed receive a fair share of the benefits of safe and efficient ML models in healthcare, there will be more health data that could be used to create better datasets. In turn, these datasets could be used in the future by tech companies and AI development, benefiting them as well as future patients and clinicians. This assumption, in short, is based onto the reasoning that if tech companies and AI developers take on moral duties that ensure responsible practices in the development process of ML models and the consequences of the potential risks posed by the models they produce, they will continue to receive a share of the benefits in the future.

Given these arguments, I propose to conceptualize the third and fourth relationships in this constellation as of secondary order, but with parallel obligations that both technology companies and AI developers have toward clinicians as end-users/patients as beneficiaries and patients as consumers, respectively. These relationships are unilateral in nature and constitute the first-order relationships of the constellation, although in practice they derive from the commercial relationship discussed earlier. The reason for conceptualizing them as first-order relationships is that they have the strongest normative weight in the constellation in terms of assessing risks and benefits. The notion of parallel duties aims to allocate moral responsibility and accountability in a way that is not only normatively justifiable, but also practically feasible.

Thus, in addition to the duties to AI developers discussed above, which have a direct impact on the developer's ability to make moral decisions in developing the model, technology companies have a general duty to provide end users and beneficiaries with robust internal frameworks that ensure, to the extent possible,

that the models are safe to use and have been developed in accordance with existing regulations. They also bear responsibility in cases where the model harms a user or beneficiary, and the malfunction or error is technical in nature. On the other hand, AI developers have an obligation to interact meaningfully with end users and beneficiaries to incorporate their preferences, needs, and insights at points in the development pipeline where it is possible or necessary, to develop their own professional deontology with a focus on their moral agency, and to take the initiative to collaborate with other developers, managers, and technical staff to decide on the best course of action from both a technical and a normative perspective.

5.2.4 Regulatory Constellation

The fourth and final constellation formulated in this dissertation is the regulatory constellation. Here I consider the relationships between the regulatory bodies that formulate policies (public organization or government agency), legislation and governance structures, the legal entities (institutions, organizations, companies and corporations) that are involved in the processes of designing, developing and/or implementing ML models in diagnostic contexts, and civil society. Compared to the previous three constellations, this constellation does not deal with individual moral agents, and the moral duties related to the assessment of potential risks and benefits can be seen as broader and, in some relationships, as actions and consequences of actions with moral relevance rather than moral duties in the strict sense. The normative work involved in developing this constellation links the fields of public policy, AI governance, and global health, and is largely left to future research, as it is beyond the scope of this dissertation. In this subsection, however, I will attempt to at least outline the main relational structure and normative considerations that make up the constellation.

To begin with the conceptualization according to the level, the relationship of primary order in this constellation is identified between the regulatory bodies and civil society (see Fig. 5.4) since this constellation's normative focus are the potential risks and benefits of ML models in diagnosis in regard to civil society. The first question that arises here is likely what constitutes these regulatory bodies and the second, who belongs to "civil society". The regulatory bodies constitute the entities that create binding legislations and regulatory frameworks that aim to secure the interests and rights of the public, i.e., civil society. They

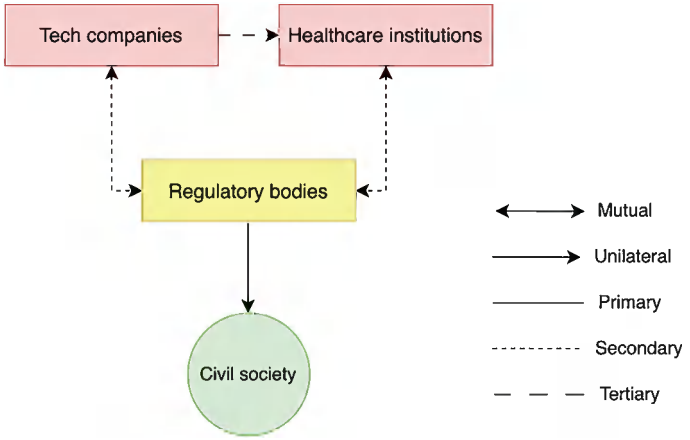

Fig. 5.4 Diagram of the relationships in the regulatory constellation. Although there are no moral agents strictly speaking, the institutions, companies and bodies have duties of moral relevance towards agents, who integrate civil society

vary depending on the country where the evaluation of the implementation takes place and can exist at different levels in the governmental structure. In the US, for example, the constellation would take into account the regulatory frameworks of the FDA and HIPAA, as both regulate different aspects of ML models and impact different rights of civil society. Similarly, in the European Union, such ML models for medical purposes will need to be approved by agencies such as the EMA, and in some cases, it would also be within the purview of the GDPR and the AI Act to intervene.

Regarding the notion of civil society, although some definitions include non-governmental organizations and social movements that seek to advance the interests of individual citizens, I take the conceptualization that refers to the space between the private and public spheres where the discussion of rights and values of society takes place. In other words, the notion of civil society aims to represent citizens as a group in matters of public interest. The two areas of interest in this constellation with respect to the evaluation of ML models in diagnostics are public health and environmental concerns. The responsibilities of regulators to civil society can be summarized in two main groups. First, to ensure that the development and implementation of ML models in clinical settings do not negatively affect the rights of civil society, and second, to provide actual and reliable

information about their functioning and use so that the risks can be avoided, and the benefits reaped.

The next set of relationships are secondary and mutual in nature. They form between the regulatory bodies and the technology companies that develop the models and the healthcare institutions that adopt them. On the one hand, technology companies and healthcare institutions have a duty to comply with the requirements established by regulatory bodies for two main reasons. First, they intervene directly in either the development process or the integration of the models in clinical settings; and second, they profit -or aim to- from the integration. On the other hand, regulatory bodies have a duty to consider and balance the interests of these entities in the regulatory framework.

Although it could be argued that they do not have the same normative importance as the interests of civil society, the commercial and professional interests of technology companies and health care institutions have an impact on the rights of individual rights holders associated with them. For example, if certain policies severely limit a company's ability to stay competitive in the marketplace, this could lead to bankruptcy and layoffs. For individual right-holders, this would mean the loss of their livelihood -and possibly that of their family-, but it could also pose a risk to society, if too many companies face the same situation, and the economic stability of the country is compromised.

The final type of relationships in this constellation are between the governance structures of the organizations that participate in the implementation process, namely, the healthcare institution and the technology company. In the first place, it should be made clear that the subject of governance and AI is intricate and has a vast body of literature that is not accounted for in this subsection. However, since it is necessary to clarify what is meant here by the notion of governance, I take the definition proposed by Mäntymäki and colleagues: "AI governance is a system of rules, practices, processes, and technological tools that are employed to ensure an organization's use of AI technologies aligns with the organization's strategies, objectives, and values; fulfills legal requirements; and meets principles of ethical AI followed by the organization"[26]. This definition focuses on the aspects of governance that are concerned with the tech companies and the implementation of their products. However, I expand on it to include the normative challenges that arise from the tension between governance structures. They can conflict with each other because they are designed within organizations with the normative aims of their members in sight. Often these governance structures do not account

[26] Matti Mäntymäki et al., "Defining Organizational AI Governance," *AI and Ethics* 2, no. 4 (November 1, 2022): 604, https://doi.org/10.1007/s43681-022-00143-x.

for the overlap between their aims and the aims of the other organization with which they want to form a commercial relationship, and this can cause hurdles for the implementation.

The technology companies must consider that the healthcare organizations that adopt the ML models have, first, their own internal governance structures and, second, that these structures are also linked to governmental and, in some cases, supranational governance frameworks. In this sense, most of the responsibility for the governance tension lies with the technology company, since it is the product, the ML model, that has to adapt to the existing clinical conditions. For example, if a company is interested in commercializing an ML model in Germany, the executives must be aware of the regulations, not only at the local level, but at the level of state and federal regulations, and then at the level of European Union regulations. Going back to the definition, Mäntymäki and colleagues emphasize that the main goal of a governance structure is to ensure that there is alignment between the use of AI and the organization's goals and values. I argue that this is only one aspect of a relational governance framework, which is centered on the normative responsibilities of institutions, organizations, corporations, and businesses to civil society. While these normative responsibilities bind these entities to the interests and rights of civil society, they are mediated by the regulatory bodies that design and deploy the regulatory and legal frameworks that guide governance structures.

5.3 Analysis of Normative Aims

The analysis conducted in each constellation helped to make clear that each group of relevant moral agents identified (AI developers, clinicians, and patients) intervenes or engages with the ML tools at different points in their design, development, and implementation processes. These points of engagement are important because they are influenced by the tension between normative goals and interests of other kinds that may be at play, such as economic interests. Chap. 4 briefly touched on the issue of normative goals surrounding the implementation of ML in diagnostics. In this section, I try to provide more clarity about what else can be said about the normative goals that exist and emerge at this intersection, and what tensions might exist between these goals.

First, however, it is necessary to outline what I mean by normative goals. Generally speaking, the notion of normative aims at the intersection of ethics, AI, and healthcare answers the question of what *ought* to be done with respect to the ML model. However, as can be surmised, there is not just one answer

to this question. In fact, not only are there different normative aims about how ML models should be *employed* in healthcare, but there are other aims that have emerged at the other different points of engagement that may also have moral significance. Identifying these goals brings together the work done in previous sections and helps to provide a clearer picture of the potential conflicts regarding the distribution of burdens, i.e. potential risks, and benefits of ML models.

Although there are normative aims that could be identified at the organizational level, i.e., healthcare institutions or tech companies in this case, I will limit the analysis to the moral agents as persons assuming that AI developers and clinicians, to a degree, already embody the normative aims of their respective organizations as they tend to correspond to their professional roles. There are three types of normative aims that I find of chief importance: First, to appropriately design and develop ML models for healthcare purposes; second, to appropriately use ML models in clinical settings, and third, to appropriately engage with ML models. These aims correspond to agents at the technical level, i.e., AI developers; at the clinical level, i.e., medical practitioners; and at the consumer or end beneficiary level, i.e., the patients. The key aspect to all three questions is the notion of "appropriate", which implies the normative consideration of what each agent or group of agents ought to do at their particular engagement points with the ML tool.

However, it must be remembered, as it has been argued throughout this dissertation, that the question for the ought necessarily implies a *can*. In other words, the normative aims must be preceded by practical feasibility. For example, if the normative aim that guides the use of a model is to improve the patients' outcomes, the model must be built with the technical capabilities to do so. However, in order to go about this, there are several normative tasks that need to be carried out. For instance, it must be defined what concrete actions are required to achieve this and also a criterion to determine that the model fulfilled the normative aim, i.e., a clinical benchmark.

5.3.1 Normative Aims at the Technical Point of Engagement

Normative aims in the design and development of ML are generally framed in three ways: first, in terms of external regulatory compliance. Second, in terms of an internal governance structure, often guided by a principled approach. Third, in terms of adherence to the best practice standards of computer science, data science, etc. In the first case, models are considered to be appropriately designed

and developed if they meet the requirements set by a country's regulatory authority. This is because it is assumed that these regulations are formulated and implemented by competent authorities that seek to ensure the protection of users' rights. In the second way, high-end principles such as fairness, transparency or accountability dictate normative maxims such as "ML should avoid unjust bias" or "ML should be transparent". To some extent, the normative aim of following these principles points in the general direction of fulfilling moral duties to users and end beneficiaries, but again they fail to concretize what agents should do in the first place. Finally, in the third way, the normative aims of ML design are framed in terms of adherence to existing best practices of development from a purely technical perspective. For example, the establishment of regular code quality checks or model validation phases are professional practices that help to detect early errors or biases embedded in the model, which would ultimately be reflected in the model's performance in real clinical practice.

Each of these approaches has difficulties. Regulatory compliance does not necessarily ensure that the rights of users and end beneficiaries are *de facto* protected, as there may be cases or regulatory loopholes that do not cover certain applications. This is particularly problematic when dealing with emerging technologies such as AI, as regulators may need time to develop a solid understanding of how the models work and what actual risks arise from their use, and therefore regulation may lag behind and not match the speed of development and deployment at the commercial level. Adherence to high-level principles, as argued, faces the triple complication that they are difficult to translate into an actionable course of action, suffer from a lack of conceptual clarity and consensus, and are non-binding in nature, meaning that actors may be unsure of what to do, disagree about what the concepts imply, or decide not to act in accordance with the principles. Finally, adherence to best practices also faces two difficulties. First, given the emerging nature of AI development as a professional field, there are still no widely agreed-upon standards of practice that are recognized by all companies and professional organizations, which leads to the second problem that these standards are also non-binding in nature, which can lead agents to disregard them when there are competing interests, such as meeting a deadline, which could be delayed if the agent has to perform a quality test.

From a relational, rights-based approach, the normative aims at this point of engagement must be framed with a focus on the end user's and beneficiaries' rights at the center. This includes recognizing the limitations of the technical solutions that AI offers for complex issues in diagnostic processes (which implies that companies should avoid marketing AI solutions in misleading or unfounded ways) and acknowledging the need to integrate clinicians and patients

as main stakeholders. Concrete actions include implementing participatory design approaches in which patients and physicians take part, the development of a collaborative governance strategy based on the rights of stakeholders to resolve tensions or conflicts, and the development of standardized best practices that comply with regulatory frameworks.

5.3.2 Normative Aims at the Clinical Point of Engagement

Although the analysis of normative aims in health care may at first glance seem easier to approach, given the intrinsically moral nature of medical practice and the long-established normative discussion in medical ethics and bioethics research, it in fact poses its own set of challenges. First, there are many normative aims that routinely coexist and sometimes conflict. Although there are some general normative aims that can be used as a guide, such as providing the patient with the best possible clinical outcome or arriving at the correct diagnosis for a set of symptoms, the aims that guide decisions in the clinical encounter depend on factors such as what the clinician can do in terms of available resources and experience, what the patient's personal goals are, and even what can be done at all in the particular situation. For example, if a patient presents with a tumor that has metastasized and is considered to be beyond the reasonable threshold for curative treatment, the normative goal is likely to be to manage the patient's discomfort and improve their quality of life as much as possible and the specific actions taken will depend on the particular aspects mentioned.

The appropriate use of ML models in the diagnostic process in its broad conception has different implications depending on the type and purpose of the model, i.e., the technical possibilities constitute the *can* in the *ought* here. These aspects can be divided into two distinct factors: what the purpose for which the model is being used, and what the intended role of the model within the clinical setting is.

The first factor seeks to clarify who are the direct beneficiaries of the use of the model in terms of practical utility. In this case, we would speak of a) clinicians as beneficiaries, for example, of models designed for laboratory testing and pathology, such as blood cell counts, administrative tools that automatically set up alerts and reminders for abnormal results, or CDSS that help clinicians manage routine clinical tasks, such as blood glucose monitoring in the ICU or drug-drug interactions; and b) patients as direct beneficiaries, such as models

like diagnostic CDSS that directly support the process and provide clinicians with recommendations or alternative analyzes, medical imaging review tools that generate a diagnostic result, or patient-facing decision support systems that give patients access to their EHR and even integrate with their wearables or mobile phones.

The second factor is to clarify what kind of effort or engagement will be required from clinicians. For instance, will the model automate a clinical or an administrative process, will it be used as a "second opinion" or as supporting evidence in a clinical task like reviewing an MRI, or will it be used as an assistant at the point of care, for instance, transcribing speech to text from a summary made by a clinician that is added to the EHR or recording consultations that perhaps can be used as part of a training module for trainees.

5.3.3 Normative Aims at the Patient Point of Engagement

The identification of normative aims at this point of engagement seeks to open a space for the analysis of the role of the patients as an active participant of the diagnostic process but also as consumers and receivers of potential risks that derive from the implementation of ML technologies. In most of the literature in this field, patients are considered as the primary subjects of protection as rights holders. However, there is little research about the role patients *should* play and what potential moral duties they may have given all the new possibilities of accessing information and resources to manage their health.

The normative aim to engage appropriately with ML models urges us to consider that there may be reasonable expectations that should be placed on patients regarding their use of the new types of information available regarding the ethical use of ML models designed as telehealth apps or wearable health devices. A normative consideration of appropriate engagement with these tools includes how patients integrate them into their interactions with clinicians and into their own care journeys. In addition, the use of such tools should not discourage patients from contacting their PCP and attending diagnostic appointments, and in the case of a tool that is integrated into the diagnostic process under the supervision of the primary care physician, should not be used to overburden the physician, e.g., by being used to contact the physician outside of appointment times when there is no emergency. It could also include aspects such as gaining knowledge about how these devices work and what can be expected of them regarding their risks.

5.4 Normative Evaluation of the Distribution of Potential Risks and Benefits

The work done so far in this dissertation has aimed to provide an analysis of the relevant normative aspects arising from the implementation of ML models in medical diagnosis with a solid philosophical foundation, but also with an emphasis on the importance of attempting to contribute to an interdisciplinary approach. As argued in Chap. 2, interdisciplinarity requires building "bridges" between disciplines and designing methods that allow for the mutual exchange of knowledge in a way that can be integrated and make the most of in each discipline. Thus, in this section I aim to address the final concern of this work, namely, its contribution to the efforts to build effective interdisciplinarity, in other words, its practical utility from the perspective of applied ethics.

Although the value of the work in moral philosophy and normative applied ethics is not merely instrumental, as there is intrinsic value in providing conceptual clarity and justifications for action in a systematic way, one of the tasks of applied ethics is in part to concretize the often-complex philosophical arguments into more accessible and practical discussions and recommendations that other disciplines and fields can use to develop their own deontology or regulatory efforts. Therefore, I seek to go a step further and present a theoretical exercise of the use of the ecosystem of constellation framework and the relational, rights-based approach to identify normative considerations in a hypothetical clinical encounter.

In this exercise, I will focus on the normative considerations starting at the clinical constellation with the clinical encounter. This entails that the assessment is not done chronologically because this would require starting with the considerations at the point of design and development of the ML model, which in the exercise are already acquired by the healthcare institution and available in practice. This exercise also assumes that the regulatory landscape matches the present one, in terms of the processes and guidelines for the approval of AI-driven medical devices by regulatory agencies, such as the FDA and EMA.

This exercise is not intended to be a checklist of requirements to be followed or a set of instructions. Rather, it is meant as a guide or blueprint to the ethical reflection process, highlighting the relevant normative issues and considerations at different points in the diagnostic process. It should also be noted that this exercise makes assumptions based on hypothetical scenarios and, therefore, it will not be able to include all aspects or situations that may arise in particular clinical scenarios.

The constellation analysis had two main normative aims: first, to identify the relevant moral agents within each constellation in the ecosystem model and the relevant moral relations between them. In the analysis of the normative aims in the previous section, I discussed the implications of the different aims that various relevant moral agents could have at the different points of engagement with the ML models, and second, what normative implications could be identified according to the purpose of the models and the role they would have -either because of the technical design or because it was assigned by the clinicians or healthcare institutions- in the clinical setting. Although I have outlined the general considerations for each question, there are other relevant derivative questions and aspects that arise in particular contexts.

5.4.1 Normative Considerations About the Patient

Since this exercise starts at the clinical encounter, it is located within the clinical constellation and, as such, the main moral agents are the patient and the physician. I start by focusing on the role of the patient. The first derivative normative matter that requires consideration is whether the patient is able to make choices of medical and moral significance unaided. There are three potential responses which have different implications: First, that the patient is indeed able to do so; second, that the patient is not able to do so; and third, that the patient is able to do so to an extent but requires a degree of assistance.

In the first scenario, the patient enters the relationship with the physician as a full-fledged moral agent and has moral duties of his own towards the physician, and potentially towards relevant relationships he has with a spouse and dependents like children or older family members under his care. The moral duties towards the physician at this point of the clinical encounter can be summarized in being truthful and forthcoming about the relevant information regarding the ailment that made the patient seek care in the first place, including past medical history not known by the physician and any habits that might have an impact on the patients' health.

In addition, according to the relational perspective, the patient is not an isolated individual and as such has separate moral duties to others in his or her close circle. Although in many scenarios these moral duties do not conflict or even overlap with the duties within the clinical constellation, i.e. the patient requires minimal care, there are other scenarios where circumstances create a tension between duties as a patient and as a parent, spouse or caregiver. In these scenarios, the duties of a patient at this point in the encounter include, for example,

being proactive in seeking care when needed and being willing to make an effort to follow instructions or recommendations agreed in partnership with the clinician. For example, making appointments for tests, following up on suggestions to change harmful habits or develop healthy ones.

There is a further set of duties that could be potentially assigned to the patient, namely, moral duties towards oneself, that would encompass doing what one can to avoid getting sick, i.e., living a healthy lifestyle and avoiding putting oneself in harm's way unnecessarily and recklessly, which could help prevent the imposition of certain burdens on others and on the healthcare system. While there are good reasons to consider these types of duties, it is necessary to consider that this could lead to the phenomenon of medical responsibilisation previously mentioned. Following an argument from Sven Ove Hansson, there ought to be a differentiation between task responsibility and blame responsibility when making patients responsible for certain unwanted health outcomes. He argues that while patients should be encouraged to take responsibility for the actions, they can take to promote their own health, they should not be blameworthy if they fail (e.g., they do not manage to stop smoking)[27]. Medical responsibilisation not only ignores systemic factors outside of the control of the patient but also affects the relationship between him and the physician.

In the second scenario, i.e. that the patient is indeed incapable of making moral and medical decisions unaided, he or she cannot enter the relationship with the doctor as a fully-fledged moral agent. In this case, the relationship is no longer mutual, but situationally unilateral, and the physician still has the same moral duties towards the patient. The first aspect to consider here is what makes the patient incapable of making moral decisions. It may be that the patient is an infant, a young child or an adolescent below the legal age of consent; it may be that the patient suffers from a serious cognitive disorder which, in practical and legal terms, requires that decisions be taken by a proxy; or it may be that the patient is in a state of unconsciousness. In such scenarios, certain responsibilities that belong to the patient are transferred to the proxy, who has been chosen by the patient—and in some cases by a court—to have the authority to make decisions when the patient lacks the capacity to do so. Although the physician's primary duty remains to the patient, a relationship of secondary order between the physician and this third party is necessary for the management of the diagnostic

[27] Sven Ove Hansson, "The Ethics of Making Patients Responsible," *Cambridge Quarterly of Healthcare Ethics* 27, no. 1 (January 2018): 89–91, https://doi.org/10.1017/S0963180117000421.

process. It is therefore necessary for the physician to establish open and accessible channels of communication and to be willing to work with the third party on behalf of the patient. Looking at this context from a relational approach, it would be necessary to take into account the concerns and insights of the proxy, carer or legal guardian and involve them in the decision-making process, as the responsibility for ensuring the patient's compliance with medical recommendations would fall largely, if not exclusively, on them.

Finally, the third scenario occurs when the patient has the cognitive capacity to make moral and medical decisions but is not completely independent and relies on the assistance of family or a carer to carry out ordinary tasks. In this setting, patients who have significantly reduced mobility, a debilitating or degenerative condition that requires constant accompaniment, or mild cognitive impairment, and other similar conditions, are included. As the patient retains the capacity to make autonomous decisions, there is no reason to involve a third party in the diagnostic process. However, the relationship with the family member or carer who provides the assistance acquires normative significance in a different way than in the previous scenario. The success of the medical recommendation, i.e. the patient's adherence to it, may depend to some extent on the support of this person. If they commit themselves to supporting the patient, they acquire a duty of care based, inter alia, on the moral significance of a promise.

It is important to remember, however, that the third pillar of the proposed rights-based approach (see Table 5.1) stipulates that the assistance someone provides must not be at the expense of their own rights, i.e. at no comparable cost to themselves, as this would place the rights of the patient above those of the carer or family member. Considering this, I argue that the patient has a moral duty towards this person or persons to help himself as much as possible to avoid placing an excessive burden on her. Of course, situations like these in reality cannot be resolved simply in terms of normative imposition of burdens. Instead, it would require that both parties, patients and carers or support family members, mutually embrace their vulnerability and communicate their needs. It would also be a continuous, open-ended process of negotiating and compromising certain aspects of the tasks of care. The point I want to make here is that carers or supporters should be considered equal right-holders at any given moment. The tasks of care are known to be significantly demanding at physical and mental levels and thus must not be just dismissed in favor of the needs of the patients.

5.4.2 Normative Considerations About the Physician

The other key moral agent in the clinical encounter is the physician. There are three derivative normative aspects to discuss regarding the physician's role in the clinical encounter. First, the level of care the physician provides; second, the level of experience the physician has; and third, the level of digital literacy and confidence regarding the use of AI systems in general.

In the first aspect, the level of care has ramifications for the tasks physicians are responsible for and the type of relationship they have with the patient and/or proxy. Furthermore, the normative implications of the use of ML models would differ in some respects because the level of care allows for the use of certain models and not others. The two levels of care considered relevant for this exercise are primary and secondary care, because it is at these levels of care that ML models with diagnostic purposes have the most for application. Primary care, as defined in the clinical constellation, is generally the first point of contact with the patient. A PCP's tasks include conducting physical exams, entering or updating patient information in the patient's EHR or the institution's own system, counseling the patient and/or the patient's representative about the patient's symptoms, and working with them to determine a course of action. However, even if the patient has already initiated the process and this is not the patient's first clinical encounter, the PCP is still the health care professional with whom the patient would partner for a significant portion of clinical encounters, for example, regarding the patient's health progress and the management of certain chronic conditions, such as asthma or diabetes, after the patient has gone through the diagnostic process with a secondary care physician.

Furthermore, if the patient needs to be referred to a secondary care specialist, the PCP has a role of normative relevance that could be classed as gatekeeper in that the physician must balance not only the interests or preferences of the patient but must also consider factors like the responsible allocation of resources. In other words, and taking the analysis done in the operative constellation, the decision of a PCP to, for example, order a testing procedure or to remit the patient to secondary care, can have an impact on the rights of other patients and has practical consequences for the clinical workflow, which may also affect indirectly the ability of other clinicians to fulfill their own moral duties to their own patients. The tasks of the secondary care physician include the authorization to conduct highly specialized testing on patient and monitoring of the results of it, provide acute care, i.e., short-term but specialized care for a severe injury, an

episode of an illness, or during the recovery period after a surgical procedure[28], and manage the transition of patients from inpatient to outpatient care.

The duties of primary and secondary care physicians to patients are similar in nature. Broadly construed, physicians have a duty to diagnose and treat patients. The normative aim of such a duty corresponds to the patient's right to health care, understood as the highest attainable level of care (see Sect. 3.3.1). However, to realize this duty, there are simpler, actionable duties, such as the duty to communicate relevant information to the patient and/or proxy in an appropriate and timely manner, the duty to respect's the patient's right to confidentiality, the duty to provide care without discriminating against the patient on the basis of age, gender, sexuality, race, socioeconomic situation, religious beliefs, or legal status. Other moral obligations include listening to the patient's concerns and recognizing the patient's values and preferences as valid, being respectful and considerate of the patient's vulnerability, working with the patients to manage their symptoms, conditions or diseases in a way that is meaningful to them, using available resources for the benefit of the patient instead of responding to pressure from external player's interests such as pharmaceutical or technology companies.[29], keeping professional skills and knowledge updated and acknowledge the limits of their own experience and capabilities.

Physicians also have secondary order duties to their other patients to manage their schedules so that every patient receives care in a fair and equitable manner. Of course, this needs to be tempered by the particular circumstances and needs of the patients. However, as argued in the previous subsection, in the same line that patients ought to be mindful of the fact that physicians are generally in charge of many patients at any given time and their rights are equally important as their own, physicians must be aware of the extent of their capacity to take on patients

[28] Jon Mark Hirshon et al., "Health Systems and Services: The Role of Acute Care," *Bulletin of the World Health Organization* 91 (May 2013): 386–88, https://doi.org/10.2471/BLT.12.112664.

[29] This point highlights the emerging concern of clinical experts that the development of AI models could lead to a situation similar to that of pharmaceutical companies, where the incentives to use certain drugs were disproportionate and led to irresponsible prescribing protocols. This has had an impact on the rights of individual patients, many of whom ended up becoming severely addicted to painkillers and has also had an impact on public health concerns, such as the opioid crisis in the US. For more about this, see: Thomas R. Frieden and Debra Houry, "Reducing the Risks of Relief—The CDC Opioid-Prescribing Guideline," *New England Journal of Medicine* 374, no. 16 (April 21, 2016): 1501–4, https://doi.org/10.1056/NEJMp1515917.

and try to manage time so that no patient is left waiting excessively or has to be rescheduled[30].

Physicians acquire certain moral obligations to other physicians and clinical staff because of the need to work in teams to provide patient care and the relational nature of their work. In clinical settings, physicians rely on nursing staff, junior clinical staff, and administrative staff to perform clinical and non-clinical tasks that influence the diagnostic process. As such, their responsibilities include using available resources responsibly, as noted above; recording and updating the patient's health information appropriately and in a timely manner so that other clinicians have access to it when needed to advance the diagnostic process; communicating effectively and respectfully with other clinicians; and respecting the schedules of other clinicians, i.e., not make excessive requests for other clinicians without considering their own workload.

Regarding the rights of clinicians, we must consider this from the relational approach. Like any other person, clinicians have a right to the necessary conditions to lead their lives (first pillar), which means the right to physical and psychological integrity and the right to freedom. Their concrete rights in the analysis of both clinical and operative constellations are to have a workload that does not risk their own health in both physical and mental senses. In practical terms, the general well-being of the clinicians has a direct impact on their ability to fulfill their own moral duties to the patients and other clinicians, has an impact on their institutions and on securing interests of society at the level of public health, since they are essential workers. However, the rights of clinicians should not be justified only in this instrumental way. As persons, they are ends in themselves and have the same claim as any other person to physical and psychological integrity. Although they accept to take on a considerable burden when they become medical professionals and take the Hippocratic Oath, this does not entail that their fundamental rights are stripped from them, nor does it mean that there is no limit to the burden that can be imposed on them.

The second aspect, namely the level of experience of physicians, may have an impact on the risk of being deskilled as a result of automation bias and other related phenomena derived from the use of ML models (see Sect. 4.2.4), which poses a risk to the right to health care of patients and has normative relevance for

[30] It is important to emphasize that while physicians may have a moral obligation to provide equal care to their patients, it is a well-known problem that health care institutions have a practice of requiring physicians to work too long shifts and to be responsible for a too great number of patients, which reduces their time to provide care, significantly affecting the quality of care and increasing the risk of medical errors. This is a systemic problem that needs to be addressed at the policy and institutional levels.

collaboration between clinicians. For example, although the use of ML models in radiology (the field of medicine where AI has had the greatest impact to date) poses a risk of automation bias for professionals regardless of their expertise, it has been found that less experienced radiologists have a significantly higher risk of making errors due to reduced reading performance[31]. These findings suggest that clinical experience makes practitioners more resilient to automation bias and more confident in their own professional opinion. However, this conclusion has been challenged by other studies that suggest that automation bias affects clinicians differently depending on, among other things, the complexity of the medical problem[32].

This issue therefore requires that experts continue to study the long-term effects of the use of AI in medical contexts. Regardless of the exact conditions for the phenomenon of deskilling of medical professionals, there is sufficient evidence of this type of bias in the medical profession, and it is reasonable to consider this a notable risk to patients' right to health care, based on the argument that ML models and automated machines cannot replace clinicians, and therefore If the essential skills of clinicians are reduced, the likelihood of errors and mistakes could increase, leading to adverse or negative clinical outcomes, since, as explained in Chap. 1, ML models are not fail-safe and may perform poorly in certain settings or with certain demographics, depending on technical factors and how the integration and onboarding process is carried out. In terms of normative implications for clinical workflow and collaboration with other clinicians, the risk of less reliable clinical skills could diminish the clinician's impact on colleagues or make teamwork seem redundant if they must frequently rely on the mode or machine for results.

The final aspect regarding the role of the physician is the level of digital literacy. There is evidence that building clinician confidence in the functioning of an AI system, i.e. understanding its technical and practical capabilities and limitations, generally leads to a more balanced interaction and can have a positive impact on the achievement of normative goals regarding its implementation[33].

---

[31] Thomas Dratsch et al., "Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance," *Radiology* 307, no. 4 (May 1, 2023): e222176, https://doi.org/10.1148/radiol.222176.

[32] Federico Cabitza et al., "Painting the Black Box White: Experimental Findings from Applying XAI to an ECG Reading Setting," *Machine Learning and Knowledge Extraction* 5, no. 1 (March 8, 2023): 269–86, https://doi.org/10.3390/make5010017.

[33] Enoch Yi-No Kang, Duan-Rung Chen, and Yen-Yuan Chen, "Associations between Literacy and Attitudes toward Artificial Intelligence–Assisted Medical Consultations: The Mediating Role of Perceived Distrust and Efficiency of Artificial Intelligence," *Computers*

One normative implication of a lack of digital literacy is that clinicians may be held personally responsible for human errors related to the use of ML models if they use them without adequate training. For example, an inexperienced clinician without proper training may not be able to recognize when the model has produced a flawed result. In this sense, there is a dual obligation between the healthcare institutions that purchase the ML-driven medical device and the medical professionals who are required or interested in using ML models in clinical practice. Healthcare institutions must provide all the elements necessary for clinicians to acquire the skills to use the model (paid time, physical space, materials, and instructors), and clinicians have an obligation to learn these skills.

5.4.3 Normative Considerations About Machine Learning Models

In the last subsection of this dissertation, I focus on the normative considerations that arise from the role of ML models in clinical settings and their purpose. There are three clinical roles based on the study by Pee et al. mentioned in Sect. 5.2.2, and three general categories of models that I classify based on the state-of-the-art review of CDSS by Sutton et al.[34]. I will map the normative considerations according to the roles the models can take in each of the four categories. The three roles used in this evaluation are: augment, the model is intended to enhance or expand the clinician's skills or actions; assist, the model is intended to support the clinician by reducing cognitive effort; and automate, the model is intended to support the clinician by reducing physical effort[35]. The three categories are: first, models that have a direct impact on diagnostic decisions; second, models that support diagnostic processes through workflow and administrative tasks; and third, models that support medical professionals in clinical but non-diagnostic tasks.

---

in Human Behavior 139 (February 1, 2023): 107529, https://doi.org/10.1016/j.chb.2022.107529.

[34] Sutton et al., "An Overview of Clinical Decision Support Systems," 5–6.

[35] In contrast to Pee et al., I do not include the role fourth role proposed, namely, of actuation that describes the role in which the model replaces the clinician. The reasoning is twofold. First, because it would require including the normative considerations derived from ML-driven robots, and this is beyond the scope of this dissertation. Second, as I have argued before, I do not consider it normatively desirable or justifiable to use AI systems to fully replace physicians in diagnostic settings.

First, in the category of directly influencing diagnostic decisions, models can play all three roles and can be used in both primary and secondary care. Models that enhance the clinician's ability to make diagnostic decisions can be used for tasks such as image analysis as a "second opinion", like with Lumify, a model developed by Philipps to detect abnormalities during pregnancy[36], or models like KardiaAI, developed by AliveCor to help detect if a patient is exhibiting symptoms from a spectrum of certain cardiac abnormalities[37]. There are models that reduce cognitive effort, such as advanced symptom checkers that can be used to perform a comprehensive differential diagnosis in primary care, as in the case of DxPlain[38]. Finally, there are models that have been certified to conduct autonomous diagnostic decisions in clinical settings, such as LumineticsCore for diabetic retinopathy[39].

In the second category, models can support the diagnostic process through administrative and operational tasks by automating or assisting clinicians. These include models that automatically alert clinicians to drug-drug interactions that require further evaluation or immediate action[40], that help manage the administration of medications to hospitalized patients[41], or that automate the process of measuring blood glucose levels at the point of care in the intensive care unit[42].

[36] Minh-Phuong T. Le et al., "Comparison of Four Handheld Point-of-Care Ultrasound Devices by Expert Users," *The Ultrasound Journal* 14, no. 1 (July 7, 2022): 27, https://doi.org/10.1186/s13089-022-00274-6.

[37] Ali Bahrami Rad et al., "A Crowdsourced AI Framework for Atrial Fibrillation Detection in Apple Watch and Kardia Mobile ECGs," *Sensors* 24, no. 17 (January 2024): 5708, https://doi.org/10.3390/s24175708.

[38] William F. Bond et al., "Differential Diagnosis Generators: An Evaluation of Currently Available Computer Programs," *Journal of General Internal Medicine* 27, no. 2 (February 2012): 213–19, https://doi.org/10.1007/s11606-011-1804-8.

[39] Risa M. Wolf et al., "Autonomous Artificial Intelligence Increases Screening and Follow-up for Diabetic Retinopathy in Youth: The ACCESS Randomized Control Trial," *Nature Communications* 15 (January 11, 2024): 421, https://doi.org/10.1038/s41467-023-44676-z.

[40] Habibollah Pirnejad et al., "Preventing Potential Drug-Drug Interactions through Alerting Decision Support Systems: A Clinical Context Based Methodology," *International Journal of Medical Informatics* 127 (July 1, 2019): 18–26, https://doi.org/10.1016/j.ijmedinf.2019.04.006.

[41] Jetske Graafsma et al., "The Use of Artificial Intelligence to Optimize Medication Alerts Generated by Clinical Decision Support Systems: A Scoping Review," Journal of the American Medical Informatics Association: JAMIA 31, no. 6 (April 19, 2024): 1411–22, https://doi.org/10.1093/jamia/ocae076.

[42] Sravani Medanki et al., "Artificial Intelligence Powered Glucose Monitoring and Controlling System: Pumping Module," *World Journal of Experimental Medicine* 14, no. 1 (March 20, 2024): 87916, https://doi.org/10.5493/wjem.v14.i1.87916.

In all of these examples, the models reduce the physical and cognitive demands on clinicians and could potentially have a positive impact on reducing clinician burnout, satisfaction, and turnover.

In the third category, the models used to support clinicians with clinical tasks but with no direct diagnostic impact can also hold the three roles. A model can augment the skills of a clinician by improving the quality of a medical image or aiding in the process of image segmentation, in which the model provides a clear area where a potential abnormality could exist[43]. Since the physician still needs to perform the diagnostic process, these types of models are categorized here. The models that either assist or automate clinical tasks belong mostly to tasks in laboratory testing and interpretation of pathology results. For instance, models that help clinicians to grade tumors[44] or automate blood-cell counting[45].

At this level, there are normative considerations that may have been mentioned earlier but need further evaluation. First, we need to ask whether the model we are evaluating is actually needed at all. This means determining whether there is a real area of opportunity, such as a significantly high rate of medical error with respect to the particular disease to be diagnosed, or the existence of a highly inefficient clinical workflow that impacts clinician scheduling, the cost of diagnosing the disease, and the efficiency of providing care to the patient, that the model is designed to help solve or completely solve. Closely related, the second aspect to consider is whether this particular model is capable of doing the job that clinicians need it to do. In other words, whether the model can fulfill the normative goal of clinicians with respect to its implementation and use. A third aspect to examine is what are the requirements for integrating the model into clinical practice are. For example, what are the clinicians' perceptions of the model and expectations of its performance? How long will it take to train clinicians to use the model optimally? How will the model's performance be measured?

---

[43] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib, "Deep Learning for Medical Image Processing: Overview, Challenges and the Future," in *Classification in BioApps: Automation of Decision Making*, ed. Nilanjan Dey, Amira S. Ashour, and Surekha Borra (Cham: Springer International Publishing, 2018), 323–50, https://doi.org/10.1007/978-3-319-65981-7_12.

[44] Eugene Vorontsov et al., "A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection," *Nature Medicine*, July 22, 2024, 1–12, https://doi.org/10.1038/s41591-024-03141-0.

[45] Francesca Isabelle F. Escobar et al., "Automated Counting of White Blood Cells in Thin Blood Smear Images," *Computers and Electrical Engineering* 108 (May 1, 2023): 108710, https://doi.org/10.1016/j.compeleceng.2023.108710.

The fourth and final aspect that needs to be evaluated is how the hospital's protocols and policies address the implementation of the model in clinical practice. In the regulatory constellation, it was mentioned that governance structures play a relevant role in ensuring the effective protection of people's rights in addition to government regulation—in this sense it is relevant to consider whether there are additional or new guidelines for cases where there are technical errors leading to undesired outcomes, errors at the intersection of human-model interaction, or human mistakes derived from the use of the model (see Sect. 4.2.5 for a more detailed account of these categories).

Considering the normative analysis presented, it becomes clear that the integration of ML models in health care requires more than technical efficiency- it requires an ethical framework based on relational rights and shared responsibilities. This approach not only recognizes the complexity of medical encounters but also insists on the need to balance technological innovation with the preservation of human dignity. Central to the argument developed in this chapter is the assertion that applied ethics must remain a dynamic, interdisciplinary field, bridging philosophy and real-world practice. The use of ML models in medical diagnosis exemplifies this need by challenging traditional roles and responsibilities within clinical settings. Patients, physicians, and caregivers are intertwined in a network of ethical relationships. Within this network, patients are empowered by their rights but constrained by the moral responsibilities they share with others, including clinicians and caregivers.

The relational rights-based approach offers a nuanced way of navigating these responsibilities. It encourages health care providers to move beyond mere compliance with technical guidelines or regulatory standards to a deeper engagement with the moral dimensions of care. Physicians must be vigilant not only about the technical accuracy of ML models, but also about their potential to disempower professionals or reinforce harmful biases. In this sense, digital literacy and continuing professional development are necessary to ensure that clinicians remain active agents in patient care rather than passive operators of technology.

Ethical incorporation of ML models requires healthcare institutions to rethink their governance structures to ensure that models are not adopted at face value but are continually evaluated for their real-world impact on patient care and clinical workflows. Protocols must be established to manage errors at the intersection of human and machine decision making, reflecting the broader ethical goal of protecting patient rights while promoting clinical effectiveness. I make the argument that the ethical integration of ML in healthcare can find guidance in the proposed

a relational, rights-based approach. The moral agents within these constellations must each recognize their shared vulnerability and interdependence and embrace a model of care that prioritizes not only the outcomes of diagnosis but also the dignity and rights of all involved.

Conclusion

6

In this dissertation, I set out to argue that there is an urgent need to make a normative assessment of the distribution of benefits and risks of implementing machine learning models in diagnostic settings. In the introduction, I justified this claim with three central arguments. First, that there is a lack of clarity about how to deal with situations where conflict between principles, values, and rights emerge. Second, that a significant portion of the discourse in the field of AI ethics takes an approach based on principles which faces major hurdles to provide actionable guidance in practical contexts and thus is not easily enforceable in regulatory contexts. Three, there is a lack of research about the convergence and potential tension between the normative aims of relevant actors at the intersection of the development of machine learning models, their implementation in healthcare settings, and their effective integration in clinical workflows.

Integrating advanced machine learning tools into medical diagnostics represents one of the most significant technological advances in healthcare today. Although an increasingly common narrative says that the future of healthcare will be largely shaped by the capabilities of these systems, this dissertation has attempted to highlight that the promise of ML in medicine must be tempered by a thoughtful, rigorous examination of its ethical, practical, and technical dimensions. In this concluding section, I summarize the key findings of the dissertation, reaffirm the importance of a normative assessment of how the distribution of the risks and benefits of this implementation should be considered, and reflect on the broader implications for future research.

The opening chapter of this dissertation presented an overview of the current state of ML in medical diagnosis, emphasizing both its potential and the challenges it introduces. From its ability to detect diseases earlier and more accurately

L. D. Dias Duran, *Machine Learning in Medical Diagnosis*, Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics,
https://doi.org/10.1007/978-3-662-71357-0_6

to its promise of supporting overwhelmed healthcare systems, ML has started to show its value in specific areas of healthcare. Yet, as the historical overview of AI and ML in medicine highlighted, the path toward realizing any potential benefit is not without obstacles. One of the central insights from this dissertation is that the application of ML in healthcare requires us to critically assess the expectations placed on these technologies. Over-enthusiasm about AI's capabilities can be dangerous, not only because it risks disappointing stakeholders, but also because it may overshadow important normative considerations surrounding patient rights. For healthcare, such disillusionment is not merely a technical setback but could have direct and harmful implications for patient care. Therefore, as this dissertation has repeatedly argued, healthcare must be approached with cautious optimism, balancing the excitement of innovation with a deep commitment to serious normative evaluation.

Chap. 2 answered two questions of justificatory nature. First, why is an evaluation of the distribution of potential risks and benefits necessary? And second, why was a normative approach based on rights selected? The answer to the first question touched on the normative complexities at the intersection of healthcare and ML and was framed by the several gaps identified in the literature: the epistemic, responsibility, conceptual, and implementation gaps. This chapter also outlined other methodological approaches to AI ethics and opened the discussion to introduce the proposed normative approach. Chap. 3 introduced the groundwork of the proposed rights-based approach to analyze and evaluate the risks and benefits of the application of ML in medical diagnosis. The framework consists of four pillars drawn from the philosophical work of Klaus Steigleder and Alan Gewirth, and a transversal floor grounded in elements of relational theory that address the shortcomings of the atomistic approaches to rights. The relational approach highlighted the normative relevance of relationships as fundamental considerations in rights theories and as essential to grasp the complexity of the contexts where medical ML models are being used. This chapter also looked at the impact of AI tools on the conceptualization and normative implications of certain relevant rights. This analysis made clear, as proposed in the second pillar, that in practical contexts rights are almost always not absolute and a rigorous normative assessment of how potential tensions arise and what elements play a role is necessary to evaluate when and under which circumstances a certain right might need to be prioritized over other or others.

Chap. 4 was concerned with identifying and analyzing the various risks that arise or could arise from the integration of ML models in diagnostic settings. While these risks vary in nature and impact, and many of them have not yet materialized, i.e., they are potential risks, they still raise concerns that should be

considered and addressed in advance to prevent harm from occurring. This is not an easy task because some of these risks are slow to develop and their effects may not be felt for years, if not decades. However, the value of the normative assessment undertaken in this chapter shows that it is not unrealistic to develop frameworks that can help prepare for these challenges. In addition to this task, the purpose of this chapter was also to provide a balanced view of existing challenges in healthcare that pose significant risks to the rights of patients worldwide and that could be solved or minimized through the use of certain AI tools. The aim of this chapter was to provide a complete picture and to emphasize that a normative evaluation must consider the risks at all levels in order to make a serious assessment of trade-offs in conflict situations.

Finally, Chap. 5 introduced the Ecosystem of Moral Constellations, a relational rights-based framework that approaches the diagnostic processes as a network of interconnected relationships between patients, healthcare providers, institutions, and AI technologies. This framework emphasizes the need for a dynamic and evolving approach to ethical decision-making in healthcare. It acknowledges that healthcare is not a static field but one that is constantly shaped by new technologies, changing societal values, and evolving patient expectations. The relational aspect of this framework is crucial because it reflects the real-world complexity of healthcare delivery. Decisions about the use of ML in diagnosis do not occur in a vacuum; they are shaped by the relationships between all the agents involved. By focusing on these relationships, the framework encourages a more holistic view of healthcare ethics—one that considers the needs and rights of all stakeholders, not just patients.

The contribution of my work lies in the creation of the Ecosystem of Moral Constellations, as a normative framework grounded in a relational, rights-based approach. In doing so, I have presented a novel approach to identifying both obvious and often overlooked normative considerations that arise in the context of applied AI ethics and medical ethics. To the best of my knowledge, this assessment is the first of its kind and contributes to the field of normative research with its interdisciplinary possibilities for in-depth analysis of nuanced normative challenges.

Despite the great potential of this approach, there are limitations and aspects that have been left for future work. This dissertation is philosophical in scope and nature, and all technical and clinical aspects and theories were acquired through theoretical research. Since there is no empirical data from direct contact with medical professionals or AI developers, it is possible that some nuances are not included or are misinterpreted. The second limitation relates to the regulatory elements mentioned at various points in the dissertation. This work did not attempt

to provide in-depth explanations of existing regulatory structures, and those mentioned were researched in the context of the specific topic at hand. Moreover, this dissertation did not cover topics that could be considered within the general scope of AI ethics in healthcare but were considered too broad and of little direct value against the time constraints. These topics include the role of medical robots in the diagnostic setting and the philosophical debate about the potential moral agency of fully automated systems. A final limitation of this dissertation, and one that will likely be continued in future work, is the integration of the normative considerations in all four constellations and conducting a general assessment of the distribution of risks and benefits at the ecosystem level.

There is much research to be done in the field of applied ethics research in connection with AI and health. This dissertation opens up a path of research that can have a significant impact on how normative evaluations of challenges in medical ethics are conducted. The Ecosystem of Moral Constellation framework can be a useful prospective tool to inform medical professionals about normative aspects that might be difficult to foresee. It could even be integrated into clinical trials designed to evaluate the medical safety and efficacy of AI tools or into longitudinal studies that are interested in ethical aspects that quantitative or qualitative methods cannot grasp.

An important lesson that the excitement of implementing ML models in healthcare has taught us is about human limitations and the importance of understanding and accepting them. This does not mean giving up on the ideals and goals of advancing medical research and practice. Rather, reflecting on these limitations is a reflection on our strengths and values. Awareness of these important issues can help us decide what role we want technologies like AI to play, and how we can work to ensure that they indeed help us fulfill our moral obligations to others.

I opened this dissertation with the question about the role of applied ethics. Although this dissertation has most assuredly not provided a final answer, I shall argue that it at least contributes to it by adding a crucial element that I discovered with each chapter I wrote. Applied ethics is -and ought to be- *prospective* and as such, one of its roles is to peer into the future challenges and provide an informed argumentation about what we can expect and what we can do to address them.

Bibliography

104th Congress of the United States. Health Insurance Portability and Accountability Act, Pub. L. No. Public Law 104–191 (1996). https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.

Abrusci, Elena, and Richard Mackenzie-Gray Scott. "The Questionable Necessity of a New Human Right against Being Subject to Automated Decision-Making." *International Journal of Law and Information Technology* 31, no. 2 (August 30, 2023): 114–43. https://doi.org/10.1093/ijlit/eaad013.

Ahmat, Adam, Sunny C. Okoroafor, Isabel Kazanga, James Avoka Asamani, Jean Jacques Salvador Millogo, Mourtala Mahaman Abdou Illou, Kasonde Mwinga, and Jennifer Nyoni. "The Health Workforce Status in the WHO African Region: Findings of a Cross-Sectional Study." *BMJ Global Health* 7, no. Suppl 1 (May 1, 2022): 1–8. https://doi.org/10.1136/bmjgh-2021-008317.

Ahn Hyeong Sik, Kim Hyun Jung, and Welch H. Gilbert. "Korea's Thyroid-Cancer 'Epidemic' — Screening and Overdiagnosis." *New England Journal of Medicine* 371, no. 19 (2014): 1765–67. https://doi.org/10.1056/NEJMp1409841.

Altman, Sam. "Planning for AGI and Beyond." *Open AI Blog* (blog), February 24, 2023. https://openai.com/index/planning-for-agi-and-beyond/.

Aluttis, Christoph, Tewabech Bishaw, and Martina W. Frank. "The Workforce for Health in a Globalized Context – Global Shortages and International Migration." *Global Health Action* 7, no. 1 (December 2014): 23611. https://doi.org/10.3402/gha.v7.23611.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.," May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ash, J. S., M. Berg, and E. Coiera. "Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System-Related Errors." *Journal of the American Medical Informatics Association* 11, no. 2 (April 2004): 104–12. https://doi.org/10.1197/jamia.M1471.

Azam, Kamran, Anwar Khan, and Muhammad Toqeer Alam. "Causes and Adverse Impact of Physician Burnout: A Systematic Review" 27 (2017).

L. D. Dias Duran, *Machine Learning in Medical Diagnosis*, Neue Wege der Angewandten Ethik / New Pathways of Applied Ethics,
https://doi.org/10.1007/978-3-662-71357-0

Babic, Boris, Sara Gerke, Theodoros Evgeniou, and I. Glenn Cohen. "Direct-to-Consumer Medical Machine Learning and Artificial Intelligence Applications." *Nature Machine Intelligence* 3, no. 4 (April 2021): 283–87. https://doi.org/10.1038/s42256-021-00331-0.

Babushkina, Dina. "Are We Justified Attributing a Mistake in Diagnosis to an AI Diagnostic System?" *AI and Ethics* 3, no. 2 (May 1, 2023): 567–84. https://doi.org/10.1007/s43681-022-00189-x.

Baird, P.A., and A.D. Sadovnick. "Survival in Infants with Anencephaly." *Clinical Pediatrics* 23, no. 5 (May 1, 1984): 268–71. https://doi.org/10.1177/000992288402300505.

Balogh, Erin P., Bryan T. Miller, and John R. Ball, eds. *Improving Diagnosis in Health Care.* Washington, D.C.: National Academies Press, 2015. http://www.nap.edu/catalog/21794.

Bastian, Brock, Charlie R. Crimston, Christoph Klebl, and Paul A. M. van Lange. "The Moral Significance of Protecting Environmental and Cultural Objects." *PLOS ONE* 18, no. 2 (February 9, 2023): 1–22. https://doi.org/10.1371/journal.pone.0280393.

Bell, Michael Z. "Why Expert Systems Fail." *Journal of the Operational Research Society* 36, no. 7 (July 1985): 613–19. https://doi.org/10.1057/jors.1985.106.

Bergum, Vangie. "Beyond Rights: The Ethical Challenge." *Phenomenology + Pedagogy*, January 1, 1992, 75–84. https://doi.org/10.29173/pandp14904.

Bergum, Vangie. "Discourse - Ethical Challenges of the 21st Century: Attending to Relations." *Canadian Journal of Nursing Research Archive*, 2002. https://cjnr.archive.mcgill.ca/article/view/1761.

Bergum, Vangie, and John Dossetor. *Relational Ethics: The Full Meaning of Respect*. Independently published, 2020.

Bernstein, Michael H., Michael K. Atalay, Elizabeth H. Dibble, Aaron W. P. Maxwell, Adib R. Karam, Saurabh Agarwal, Robert C. Ward, Terrance T. Healey, and Grayson L. Baird. "Can Incorrect Artificial Intelligence (AI) Results Impact Radiologists, and If so, What Can We Do about It? A Multi-Reader Pilot Study of Lung Cancer Detection with Chest Radiography." *European Radiology* 33, no. 11 (June 2, 2023): 8263–69. https://doi.org/10.1007/s00330-023-09747-1.

Berry, Shandeigh N. "Providing Palliative Care to Neonates With Anencephaly in the Home Setting." *Journal of Hospice & Palliative Nursing* 23, no. 4 (August 2021): 367. https://doi.org/10.1097/NJH.0000000000000770.

Bettinsoli, Maria Laura, Daniela Di Riso, Jaime L. Napier, Lorenzo Moretti, Pierfrancesco Bettinsoli, Michelangelo Delmedico, Andrea Piazzolla, and Biagio Moretti. "Mental Health Conditions of Italian Healthcare Professionals during the COVID-19 Disease Outbreak." *Applied Psychology. Health and Well-Being* 12, no. 4 (December 2020): 1054–73. https://doi.org/10.1111/aphw.12239.

Bietti, Elettra. "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–19. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372860.

Birhane, Abeba. "Algorithmic Injustice: A Relational Ethics Approach." *Patterns* 2, no. 2 (February 2021): 100205. https://doi.org/10.1016/j.patter.2021.100205.

Bleher, Hannah, and Matthias Braun. "Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice." *Science and Engineering Ethics* 29, no. 3 (May 26, 2023): 1–21. https://doi.org/10.1007/s11948-023-00443-3.

Boer, Bas de, and Olya Kudina. "What Is Morally at Stake When Using Algorithms to Make Medical Diagnoses? Expanding the Discussion beyond Risks and Harms." *Theoretical Medicine and Bioethics* 42, no. 5 (December 1, 2021): 245–66. https://doi.org/10.1007/s11017-021-09553-0.

Bolitho, Marc. "Powering The Future: Meet The Scaling Energy Demands Of Generative AI." *Forbes*, April 22, 2024. https://www.forbes.com/councils/forbestechcouncil/2024/04/22/powering-the-future-meet-the-scaling-energy-demands-of-generative-ai/.

Bongaarts, John. "Human Population Growth and the Demographic Transition." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, no. 1532 (October 27, 2009): 2985–90. https://doi.org/10.1098/rstb.2009.0137.

Boniol, Mathieu, Teena Kunjumen, Tapas Sadasivan Nair, Amani Siyam, James Campbell, and Khassoum Diallo. "The Global Health Workforce Stock and Distribution in 2020 and 2030: A Threat to Equity and 'Universal' Health Coverage?" *BMJ Global Health* 7, no. 6 (June 2022): 1–8. https://doi.org/10.1136/bmjgh-2022-009316.

Borenstein, Jason, Frances S. Grodzinsky, Ayanna Howard, Keith W. Miller, and Marty J. Wolf. "AI Ethics: A Long History and a Recent Burst of Attention." *Computer* 54, no. 1 (January 2021): 96–102. https://doi.org/10.1109/MC.2020.3034950.

Bovens, Mark. "Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism." *West European Politics* 33, no. 5 (September 2010): 946–67. https://doi.org/10.1080/01402382.2010.486119.

Braverman, Harry. *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. 25th anniversary ed. New York: Monthly Review Press, 1998.

Brodersen, John, Lisa M. Schwartz, Carl Heneghan, Jack William O'Sullivan, Jeffrey K. Aronson, and Steven Woloshin. "Overdiagnosis: What It Is and What It Isn't." *BMJ Evidence-Based Medicine* 23, no. 1 (February 1, 2018): 1–3. https://doi.org/10.1136/ebmed-2017-110886.

Brownstein, Michael. "Implicit Bias." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University, 2019. https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/.

Brugger, Florian, and Christian Gehrke. "Skilling and Deskilling: Technological Change in Classical Economic Theory and Its Empirical Evidence." *Theory and Society* 47, no. 5 (October 1, 2018): 663–89. https://doi.org/10.1007/s11186-018-9325-7.

Buchanan, B G, and R G Smith. "Fundamentals of Expert Systems." *Annual Review of Computer Science* 3, no. 1 (June 1988): 23–58. https://doi.org/10.1146/annurev.cs.03.060188.000323.

Budd, Samuel, Emma C. Robinson, and Bernhard Kainz. "A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis." *Medical Image Analysis* 71 (July 1, 2021): 102062. https://doi.org/10.1016/j.media.2021.102062.

Bundesverfassungsgericht. "Zur Gewährleistung wirkungsvollen Grundrechtsschutzes bei der Übertragung von Hoheitsrechten an supranationale Organisationen," July 24, 2018. https://www.bverfg.de/e/rs20180724_2bvr196109.html.

Cabitza, Federico, Andrea Campagner, and Clara Balsano. "Bridging the 'Last Mile' Gap between AI Implementation and Operation: 'Data Awareness' That Matters." *Annals of Translational Medicine* 8, no. 7 (April 2020): 501–10. https://doi.org/10.21037/atm.2020.03.63.

Cabitza, Federico, Andrea Campagner, Chiara Natali, Enea Parimbelli, Luca Ronzio, and Matteo Cameli. "Painting the Black Box White: Experimental Findings from Applying XAI to an ECG Reading Setting." *Machine Learning and Knowledge Extraction* 5, no. 1 (March 8, 2023): 269–86. https://doi.org/10.3390/make5010017.

Casey, Bryan, Ashkon Farhangi, and Roland Vogl. "Rethinking Explainable Machines: The GDPR's Right to Explanation Debate and the Rise of Algorithmic Audits in Enterprise." *Berkeley Technology Law Journal* 31, no. 1 (2019): 1–143. https://doi.org/10.15779/Z38M32N986.

Chen, Huaming, and M. Ali Babar. "Security for Machine Learning-Based Software Systems: A Survey of Threats, Practices, and Challenges." *ACM Computing Surveys* 56, no. 6 (February 23, 2024): 151:1–151:38. https://doi.org/10.1145/3638531.

Chen, Jonathan H., and Steven M. Asch. "Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations." *The New England Journal of Medicine* 376, no. 26 (June 29, 2017): 2507–9. https://doi.org/10.1056/NEJMp1702071.

Cheng, Fu-Yuan, Himanshu Joshi, Pranai Tandon, Robert Freeman, David L. Reich, Madhu Mazumdar, Roopa Kohli-Seth, Matthew A. Levin, Prem Timsina, and Arash Kia. "Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients." *Journal of Clinical Medicine* 9, no. 6 (June 2020): 1668. https://doi.org/10.3390/jcm9061668.

Chew, Nicholas W.S., Grace K.H. Lee, Benjamin Y.Q. Tan, Mingxue Jing, Yihui Goh, Nicholas J.H. Ngiam, Leonard L.L. Yeo, et al. "A Multinational, Multicentre Study on the Psychological Outcomes and Associated Physical Symptoms amongst Healthcare Workers during COVID-19 Outbreak." *Brain, Behavior, and Immunity* 88 (August 2020): 559–65. https://doi.org/10.1016/j.bbi.2020.04.049.

Citron, Danielle Keats. "The Surveilled Student." SSRN Scholarly Paper. Rochester, NY, August 25, 2023. https://papers.ssrn.com/abstract=4552267.

Clough, Reece Alexander James, William Anthony Sparkes, Oliver Thomas Clough, Joshua Thomas Sykes, Alexander Thomas Steventon, and Kate King. "Transforming Healthcare Documentation: Harnessing the Potential of AI to Generate Discharge Summaries." *BJGP Open* 8, no. 1 (April 2024): BJGPO.2023.0116. https://doi.org/10.3399/BJGPO.2023.0116.

Clusmann, Jan, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, et al. "The Future Landscape of Large Language Models in Medicine." *Communications Medicine* 3, no. 1 (October 10, 2023): 1–8. https://doi.org/10.1038/s43856-023-00370-1.

Coats, Pamela K. "Why Expert Systems Fail." *Financial Management* 17, no. 3 (1988): 77. https://doi.org/10.2307/3666074.

Cohen, Joseph Paul, Tianshi Cao, Joseph D. Viviano, Chin-Wei Huang, Michael Fralick, Marzyeh Ghassemi, Muhammad Mamdani, Russell Greiner, and Yoshua Bengio. "Problems in the Deployment of Machine-Learned Models in Health Care." *CMAJ : Canadian Medical Association Journal* 193, no. 35 (September 7, 2021): E1391–94. https://doi.org/10.1503/cmaj.202066.

Cohen, Joshua, and Peter Neumann. "Cost Savings and Cost-Effectiveness of Clinical Preventive Care." The Synthesis Project. The Robert Wood Johnson Foundation, September 1, 2009.

Convergence: Facilitating Transdisciplinary Integration of Life Sciences, Physical Sciences, Engineering, and Beyond. Washington, D.C.: National Academies Press, 2014. https://doi.org/10.17226/18722.

Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, et al. "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance." *Patterns* 4, no. 10 (October 13, 2023). https://doi.org/10.1016/j.patter.2023.100857.

Corselli-Nordblad, Louise, and Helene Strandell. *Ageing Europe: Looking at the Lives of Older People in the EU*. 2020 Edition. Publications Office of the European Union, 2020. https://data.europa.eu/doi/10.2785/628105.

Crabtree, Andy, Lachlan Urquhart, and Jiahong Chen. "Right to an Explanation Considered Harmful." *SSRN Electronic Journal*, 2019. https://doi.org/10.2139/ssrn.3384790.

Crawford, Kate. *Atlas of AI - Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven London: Yale University Press, 2021.

Crawford, Kate, and Trevor Paglen. "Correction to: Excavating AI: The Politics of Images in Machine Learning Training Sets." *AI & SOCIETY* 36, no. 4 (December 1, 2021): 1399–1399. https://doi.org/10.1007/s00146-021-01301-1.

Criddle, Cristina, and Madhumita Murgia. "Big Tech Companies Cut AI Ethics Staff, Raising Safety Concerns." *Financial Times*, March 29, 2023, sec. Artificial Intelligence. https://www.ft.com/content/26372287-6fb3-457b-9e9c-f722027f36b3.

Crosby, David, Sangeeta Bhatia, Kevin M. Brindle, Lisa M. Coussens, Caroline Dive, Mark Emberton, Sadik Esener, et al. "Early Detection of Cancer." *Science* 375, no. 6586 (March 18, 2022): eaay9040. https://doi.org/10.1126/science.aay9040.

Cullen, Kevin, N. S. Stenhouse, K. L. Wearne, and T. A. Welborn. "Multiple Regression Analysis of Risk Factors for Cardiovascular Disease and Cancer Mortality in Busselton, Western Australia—13-Year Study." *Journal of Chronic Diseases* 36, no. 5 (January 1, 1983): 371–77. https://doi.org/10.1016/0021-9681(83)90169-8.

Cuthbertson Amy. "Scientists Warn New AI May Be 'Slightly Conscious.'" *The Independent*, February 18, 2022, sec. Tech. https://www.independent.co.uk/tech/artificial-intelligence-conciousness-ai-deepmind-b2017393.html.

Danaher, John. "The Rise of the Robots and the Crisis of Moral Patiency." *AI & SOCIETY* 34, no. 1 (March 1, 2019): 129–36. https://doi.org/10.1007/s00146-017-0773-9.

Dang, Thai-Thanh, Pablo Antolin, and Howard Oxley. "Fiscal Implication of Ageing: Projections of Age-Related Spending." SSRN Scholarly Paper. Rochester, NY, September 1, 2001. https://doi.org/10.2139/ssrn.607122.

Dau-Lin, Hsü. "The Myth of the 'Five Human Relations' of Confucius." *Monumenta Serica* 29 (1970): 27–37. https://www.jstor.org/stable/40725916.

Davenport, Thomas, and Ravi Kalakota. "The Potential for Artificial Intelligence in Healthcare." *Future Healthcare Journal* 6, no. 2 (June 2019): 94–98. https://doi.org/10.7861/futurehosp.6-2-94.

De George, Richard T. "The Moral Responsibility of the Hospital." *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 7, no. 1 (February 1, 1982): 87–100. https://doi.org/10.1093/jmp/7.1.87.

De Posada, Francisco González, and Francisco A. González Redondo. "En Torno al «Astra-Torres XIV», El «autómata Ajedrecista» y Los Ensayos Sobre Automática: Leonardo

Torres Quevedo, 1913–2013." *Llull: Revista de La Sociedad Espanola de Historia de Las Ciencias y de Las Tecnicas* 36, no. 78 (December 2013): 456–65.

Dedehayir, Ozgur, and Martin Steinert. "The Hype Cycle Model: A Review and Future Directions." *Technological Forecasting and Social Change* 108 (July 1, 2016): 28–41. https://doi.org/10.1016/j.techfore.2016.04.005.

Descartes, René. *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences.* Translated by Ian Maclean. Oxford World's Classics. Oxford ; New York: Oxford University Press, 2006.

Dratsch, Thomas, Xue Chen, Mohammad Rezazade Mehrizi, Roman Kloeckner, Aline Mähringer-Kunz, Michael Püsken, Bettina Baeßler, Stephanie Sauer, David Maintz, and Daniel Pinto Dos Santos. "Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance." *Radiology* 307, no. 4 (May 1, 2023): e222176. https://doi.org/10.1148/radiol.222176.

Eitel-Porter, Ray and Grosskopf, Ulf. "From AI Compliance to Competitive Advantage." Accenture, June 30, 2022. https://www.accenture.com/us-en/insights/artificial-intelligence/ai-compliance-competitive-advantage.

Elder, Andrew. "Clinical Skills Assessment in the Twenty-First Century." *Medical Clinics of North America* 102, no. 3 (May 2018): 545–58. https://doi.org/10.1016/j.mcna.2017.12.014.

Elder, Andrew, Chris McManus, Lawrence McAlpine, and Jane Dacre. "What Skills Are Tested in the New PACES Examination?" *Annals of the Academy of Medicine, Singapore* 40, no. 3 (March 15, 2011): 119–25. https://doi.org/10.47102/annals-acadmedsg.V40N3p119.

Escobar, Francesca Isabelle F., Jacqueline Rose T. Alipo-on, Jemima Louise U. Novia, Myles Joshua T. Tan, Hezerul Abdul Karim, and Nouar AlDahoul. "Automated Counting of White Blood Cells in Thin Blood Smear Images." *Computers and Electrical Engineering* 108 (May 1, 2023): 108710. https://doi.org/10.1016/j.compeleceng.2023.108710.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542, no. 7639 (February 2017): 115–18. https://doi.org/10.1038/nature21056.

European Commission and Directorate-General for Health and Food Safety. Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space, Pub. L. No. 2022/0140/COD, 13.20.60.00, 15.30.00.00 (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197.

European Parliament. "Directive 95/46/EC | European Data Protection Supervisor," September 20, 2024. https://www.edps.europa.eu/data-protection/our-work/publications/legislation/directive-9546ec_en.

Evgeny Morozov. "Solutionism and Its Discontents." In *To Save Everything, Click Here: The Folly of Technological Solutionism*, 1st ed., 1–16. New York, NY: PublicAffairs, 2014.

Fenwick, Mark, and Paulius Jurcys. "Originality and the Future of Copyright in an Age of Generative AI." *Computer Law & Security Review* 51 (November 2023): 105892. https://doi.org/10.1016/j.clsr.2023.105892.

Ferretti, Maria Paola. "Risk and Distributive Justice: The Case of Regulating New Technologies." *Science and Engineering Ethics* 16, no. 3 (September 2010): 501–15. https://doi.org/10.1007/s11948-009-9172-z.

Ferris, Robert. "Elon Musk: Tesla Will Have All Its Self-Driving Car Features by the End of the Year." *CNBC*, February 19, 2019, sec. Autos. https://www.cnbc.com/2019/02/19/elon-musk-tesla-will-have-all-its-self-driving-car-features-by-the-end-of-the-year.html.

Feuerriegel, Stefan, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. "Generative AI." *Business & Information Systems Engineering* 66, no. 1 (February 2024): 111–26. https://doi.org/10.1007/s12599-023-00834-7.

Flaherty, Stephen, E David Zepeda, Koenraad Mortele, and Gary J Young. "Magnitude and Financial Implications of Inappropriate Diagnostic Imaging for Three Common Clinical Conditions." *International Journal for Quality in Health Care*, January 23, 2019. https://doi.org/10.1093/intqhc/mzy248.

Frey, Melissa K., Muhammad Danyal Ahsan, Hannah Bergeron, Jenny Lin, Xuan Li, Rana K. Fowlkes, Priyanka Narayan, et al. "Cascade Testing for Hereditary Cancer Syndromes: Should We Move Toward Direct Relative Contact? A Systematic Review and Meta-Analysis." *Journal of Clinical Oncology* 40, no. 35 (December 10, 2022): 4129–43. https://doi.org/10.1200/JCO.22.00303.

Frieden, Thomas R., and Debra Houry. "Reducing the Risks of Relief — The CDC Opioid-Prescribing Guideline." *New England Journal of Medicine* 374, no. 16 (April 21, 2016): 1501–4. https://doi.org/10.1056/NEJMp1515917.

Friedman, Batya. "Moral Responsibility and Computer Technology," April 1990. https://eric.ed.gov/?id=ED321737.

Friedman, Batya, Peter H. Kahn, Alan Borning, and Alina Huldtgren. "Value Sensitive Design and Information Systems." In *Early Engagement and New Technologies: Opening up the Laboratory*, edited by Neelke Doorn, Daan Schuurbiers, Ibo Van De Poel, and Michael E. Gorman, 16:55–95. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands, 2013. https://doi.org/10.1007/978-94-007-7844-3_4.

Fries, James F., Bonnie Bruce, and Eliza Chakravarty. "Compression of Morbidity 1980–2011: A Focused Review of Paradigms and Progress." *Journal of Aging Research* 2011 (2011): 261702. https://doi.org/10.4061/2011/261702.

Galligan, Claire, Hannah Rosenfeld, Molly Kleinman, and Shobita Parthasarathy. "Cameras in the Classroom: Facial Recognition Technology in Schools." Technology Assessment Project. Michigan: Gerald R. Ford School of Public Policy. University of Michigan, 2020. http://deepblue.lib.umich.edu/handle/2027.42/191755.

Gans, Simon T. de, Gerdinique C. Maessen, Marjolein H. J. van de Pol, Marjan J. van Apeldoorn, Margot A. L. van Ingen-Stokbroekx, Niels van der Sloot, Carolina J. P. W. Keijsers, and Babette C. van der Zwaard. "Effect of Interprofessional and Intraprofessional Clinical Collaboration on Patient Related Outcomes in Multimorbid Older Patients – a Retrospective Cohort Study on the Intensive Collaboration Ward." *BMC Geriatrics* 23, no. 1 (August 26, 2023): 519. https://doi.org/10.1186/s12877-023-04232-2.

Gao, Yue, Guang-Yao Cai, Wei Fang, Hua-Yi Li, Si-Yuan Wang, Lingxi Chen, Yang Yu, et al. "Machine Learning Based Early Warning System Enables Accurate Mortality Risk Prediction for COVID-19." *Nature Communications* 11, no. 1 (October 6, 2020): 5033. https://doi.org/10.1038/s41467-020-18684-2.

Gardner, John, and Narelle Warren. "Learning from Deep Brain Stimulation: The Fallacy of Techno-Solutionism and the Need for 'Regimes of Care.'" *Medicine, Health Care and Philosophy* 22, no. 3 (September 1, 2019): 363–74. https://doi.org/10.1007/s11019-018-9858-6.

Garg, Amit. "Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review." *DECISION SUPPORT*, 2005, 16. https://doi.org/10.1001/jama.293.10.1223.

General Data Protection Regulation (GDPR). "Art. 12 GDPR – Transparent Information, Communication and Modalities for the Exercise of the Rights of the Data Subject." Accessed September 23, 2024. https://gdpr-info.eu/art-12-gdpr/.

General Data Protection Regulation (GDPR). "Art. 21 GDPR – Right to Object." Accessed September 23, 2024. https://gdpr-info.eu/art-21-gdpr/.

General Data Protection Regulation (GDPR). "Recital 35 - Health Data." Accessed September 23, 2024. https://gdpr-info.eu/recitals/no-35/.

Gewirth, Alan. "Are There Any Absolute Rights?" *The Philosophical Quarterly (1950-)* 31, no. 122 (1981): 1–16. https://doi.org/10.2307/2218674.

Gewirth, Alan. *Reason and Morality*. Chicago: University of Chicago Press, 1978.

Gewirth, Alan. *The Community of Rights*. Chicago, IL: University of Chicago Press, 1996. https://press.uchicago.edu/ucp/books/book/chicago/C/bo3645650.html.

Google. "2023 Environmental Report," 2023. https://sustainability.google/reports/google-2023-environmental-report/.

Gordillo, Nelly, Eduard Montseny, and Pilar Sobrevilla. "State of the Art Survey on MRI Brain Tumor Segmentation." *Magnetic Resonance Imaging* 31, no. 8 (October 1, 2013): 1426–38. https://doi.org/10.1016/j.mri.2013.05.002.

Graber, Mark L. "Progress Understanding Diagnosis and Diagnostic Errors: Thoughts at Year 10." *Diagnosis* 7, no. 3 (August 27, 2020): 151–59. https://doi.org/10.1515/dx-2020-0055.

Graber, Mark L., Nancy Franklin, and Ruthanna Gordon. "Diagnostic Error in Internal Medicine." *Archives of Internal Medicine* 165, no. 13 (July 11, 2005): 1493–99. https://doi.org/10.1001/archinte.165.13.1493.

Greenle, Meredith MacKenzie, Karen B. Hirschman, Ken Coburn, Sherry Marcantonio, Alexandra L. Hanlon, Mary Naylor, Elizabeth Mauer, and Connie Ulrich. "End-of-Life Health-Care Utilization Patterns Among Chronically Ill Older Adults." *American Journal of Hospice and Palliative Medicine®* 36, no. 6 (June 2019): 507–12. https://doi.org/10.1177/1049909118824962.

Gregersen, Fredrik Alexander. "The Impact of Ageing on Health Care Expenditures: A Study of Steepening." *The European Journal of Health Economics* 15, no. 9 (December 2014): 979–89. https://doi.org/10.1007/s10198-013-0541-9.

Griffin, Tricia A., Brian P. Green, and Jos V.M. Welie. "The Ethical Wisdom of AI Developers." *AI and Ethics*, March 20, 2024, 1–11. https://doi.org/10.1007/s43681-024-00458-x.

Griffin, Tricia A., Brian Patrick Green, and Jos V. M. Welie. "The Ethical Agency of AI Developers." *AI and Ethics*, January 9, 2023, 1–9. https://doi.org/10.1007/s43681-022-00256-3.

Gruenberg, Ernest M. "The Failures of Success." *The Milbank Quarterly* 83, no. 4 (December 2005): 779–800. https://doi.org/10.1111/j.1468-0009.2005.00400.x.

Gryz, Jarek, and Marcin Rojszczak. "Black Box Algorithms and the Rights of Individuals: No Easy Solution to the 'Explainability' Problem." *Internet Policy Review* 10, no. 2 (June 30, 2021): 1–24. https://doi.org/10.14763/2021.2.1564.

Hagendorff, Thilo. "Blind Spots in AI Ethics." *AI and Ethics* 2, no. 4 (November 1, 2022): 851–67. https://doi.org/10.1007/s43681-021-00122-8.

Hansson, Sven Ove. "The Ethics of Making Patients Responsible." *Cambridge Quarterly of Healthcare Ethics* 27, no. 1 (January 2018): 87–92. https://doi.org/10.1017/S0963180117000421.

Hao, Karen. "AI Is Taking Water From the Desert." *The Atlantic* (blog), March 1, 2024. https://www.theatlantic.com/technology/archive/2024/03/ai-water-climate-microsoft/677602/.

Haselton, Martie G., Daniel Nettle, and Paul W. Andrews. "The Evolution of Cognitive Bias." In *The Handbook of Evolutionary Psychology*, edited by David M. Buss, 1st ed., 724–46. Wiley, 2015. https://doi.org/10.1002/9780470939376.ch25.

Hauer, Barbara. "Data and Information Leakage Prevention Within the Scope of Information Security." *IEEE Access* 3 (2015): 2554–65. https://doi.org/10.1109/ACCESS.2015.2506185.

Hemberg, Jessica, and Håkan Hemberg. "Ethical Competence in a Profession: Healthcare Professionals' Views." *Nursing Open* 7, no. 4 (July 2020): 1249–59. https://doi.org/10.1002/nop2.501.

Herring, Jonathan. "Forging a Relational Approach: Best Interests or Human Rights?" *Medical Law International* 13, no. 1 (March 1, 2013): 32–54. https://doi.org/10.1177/0968533213486542.

Hirshon, Jon Mark, Nicholas Risko, Emilie JB Calvello, Sarah Stewart de Ramirez, Mayur Narayan, Christian Theodosis, and Joseph O'Neill. "Health Systems and Services: The Role of Acute Care." *Bulletin of the World Health Organization* 91 (May 2013): 386–88. https://doi.org/10.2471/BLT.12.112664.

Ho, Emily, Matthew Ferguson, Jeremy Ive, and Michael Schulter. *Relational Rights*. 1st ed., 2021. https://www.relationalresearch.org/product/relational-rights-book/.

Hoff, Timothy. "Deskilling and Adaptation among Primary Care Physicians Using Two Work Innovations." *Health Care Management Review* 36, no. 4 (October 2011): 338–48. https://doi.org/10.1097/HMR.0b013e31821826a1.

Hohfeld, Wesley Newcomb. "Some Fundamental Legal Conceptions as Applied in Judicial Reasoning." *The Yale Law Journal* 23, no. 1 (1913): 16. https://doi.org/10.2307/785533.

Holst Jensen, Mads, Marie Villumsen, and Thomas Døcker Petersen. *The AAAQ Framework and the Right to Water: International Indicators for Availability, Accessibility, Acceptability and Quality ; an Issue Paper of the AAAQ Toolbox*. Copenhagen: Danish Institute for Human Rights, 2014.

Houssami, Nehmat. "Overdiagnosis of Breast Cancer in Population Screening: Does It Make Breast Screening Worthless?" *Cancer Biology & Medicine* 14, no. 1 (February 2017): 1–8. https://doi.org/10.20892/j.issn.2095-3941.2016.0050.

Howdon, Daniel, and Nigel Rice. "Health Care Expenditures, Age, Proximity to Death and Morbidity: Implications for an Ageing Population." *Journal of Health Economics* 57 (January 1, 2018): 60–74. https://doi.org/10.1016/j.jhealeco.2017.11.001.

Hunt, Paul. "The Human Right to the Highest Attainable Standard of Health: New Opportunities and Challenges." *Transactions of The Royal Society of Tropical Medicine and Hygiene* 100, no. 7 (July 1, 2006): 603–7. https://doi.org/10.1016/j.trstmh.2006.03.001.

Hunt, Paul, and Gunilla Backman. "Health Systems and the Right to the Highest Attainable Standard of Health." *Health and Human Rights* 10, no. 1 (2008): 81–92. https://doi.org/10.2307/20460089.

Huq, Aziz Z. "A Right to a Human Decision." *Virginia Law Review*, Technology, 106, no. 3 (2020): 611–88.

Hutchison, Jacqueline, and Julia Holdsworth. "What Choice? Risk and Responsibilisation in Cardiovascular Health Policy." *Health* 25, no. 3 (May 1, 2021): 288–305. https://doi.org/10.1177/1363459319886106.

Ihara, Craig K. "Are Individual Rights Necessary? A Confucian Perspective." In *Confucian Ethics: A Comparative Study of Self, Autonomy, and Community*, edited by David B. Wong and Kwong-Loi Shun, 11–30. Cambridge: Cambridge University Press, 2004. https://doi.org/10.1017/CBO9780511606960.003.

Institute of Medicine (US) Committee on the Future of Primary Care. "Defining Primary Care." In *Primary Care: America's Health in a New Era*, edited by Molla S. Donaldson, Karl D. Yordy, Kathleen N. Lohr, and Neal A. Vanselow. National Academies Press (US), 1996. https://www.ncbi.nlm.nih.gov/books/NBK232631/.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1, no. 9 (September 2, 2019): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Joseph, Bellal, Hamidreza Hosseinpour, Joseph Sakran, Tanya Anand, Christina Colosimo, Adam Nelson, Collin Stewart, Audrey L. Spencer, Bo Zhang, and Louis J. Magnotti. "Defining the Problem: 53 Years of Firearm Violence Afflicting America's Schools." *Journal of the American College of Surgeons* 238, no. 4 (April 2024): 671. https://doi.org/10.1097/XCS.0000000000000955.

Kaissis, Georgios A., Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. "Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging." *Nature Machine Intelligence* 2, no. 6 (June 2020): 305–11. https://doi.org/10.1038/s42256-020-0186-1.

Kallus, Nathan, and Angela Zhou. "Residual Unfairness in Fair Machine Learning from Prejudiced Data." In *Proceedings of the 35th International Conference on Machine Learning*, 2439–48. PMLR, 2018. https://proceedings.mlr.press/v80/kallus18a.html.

Kan, Wing Shan, and Raul P. Lejano. "Relationality: The Role of Connectedness in the Social Ecology of Resilience." *International Journal of Environmental Research and Public Health* 20, no. 5 (1–7): 3865. https://doi.org/10.3390/ijerph20053865.

Kaul, Vivek, Sarah Enslin, and Seth A. Gross. "History of Artificial Intelligence in Medicine." *Gastrointestinal Endoscopy* 92, no. 4 (October 1, 2020): 807–12. https://doi.org/10.1016/j.gie.2020.06.040.

Kayaalp, Mehmet. "Modes of De-Identification." *AMIA Annual Symposium Proceedings* 2017 (April 16, 2018): 1044–50.

Kelly, Christopher J., Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. "Key Challenges for Delivering Clinical Impact with Artificial Intelligence." *BMC Medicine* 17, no. 1 (December 2019): 195. https://doi.org/10.1186/s12916-019-1426-2.

Kieval, Hillel J. "Pursuing the Golem of Prague: Jewish Culture and the Invention of a Tradition." *Modern Judaism* 17, no. 1 (1997): 1–23.

Kitchin, Rob, and Gavin McArdle. "What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets." *Big Data & Society* 3, no. 1 (June 1, 2016): 2053951716631130. https://doi.org/10.1177/2053951716631130.

Kneese, Tamara. "Climate Justice and Labor Rights | Part I: AI Supply Chains and Workflows." *AI Now Institute* (blog), August 2, 2023. https://ainowinstitute.org/general/climate-justice-and-labor-rights-part-i-ai-supply-chains-and-workflows.

Krittanawong, Chayakrit, HongJu Zhang, Zhen Wang, Mehmet Aydar, and Takeshi Kitai. "Artificial Intelligence in Precision Cardiovascular Medicine." *Journal of the American College of Cardiology* 69, no. 21 (May 30, 2017): 2657–64. https://doi.org/10.1016/j.jacc.2017.03.571.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Commun. ACM* 60, no. 6 (May 24, 2017): 84–90. https://doi.org/10.1145/3065386.

Kushida, Clete A., Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. "Strategies for De-Identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies." *Medical Care* 50 (July 2012): S82. https://doi.org/10.1097/MLR.0b013e3182585355.

LaGrandeur, Kevin. "The Consequences of AI Hype." *AI and Ethics* 4, no. 3 (August 1, 2024): 653–56. https://doi.org/10.1007/s43681-023-00352-y.

Laney, Douglas B. "AI Ethics Essentials: Lawsuit Over AI Denial of Healthcare." *Forbes*. Accessed May 23, 2024. https://www.forbes.com/sites/douglaslaney/2023/11/16/ai-ethics-essentials-lawsuit-over-ai-denial-of-healthcare/.

Langlotz, Curtis P. "Will Artificial Intelligence Replace Radiologists?" *Radiology: Artificial Intelligence* 1, no. 3 (May 2019): 1–3. https://doi.org/10.1148/ryai.2019190058.

Lauer, Dave. "You Cannot Have AI Ethics without Ethics." *AI and Ethics* 1, no. 1 (February 1, 2021): 21–25. https://doi.org/10.1007/s43681-020-00013-4.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." *Nature* 521, no. 7553 (May 2015): 436–44. https://doi.org/10.1038/nature14539.

Li, Pengfei, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models." arXiv, October 29, 2023. http://arxiv.org/abs/2304.03271.

Liberati, Elisa G., Francesca Ruggiero, Laura Galuppo, Mara Gorli, Marien González-Lorenzo, Marco Maraldi, Pietro Ruggieri, et al. "What Hinders the Uptake of Computerized Decision Support Systems in Hospitals? A Qualitative Study and Framework for Implementation." *Implementation Science* 12, no. 1 (December 2017): 113. https://doi.org/10.1186/s13012-017-0644-2.

Lipton, Zachary C. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery." *Queue* 16, no. 3 (June 2018): 31–57. https://doi.org/10.1145/3236386.3241340.

Looi, Mun-Keat. "The European Healthcare Workforce Crisis: How Bad Is It?" *BMJ* 384 (January 19, 2024): q8. https://doi.org/10.1136/bmj.q8.

Ludewigs, Sophie, Jonas Narchi, Lukas Kiefer, and Eva C. Winkler. "Ethics of the Fiduciary Relationship between Patient and Physician: The Case of Informed Consent." *Journal of Medical Ethics*, December 23, 2022, 1–8. https://doi.org/10.1136/jme-2022-108539.

Ludvigsen, Kasper Groes Albin. "The Carbon Footprint of GPT-4." *Medium* (blog), July 18, 2023. https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae.

Maher, Nicole A., Joeky T. Senders, Alexander F. C. Hulsbergen, Nayan Lamba, Michael Parker, Jukka-Pekka Onnela, Annelien L. Bredenoord, Timothy R. Smith, and Marike L. D. Broekman. "Passive Data Collection and Use in Healthcare: A Systematic Review of Ethical Issues." *International Journal of Medical Informatics* 129 (September 1, 2019): 242–47. https://doi.org/10.1016/j.ijmedinf.2019.06.015.

Maina, Ivy W., Tanisha D. Belton, Sara Ginzberg, Ajit Singh, and Tiffani J. Johnson. "A Decade of Studying Implicit Racial/Ethnic Bias in Healthcare Providers Using the Implicit Association Test." *Social Science & Medicine*, The role of Racism in Health Inequalities: Integrating Approaches from Across Disciplines, 199 (February 1, 2018): 219–29. https://doi.org/10.1016/j.socscimed.2017.05.009.

Mäntymäki, Matti, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. "Defining Organizational AI Governance." *AI and Ethics* 2, no. 4 (November 1, 2022): 603–9. https://doi.org/10.1007/s43681-022-00143-x.

Martin, Jacqueline S., Wolfgang Ummenhofer, Tanja Manser, and Rebecca Spirig. "Interprofessional Collaboration among Nurses and Physicians: Making a Difference in Patient Outcome." *Swiss Medical Weekly* 140, no. 1718 (May 8, 2010): 1–12. https://doi.org/10.4414/smw.2010.12648.

Masselink, Leah E., Shoou-Yih D. Lee, and Thomas R. Konrad. "Workplace Relational Factors and Physicians' Intention to Withdraw from Practice." *Health Care Management Review* 33, no. 2 (June 2008): 178. https://doi.org/10.1097/01.HMR.0000304507.50674.28.

Maturana, Carles Rubio, Allisson Dantas De Oliveira, Sergi Nadal, Besim Bilalli, Francesc Zarzuela Serrat, Mateu Espasa Soley, Elena Sulleiro Igual, et al. "Advances and Challenges in Automated Malaria Diagnosis Using Digital Microscopy Imaging with Artificial Intelligence Tools: A Review." *Frontiers in Microbiology* 13 (November 15, 2022): 1006659. https://doi.org/10.3389/fmicb.2022.1006659.

Mauri, Lara, and Ernesto Damiani. "Modeling Threats to AI-ML Systems Using STRIDE." *Sensors* 22, no. 17 (January 2022): 6662. https://doi.org/10.3390/s22176662.

McAlister, Scott, Forbes McGain, Matilde Breth-Petersen, David Story, Kate Charlesworth, Glenn Ison, and Alexandra Barratt. "The Carbon Footprint of Hospital Diagnostic Imaging in Australia." *The Lancet Regional Health – Western Pacific* 24 (July 1, 2022). https://doi.org/10.1016/j.lanwpc.2022.100459.

McCarthy, John. "What Is Artificial Intelligence?" Stanford University, 2004.

McCorduck, Pamela. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 2nd ed. An A K Peters Book. Boca Raton. FL: CRC Press, 2018.

McCormick, Erin. "What Happened When a 'Wildly Irrational' Algorithm Made Crucial Healthcare Decisions." *The Guardian*, July 2, 2021, sec. US news. https://www.theguardian.com/us-news/2021/jul/02/algorithm-crucial-healthcare-decisions.

McKenna, Fitzgerald, Aaron Brody, and Seth D. Baum. "2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy," 2020. https://www.ssrn.com/abstract=3070741.

McLean, Scott, Gemma J. M. Read, Jason Thompson, Chris Baber, Neville A. Stanton, and Paul M. Salmon. "The Risks Associated with Artificial General Intelligence: A Systematic Review." *Journal of Experimental & Theoretical Artificial Intelligence* 35, no. 5 (July 4, 2023): 649–63. https://doi.org/10.1080/0952813X.2021.1964003.

McLennan, Stuart, Amelia Fiske, Leo Anthony Celi, Ruth Müller, Jan Harder, Konstantin Ritt, Sami Haddadin, and Alena Buyx. "An Embedded Ethics Approach for AI Development." *Nature Machine Intelligence* 2, no. 9 (September 2020): 488–90. https://doi.org/10.1038/s42256-020-0214-1.

Medanki, Sravani, Nikhil Dommati, Hema Harshitha Bodapati, Venkata Naga Sai Kowsik Katru, Gollapalli Moses, Abhishek Komaraju, Nanda Sai Donepudi, Dhanya Yalamanchili, Jasti Sateesh, and Pratap Turimerla. "Artificial Intelligence Powered Glucose Monitoring and Controlling System: Pumping Module." *World Journal of Experimental Medicine* 14, no. 1 (March 20, 2024): 87916. https://doi.org/10.5493/wjem.v14.i1.87916.

Meskó, Bertalan, Gergely Hetényi, and Zsuzsanna Győrffy. "Will Artificial Intelligence Solve the Human Resource Crisis in Healthcare?" *BMC Health Services Research* 18, no. 1 (July 13, 2018): 545. https://doi.org/10.1186/s12913-018-3359-4.

Metz, Cade, and Daisuke Wakabayashi. "Google Researcher Says She Was Fired Over Paper Highlighting Bias in AI" *The New York Times*, December 3, 2020, sec. Technology. https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html.

Metz, Thaddeus. "An African Theory of Moral Status: A Relational Alternative to Individualism and Holism." *Ethical Theory and Moral Practice* 15, no. 3 (June 1, 2012): 387–402. https://doi.org/10.1007/s10677-011-9302-y.

Metz, Thaddeus, and Sarah Clark Miller. "Relational Ethics." In *The International Encyclopedia of Ethics*, edited by Hugh LaFollette, 1–10. Blackwell, 2013. https://philarchive.org/rec/METR-7.

Mhlambi, Sabelo. "From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance." *Carr Center for Human Rights Policy*, Human Rights Policy Discussion, July 8, 2020, 1–27.

Milmo, Dan, and Phillip Inman. "Why Has Nvidia Driven Stock Markets to Record Highs?" *The Guardian*, February 23, 2024, sec. Technology. https://www.theguardian.com/technology/2024/feb/23/why-has-nvidia-driven-stock-markets-to-record-highs.

Milosevic, Marina, Dragan Jankovic, Aleksandar Milenkovic, and Dragan Stojanov. "Early Diagnosis and Detection of Breast Cancer." *Technology and Health Care* 26, no. 4 (September 27, 2018): 729–59. https://doi.org/10.3233/THC-181277.

Minsky, Marvin Lee. *Semantic Information Processing*. Reprint. Cambridge (Mass.) London: MIT Press, 1988.

Misau, Yusuf Abdu, Nabilla Al-Sadat, and Adamu Bakari Gerei. "Brain-Drain and Health Care Delivery in Developing Countries." *Journal of Public Health in Africa* 1, no. 1 (August 19, 2010): 20–21. https://doi.org/10.4081/jphia.2010.e6.

Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1, no. 11 (November 2019): 501–7. https://doi.org/10.1038/s42256-019-0114-4.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." *Science and Engineering Ethics* 26, no. 4 (August 2020): 2141–68. https://doi.org/10.1007/s11948-019-00165-5.

Morley, Jessica, Caio C.V. Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. "The Ethics of AI in Health Care: A Mapping Review." *Social Science & Medicine* 260 (September 2020): 113172. https://doi.org/10.1016/j.socscimed.2020.113172.

Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism.* Paperback 1. publ. New York, NY: PublicAffairs, 2014.

Mosqueira-Rey, Eduardo, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. "Human-in-the-Loop Machine Learning: A State of the Art." *Artificial Intelligence Review* 56, no. 4 (April 1, 2023): 3005–54. https://doi.org/10.1007/s10462-022-10246-w.

Mozaffari-Kermani, Mehran, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha. "Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare." *IEEE Journal of Biomedical and Health Informatics* 19, no. 6 (November 2015): 1893–1905. https://doi.org/10.1109/JBHI.2014.2344095.

Munn, Luke. "The Uselessness of AI Ethics." *AI and Ethics* 3, no. 3 (August 1, 2023): 869–77. https://doi.org/10.1007/s43681-022-00209-w.

Murdoch, Blake. "Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era." *BMC Medical Ethics* 22, no. 1 (December 2021): 122. https://doi.org/10.1186/s12910-021-00687-3.

Natali, Chiara, Luca Marconi, Leslye Denisse Dias Duran, Massimo Miglioretti, and Federico Cabitza. "AI-Induced Deskilling in Medicine: A Mixed Method Literature Review for Setting a New Research Agenda." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, (March 5, 2025). https://doi.org/10.2139/ssrn.5166364.

National Cancer Institute. "Adjuvant Therapy," February 2, 2011. https://www.cancer.gov/publications/dictionaries/cancer-terms/.

National Highway Traffic Safety Administration. "Summary Report: Standing General Order on Crash Reporting for Automated Driving Systems." U.S Department of Transportation, June 2022.

National Institute of Biomedical Imaging and Bioengineering. "Mammography." Accessed September 16, 2024. https://www.nibib.nih.gov/science-education/science-topics/mammography.

Nedelsky, Jennifer. "Reconceiving Rights and Constitutionalism." *Journal of Human Rights* 7, no. 2 (June 17, 2008): 139–73. https://doi.org/10.1080/14754830802071950.

Nedelsky, Jennifer. "Reconceiving Rights as Relationship." In *Explorations in Difference*, edited by Hart, Jonathan and Bauman, Richard W., 1st ed., 67–88. Routledge, 1996.

Needham, Joseph. *Science and Civilisation in China.* Vol. 2. Cambridge: Cambridge university press, 1991.

Nehmat, Houssami, and Houssami Nehmat. "Overdiagnosis of Breast Cancer in Population Screening: Does It Make Breast Screening Worthless?" *Cancer Biology & Medicine* 14, no. 1 (2017): 1–8. https://doi.org/10.20892/j.issn.2095-3941.2016.0050.

Neumann, Peter J, and Joshua T Cohen. "Cost Savings and Cost-Effectiveness of Clinical Preventive Care." *The Synthesis Project Research Synthesis Report*, no. 18 (September 1, 2009). https://doi.org/48508.

Newman-Toker, David E., Najlla Nassery, Adam C. Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D. Clemens, Zheyu Wang, Yuxin Zhu, et al. "Burden of Serious Harms from Diagnostic Error in the USA." *BMJ Quality & Safety* 33, no. 2 (February 1, 2024): 109–20. https://doi.org/10.1136/bmjqs-2021-014130.

Nguyen, T. V., S. M. Diakiw, M. D. VerMilyea, A. W. Dinsmore, M. Perugini, D. Perugini, and J. M. M. Hall. "Efficient Automated Error Detection in Medical Data Using Deep-Learning and Label-Clustering." *Scientific Reports* 13, no. 1 (November 9, 2023): 1–19. https://doi.org/10.1038/s41598-023-45946-y.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366, no. 6464 (October 25, 2019): 447–53. https://doi.org/10.1126/science.aax2342.

OECD. "Explanatory Memorandum on the Updated OECD Definition of an AI System." Paris: OECD, March 5, 2024. https://doi.org/10.1787/623da898-en.

OECD. "Working Age Population." OECD. Accessed September 20, 2024. https://doi.org/10.1787/d339918b-en.

Olaye, Iredia M., and Azizi A. Seixas. "The Gap Between AI and Bedside: Participatory Workshop on the Barriers to the Integration, Translation, and Adoption of Digital Health Care and AI Startup Technology Into Clinical Practice." *Journal of Medical Internet Research* 25, no. 1 (May 2, 2023): 1–10. https://doi.org/10.2196/32962.

O'Reilly, Michael. "Council Post: The Unseen Data Conundrum." Forbes. Accessed August 7, 2024. https://www.forbes.com/sites/forbestechcouncil/2022/02/03/the-unseen-data-conundrum/.

Oshana, Marina. "Moral Accountability." *Philosophical Topics* 32, no. 1/2 (2004): 255–74.

Oshana, Marina. "Relational Autonomy." In *International Encyclopedia of Ethics*, 1–13. John Wiley & Sons, Ltd, 2020. https://doi.org/10.1002/9781444367072.wbiee921.

Oxford English Dictionary. "Error, n., Etymology." Oxford University Press, June 2024. Oxford English Dictionary. https://doi.org/10.1093/OED/3627921224.

Oxford English Dictionary. "Relational, Adj., Sense 2." Oxford University Press, September 2024. Oxford English Dictionary. https://doi.org/10.1093/OED/5956498667.

Packhäuser, Kai, Sebastian Gündel, Nicolas Münster, Christopher Syben, Vincent Christlein, and Andreas Maier. "Deep Learning-Based Patient Re-Identification Is Able to Exploit the Biometric Nature of Medical Chest X-Ray Data." *Scientific Reports* 12, no. 1 (September 1, 2022): 14851. https://doi.org/10.1038/s41598-022-19045-3.

Parikh, Rajul, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. "Understanding and Using Sensitivity, Specificity and Predictive Values." *Indian Journal of Ophthalmology* 56, no. 1 (2008): 45–50.

Pearl, Judea. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. The Addison-Wesley Series in Artificial Intelligence. Reading, Mass: Addison-Wesley Pub. Co, 1984.

Pedwell, Carolyn. *Feminism, Culture and Embodied Practice: The Rhetorics of Comparison*. First issued in paperback. Transformations. London: Routledge, 2012.

Pee, L. G., Shan L. Pan, and Lili Cui. "Artificial Intelligence in Healthcare Robots: A Social Informatics Study of Knowledge Embodiment." *Journal of the Association for Information Science and Technology* 70, no. 4 (2019): 351–69. https://doi.org/10.1002/asi.24145.

Pei, Jing, Lei Deng, Sen Song, Mingguo Zhao, Youhui Zhang, Shuang Wu, Guanrui Wang, et al. "Towards Artificial General Intelligence with Hybrid Tianjic Chip Architecture." *Nature* 572, no. 7767 (August 2019): 106–11. https://doi.org/10.1038/s41586-019-1424-8.

Pellegrino, Edmund. "The 'Telos' of Medicine and the Good of the Patient." In *Clinical Bioethics*, edited by David C. Thomasma, David N. Weisstub, Thomasine Kimbrough Kushner, and Corrado Viafora, 26:21–32. International Library of Ethics, Law, and the New Medicine. Berlin/Heidelberg: Springer-Verlag, 2005. http://link.springer.com/10.1007/1-4020-3593-4_2.

Perrigo, Billy. "Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer." *TIME*, January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Persson, Ingmar, and Julian Savulescu. "The Impossibility of a Moral Right to Privacy." *Neuroethics* 15, no. 2 (June 28, 2022): 1–5. https://doi.org/10.1007/s12152-022-09500-3.

Petersson, Lena, Ingrid Larsson, Jens M. Nygren, Per Nilsen, Margit Neher, Julie E. Reed, Daniel Tyskbo, and Petra Svedberg. "Challenges to Implementing Artificial Intelligence in Healthcare: A Qualitative Interview Study with Healthcare Leaders in Sweden." *BMC Health Services Research* 22, no. 1 (July 1, 2022): 850. https://doi.org/10.1186/s12913-022-08215-8.

Plakht, Ygal, Harel Gilutz, Jonathan Eli Arbelle, Dan Greenberg, and Arthur Shiyovich. "Healthcare Resources Utilization throughout the Last Year of Life after Acute Myocardial Infarction." *Journal of Clinical Medicine* 12, no. 8 (April 8, 2023): 2773. https://doi.org/10.3390/jcm12082773.

Poel, Ibo van de. "The Problem of Many Hands." In *Moral Responsibility and the Problem of Many Hands*, by Van De Poel, Lamber Royakkers, and Sjoerd D. Zwart, 50–92, 1st ed. Routledge Studies in Ethics and Moral Theory. New York: Routledge, 2015. https://doi.org/10.4324/9781315734217.

Power, Michael, Greg Fell, and Michael Wright. "Principles for High-Quality, High-Value Testing." *BMJ Evidence-Based Medicine* 18, no. 1 (February 1, 2013): 5–10. https://doi.org/10.1136/eb-2012-100645.

Prem, Erich. "From Ethical AI Frameworks to Tools: A Review of Approaches." *AI and Ethics* 3, no. 3 (August 1, 2023): 699–716. https://doi.org/10.1007/s43681-023-00258-9.

Price Ii, William Nicholson, and I. Glenn Cohen. "Locating Liability for Medical AI." *SSRN Electronic Journal*, 2023. https://doi.org/10.2139/ssrn.4517740.

Price, W. Nicholson, and I. Glenn Cohen. "Privacy in the Age of Medical Big Data." *Nature Medicine* 25, no. 1 (January 2019): 37–43. https://doi.org/10.1038/s41591-018-0272-7.

Priester, Alan, Richard E. Fan, Joshua Shubert, Mirabela Rusu, Sulaiman Vesal, Wei Shao, Yash Samir Khandwala, Leonard S. Marks, Shyam Natarajan, and Geoffrey A. Sonn. "Prediction and Mapping of Intraprostatic Tumor Extent with Artificial Intelligence." *European Urology Open Science* 54 (August 2023): 20–27. https://doi.org/10.1016/j.euros.2023.05.018.

Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine Learning in Medicine." *New England Journal of Medicine* 380, no. 14 (April 4, 2019): 1347–58. https://doi.org/10.1056/NEJMra1814259.

Razzak, Muhammad Imran, Saeeda Naz, and Ahmad Zaib. "Deep Learning for Medical Image Processing: Overview, Challenges and the Future." In *Classification in BioApps: Automation of Decision Making*, edited by Nilanjan Dey, Amira S. Ashour, and Surekha Borra, 323–50. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-65981-7_12.

Reif, Linda. "Building Democratic Institutions: The Role of National Human Rights Institutions in Good Governance and Human Rights Protection." *Harvard H.R.J*, 2000, 1–69.

Ricci, Sante Basso, and Ugo Cerchiari. "Spontaneous Regression of Malignant Tumors: Importance of the Immune System and Other Factors (Review)." *Oncology Letters* 1, no. 6 (November 2010): 941–45. https://doi.org/10.3892/ol.2010.176.

Rinard, Ruth G. "Technology, Deskilling, and Nurses: The Impact of the Technologically Changing Environment." *Advances in Nursing Science* 18, no. 4 (June 1996): 60–69.

Ropohl, Günter and Society for Philosophy and Technology. "Philosophy of Socio-Technical Systems." *Society for Philosophy and Technology Quarterly Electronic Journal* 4, no. 3 (1999): 186–94. https://doi.org/10.5840/techne19994311.

Saban, Mor, Barniv Hava, Patito Heli, Shachar Tal, Haber Reuben, Salama Rabia, and Darawsha Aziz. "Choosing Wisely in the ED: The Diagnostic Cascade of Needless Medical Testing in a Two-Level Study." *The American Journal of Emergency Medicine* 37, no. 9 (September 1, 2019): 1705–8. https://doi.org/10.1016/j.ajem.2018.12.017.

Samonas, Spyridon, and David Coss. "The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security." *Journal of Information Systems Security* 10, no. 3 (2014): 21–45.

Samuel, Arthur L. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3, no. 3 (July 1959): 210–29. https://doi.org/10.1147/rd.33.0210.

Schillemans, Thomas, and Mark Bovens. "The Challenge of Multiple Accountability: Does Redundancy Lead to Overload?" In *Accountable Governance: Problems and Promises*, edited by Melvin J. Dubnick and H. George Frederickson, 1st ed., 19. Routledge, 2011.

Schoffer, Olaf, Dirk Schriefer, Andreas Werblow, Andrea Gottschalk, Peter Peschel, Linda A. Liang, Alexander Karmann, and Stefanie J. Klug. "Modelling the Effect of Demographic Change and Healthcare Infrastructure on the Patient Structure in German Hospitals – a Longitudinal National Study Based on Official Hospital Statistics." *BMC Health Services Research* 23, no. 1 (October 11, 2023): 1081. https://doi.org/10.1186/s12913-023-10056-y.

Schönfeld, Martin. "Who or What Has Moral Standing?" *American Philosophical Quarterly* 29, no. 4 (1992): 353–62.

Schwartz, L. M., and S. Woloshin. "Changing Disease Definitions: Implications for Disease Prevalence. Analysis of the Third National Health and Nutrition Examination Survey, 1988–1994." *Effective Clinical Practice: ECP* 2, no. 2 (1999): 76–85.

Seager, Alexander, Linda Sharp, Laura J. Neilson, Andrew Brand, James S. Hampton, Tom J. W. Lee, Rachel Evans, et al. "Polyp Detection with Colonoscopy Assisted by the GI Genius Artificial Intelligence Endoscopy Module Compared with Standard Colonoscopy in Routine Colonoscopy Practice (COLO-DETECT): A Multicentre, Open-Label, Parallel-Arm, Pragmatic Randomised Controlled Trial." *The Lancet Gastroenterology & Hepatology* 9, no. 10 (October 1, 2024): 911–23. https://doi.org/10.1016/S2468-1253(24)00161-4.

Shahid, Rabia, Muhammad Shoker, Luan Manh Chu, Ryan Frehlick, Heather Ward, and Punam Pahwa. "Impact of Low Health Literacy on Patients' Health Outcomes: A Multicenter Cohort Study." *BMC Health Services Research* 22, no. 1 (September 12, 2022): 1148. https://doi.org/10.1186/s12913-022-08527-9.

Shany, Yuval. "The Case for a New Right to a Human Decision Under International Human Rights Law." SSRN Scholarly Paper. Rochester, NY, October 4, 2023. https://doi.org/10.2139/ssrn.4592244.

Shaw, Elisabeth. "Ethical Decision-Making from a Relational Perspective." In *Ethics and Professional Issues in Couple and Family Therapy*, 2nd ed. Routledge, 2016.

Shortliffe, Edward H., Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green, and Stanley N. Cohen. "Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN System." *Computers and Biomedical Research* 8, no. 4 (August 1, 1975): 303–20. https://doi.org/10.1016/0010-4809(75)90009-9.

Shue, Henry. *Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy: 40th Anniversary Edition.* Princeton University Press, 1997. https://doi.org/10.1515/9780691200835.

Simon, Gregory E., Susan M. Shortreed, R. Yates Coley, Robert B. Penfold, Rebecca C. Rossom, Beth E. Waitzfelder, Katherine Sanchez, and Frances L. Lynch. "Assessing and Minimizing Re-Identification Risk in Research Data Derived from Health Care Records." *eGEMs* 7, no. 1 (n.d.): 6. https://doi.org/10.5334/egems.270.

Singh, Hardeep, Gordon D Schiff, Mark L Graber, Igho Onakpoya, and Matthew J Thompson. "The Global Burden of Diagnostic Errors in Primary Care." *BMJ Quality & Safety* 26, no. 6 (June 2017): 484–94. https://doi.org/10.1136/bmjqs-2016-005401.

Singh, Hardeep, and Saul N. Weingart. "Diagnostic Errors in Ambulatory Care: Dimensions and Preventive Strategies." *Advances in Health Sciences Education: Theory and Practice* 14 Suppl 1, no. 0 1 (September 2009): 57–61. https://doi.org/10.1007/s10459-009-9177-z.

Sivaraman, Venkatesh, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. "Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. Hamburg Germany: ACM, 2023. https://doi.org/10.1145/3544548.3581075.

Smith, Helen. "Clinical AI: Opacity, Accountability, Responsibility and Liability." *AI & SOCIETY* 36, no. 2 (June 1, 2021): 535–45. https://doi.org/10.1007/s00146-020-01019-6.

Statistisches Bundesamt (Destatis). "Press Release No. 330: 12.9 Million Economically Active People Will Reach Statutory Retirement Age in the next 15 Years." Federal Statistical Office. Accessed September 23, 2024. https://www.destatis.de/EN/Press/2022/08/PE22_330_13.html.

Steigleder, Klaus. "Climate Risks, Climate Economics, and the Foundations of Rights-Based Risk Ethics." *Journal of Human Rights* 15, no. 2 (April 2, 2016): 251–71. https://doi.org/10.1080/14754835.2015.1083849.

Steigleder, Klaus. "On the Criteria of the Rightful Imposition of Otherwise Impermissible Risks." *Ethical Perspectives*, no. 3 (2018): 471–95. https://doi.org/10.2143/EP.25.3.3285426.

Steigleder, Klaus. "Risk and Rights: Towards a Rights-Based Risk Ethics." Bochum, 2012.

Steigleder, Klaus, and Johannes Graf Keyserlingk. "Public Tasks During Contagious Disease Pandemics: A Rights-Based Perspective." In *Ethical Public Health Policy Within Pandemics: Theory and Practice in Ethical Pandemic Administration*, edited by Michael Boylan, 149–66. The International Library of Bioethics. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-99692-5_8.

Stoneham, Sophie, Amy Livesey, Hywel Cooper, and Charles Mitchell. "ChatGPT versus Clinician: Challenging the Diagnostic Capabilities of Artificial Intelligence in Dermatology." *Clinical and Experimental Dermatology*, November 19, 2023, llad402. https://doi.org/10.1093/ced/llad402.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 3645–50. Florence, Italy: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/P19-1355.

Susienka, Christine. "Human Responsibilities: A Relational Account of Human Rights." PhD diss., Columbia University, 2017. https://doi.org/10.7916/D84J0SP8.

Sutton, Reed T., David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success." *Npj Digital Medicine* 3, no. 1 (February 6, 2020): 1–10. https://doi.org/10.1038/s41746-020-0221-y.

Swetlitz, Casey Ross, Ike. "IBM Pitched Its Watson Supercomputer as a Revolution in Cancer Care. It's Nowhere Close." *STAT* (blog), September 5, 2017. https://www.statnews.com/2017/09/05/watson-ibm-cancer/.

Swierstra, Tsjalling, and Hedwig te Molder. "Risk and Soft Impacts." In *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, edited by Sabine Roeser, Rafaela Hillerbrand, Per Sandin, and Martin Peterson, 1049–66. Dordrecht: Springer Netherlands, 2012. https://doi.org/10.1007/978-94-007-1433-5_42.

Tajirian, Tania, Vicky Stergiopoulos, Gillian Strudwick, Lydia Sequeira, Marcos Sanches, Jessica Kemp, Karishini Ramamoorthi, Timothy Zhang, and Damian Jankowicz. "The Influence of Electronic Health Record Use on Physician Burnout: Cross-Sectional Survey." *Journal of Medical Internet Research* 22, no. 7 (July 15, 2020): 1–13. https://doi.org/10.2196/19274.

Taylor, Josh. "Photos of Australian Children Used in Dataset to Train AI, Human Rights Group Says." *The Guardian*, July 2, 2024, sec. Technology. https://www.theguardian.com/technology/article/2024/jul/03/australian-children-used-ai-data-stability-midjourney.

Tchou, Philip M., Tamara Miner Haygood, E. Neely Atkinson, Tanya W. Stephens, Paul L. Davis, Elsa M. Arribas, William R. Geiser, and Gary J. Whitman. "Interpretation Time of Computer-Aided Detection at Screening Mammography." *Radiology* 257, no. 1 (October 2010): 40–46. https://doi.org/10.1148/radiol.10092170.

Thavorn, K., Z. Wang, D. Fergusson, S. Van Katwyk, A. Arnaout, and M. Clemons. "Cost Implications of Unwarranted Imaging for Distant Metastasis in Women with Early-Stage Breast Cancer in Ontario." *Current Oncology* 23, no. 11 (February 1, 2016): 52–55. https://doi.org/10.3747/co.23.2977.

Thomson, Judith Jarvis. "The Right to Privacy." *Philosophy & Public Affairs* 4, no. 4 (1975): 295–314.

Thomson, Judith Jarvis, and William Parent. *Rights, Restitution, and Risk: Essays, in Moral Theory*. Cambridge, Mass: Harvard University Press, 1986.

Topol, Eric J. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. First edition. New York, NY: Basic Books, 2019.

Trusov, Michael, Liye Ma, and Zainab Jamal. "Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting." *Marketing Science* 35, no. 3 (May 2016): 405–26. https://doi.org/10.1287/mksc.2015.0956.

Turing, Alan. "I.—Computing Machinery and Intelligence." *Mind* LIX, no. 236 (October 1, 1950): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Ullah, Ehsan, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. "Challenges and Barriers of Using Large Language Models (LLM) Such as ChatGPT for Diagnostic Medicine with a Focus on Digital Pathology – a Recent Scoping Review." *Diagnostic Pathology* 19, no. 1 (February 27, 2024): 43. https://doi.org/10.1186/s13000-024-01464-7.

UN Committee on Economic and Social and Cultural Rights (22nd sess : 2000 : Geneva). "General Comment No. 14 (2000), The Right to the Highest Attainable Standard of Health (Article 12 of the International Covenant on Economic, Social and Cultural Rights)," August 11, 2000. https://digitallibrary.un.org/record/425041.

UN General Assembly. "International Covenant on Economic, Social and Cultural Rights." Accessed September 18, 2024. https://www.refworld.org/legal/agreements/unga/1966/en/33423.

United Nations. "Universal Declaration of Human Rights." United Nations. Accessed September 18, 2024. https://www.un.org/en/about-us/universal-declaration-of-human-rights.

United Nations Department of Economic and Social Affairs. *World Social Report 2023: Leaving No One Behind in an Ageing World*. World Social Report. United Nations, 2023. https://doi.org/10.18356/9789210019682.

U.S. Energy Information Administration (EIA). "Electricity Use in Homes." Accessed September 20, 2024. https://www.eia.gov/energyexplained/use-of-energy/electricity-use-in-homes.php.

Vallor, Shannon. "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character." *Philosophy & Technology* 28, no. 1 (March 2015): 107–24. https://doi.org/10.1007/s13347-014-0156-9.

Van Ryn, Michelle, Rachel Hardeman, Sean M. Phelan, Diana J. Burgess PhD, John F. Dovidio, Jeph Herrin, Sara E. Burke, et al. "Medical School Experiences Associated with Change in Implicit Racial Bias Among 3547 Students: A Medical Student CHANGES Study Report." *Journal of General Internal Medicine* 30, no. 12 (December 2015): 1748–56. https://doi.org/10.1007/s11606-015-3447-7.

Vorontsov, Eugene, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, et al. "A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection." *Nature Medicine*, July 22, 2024, 1–12. https://doi.org/10.1038/s41591-024-03141-0.

Vynck, Gerrit De, and Will Oremus. "As AI Booms, Tech Firms Are Laying off Their Ethicists." *Washington Post*, March 30, 2023. https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/.

Wachter, Robert. *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age*, 2017.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7, no. 2 (May 2017): 76–99. https://doi.org/10.1093/idpl/ipx005.

Waddimba, Anthony C., Melinda A. Nieves, Melissa Scribani, Nicole Krupa, Paul Jenkins, and John J. May. "Predictors of Burnout among Physicians and Advanced-Practice Clinicians in Central New York." *Journal of Hospital Administration* 4, no. 6 (August 9, 2015): 21–30. https://doi.org/10.5430/jha.v4n6p21.

Wahl, Brian, Aline Cossy-Gantner, Stefan Germann, and Nina R. Schwalbe. "Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings?" *BMJ Global Health* 3, no. 4 (August 1, 2018): e000798. https://doi.org/10.1136/bmjgh-2018-000798.

Warren, Samuel D., and Louis D. Brandeis. "The Right to Privacy." *Harvard Law Review* 4, no. 5 (1890): 193–220. https://doi.org/10.2307/1321160.

Weber, Griffin M., Kenneth D. Mandl, and Isaac S. Kohane. "Finding the Missing Link for Big Biomedical Data." *JAMA* 311, no. 24 (June 25, 2014): 2479–80. https://doi.org/10.1001/jama.2014.4228.

Weiss, Sholom, Casimir A. Kulikowski, and Aran Safir. "Glaucoma Consultation by Computer." *Computers in Biology and Medicine* 8, no. 1 (January 1, 1978): 25–40. https://doi.org/10.1016/0010-4825(78)90011-2.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave. "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. AIES '19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3306618.3314289.

Widder, David Gray, and Dawn Nafus. "Dislocated Accountabilities in the 'AI Supply Chain': Modularity and Developers' Notions of Responsibility." *Big Data & Society* 10, no. 1 (January 1, 2023): 20539517231177620. https://doi.org/10.1177/20539517231177620.

Wierzbicka, Anna. "The Alphabet of Human Thoughts." In *The Alphabet of Human Thoughts*, 23–52. De Gruyter Mouton, 2011. https://doi.org/10.1515/9783110857108.23.

Wolf, Risa M., Roomasa Channa, T. Y. Alvin Liu, Anum Zehra, Lee Bromberger, Dhruva Patel, Ajaykarthik Ananthakrishnan, et al. "Autonomous Artificial Intelligence Increases Screening and Follow-up for Diabetic Retinopathy in Youth: The ACCESS Randomized Control Trial." *Nature Communications* 15 (January 11, 2024): 421. https://doi.org/10.1038/s41467-023-44676-z.

World Health Organization. "Ageing and Health." *Newsroom* (blog). Accessed September 23, 2024. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

World Health Organization. *Global Strategy on Human Resources for Health: Workforce 2030*. Geneva: World Health Organization, 2016. https://iris.who.int/handle/10665/250368.

World Medical Association. "WMA International Code of Medical Ethics." World Medical Association, April 14, 2023. https://www.wma.net/policies-post/wma-international-code-of-medical-ethics/.

Yanase, Juri, and Evangelos Triantaphyllou. "A Systematic Survey of Computer-Aided Diagnosis in Medicine: Past and Present Developments." *Expert Systems with Applications* 138 (December 2019): 112821. https://doi.org/10.1016/j.eswa.2019.112821.

Yang, Maya. "New York City Schools Ban AI Chatbot That Writes Essays and Answers Prompts." *The Guardian*, January 6, 2023, sec. US news. https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt.

Ying, Xue. "An Overview of Overfitting and Its Solutions." *Journal of Physics: Conference Series* 1168, no. 2 (February 2019): 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.

Yi-No Kang, Enoch, Duan-Rung Chen, and Yen-Yuan Chen. "Associations between Literacy and Attitudes toward Artificial Intelligence–Assisted Medical Consultations: The Mediating Role of Perceived Distrust and Efficiency of Artificial Intelligence." *Computers in Human Behavior* 139 (February 1, 2023): 107529. https://doi.org/10.1016/j.chb.2022.107529.

Zwaan, Laura, and Hardeep Singh. "The Challenges in Defining and Measuring Diagnostic Error." *Diagnosis (Berlin, Germany)* 2, no. 2 (2015): 97–103. https://doi.org/10.1515/dx-2014-0069.

Zweifel, Peter, Stefan Felder, and Markus Meiers. "Ageing of Population and Health Care Expenditure: A Red Herring?" *Health Economics* 8, no. 6 (1999): 485–96. https://doi.org/10.1002/(SICI)1099-1050(199909)8:6<485::AID-HEC461>3.0.CO;2-4.