Technische Universiteit Delft

Towards a Responsible Implementation of Artificial Intelligence in Healthcare

The case of Royal Philips

Juan Camilo Ruiz Reina

TU Delft

PHILIPS

This page is intentionally left blank

Towards a Responsible Implementation of Artificial Intelligence in Healthcare

The case of Royal Philips

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Management of Technology**

Faculty of Technology, Policy and Management

by

Juan Camilo Ruiz Reina

Student number: 4743148

To be defended in public on October 28th, 2019

Graduation committee

First Supervisor	: Prof. dr. ir. IR (Ibo) van de Poel, Section EPT
Second Supervisor	: Dr. G. (Geerten) van de Kaa, Section ETI
Advisor	: Dr. JM. (Juan) Durán, Section EPT
External Supervisor	: MSc. Andrea Franco, Philips

TU Delft

PHILIPS

This page is intentionally left blank

Dedicated to my beloved family:
My parents Juan Antonio and Maria Cristina;
My sisters Patricia, Liliana and Alejandra;
My nieces Tatiana, Estefania and Valery;
My nephew Felipe.

This page is intentionally left blank

Preface

This thesis represents the final step of a two-year amazing journey at TU Delft. The master in Management of Technology (MoT) has been an enriching experience that allowed me to understand why managers, engineers and society should work together to shape the development of technological innovations in order to achieve social and economic benefits. In this work, I explored a technology that has drawn my attention for a long time: Artificial Intelligence (AI). I truly believe that this technology will change our lives and that its potential applications are manifold. However, I am also conscious that AI development must be done in a responsible manner to ensure that it represents a benefit instead of a threat to future generations. The following pages present an initial approach to responsibly introduce this technology in the field of healthcare.

This research project would not have been possible without the support of people whom I want to acknowledge. First, I would like to thank my first supervisor Prof.dr.ir Ibo van de Poel and my second supervisor Dr. Geerten van de Kaa for the valuable suggestions and feedback that I have received during the last six months. The useful guidance given in the different meetings significantly helped me to keep the thesis on a good track. Besides, I am grateful to my advisor Dr. Juan Durán for being interested in this work since the beginning and for the valuable recommendations given during the whole journey.

I would also like to thank my company supervisor Andrea Franco for his continuous support and insightful recommendations. I highly appreciate that since the first day you were committed to helping me to get the best possible result for this project. Furthermore, I would like to thank the Design for Excellence (DfX) team in Philips for offering me the opportunity to do my internship and thesis in the company. It was a remarkable experience to work with the team and to have an initial glimpse of the working culture in the Netherlands.

To my family, thank you for supporting me in every decision that I have taken in life. Without your help and love, I would have never been able to finish a master's degree on the other side of the world. Last, but not least, I would like to thank the friends that I have made in the last two years for making this experience unforgettable.

Juan Camilo Ruiz Reina
Eindhoven, September 2019

This page is intentionally left blank

Executive Summary

The introduction of artificial intelligence (AI) technologies in healthcare is expected to set a paradigm shift to medical practice because these systems will have a significant role in applications such as diagnosis-support and image analysis. However, this implementation does not come without risks. There are important ethical concerns that should be addressed beforehand to ensure public trust and acceptability. Privacy, safety, transparency, reliability and potential biases are some of the issues to consider. Responsible research and innovation (RRI) frameworks have been designed by academics to tackle this sort of problems but there is no application of these frameworks in the field of AI in healthcare. This problem is even more salient in the private sector, due to the reduced interest of companies to engage in RRI practices.

Consequently, one of the research objectives of this project was to offer recommendations on how to implement responsible research and innovation (RRI) practices to avoid potential risks and to improve the social acceptability of AI products in healthcare. For this, we studied the case of Philips by carrying out interviews with the company's experts in AI and corporate social responsibility (CSR). This information was complemented with a comprehensive study of the literature on topics related to AI in healthcare and RRI. The results from these activities were used to create a roadmap to introduce RRI practices in the AI innovation activities within Philips. This roadmap was based on PRISMA, a project developed to guide the implementation of RRI actions in the CSR practices of companies. The goal was to evaluate to what extent RRI practices represent the best approach to deal with the ethical risks that AI might pose. Besides, we analyzed the suitability of the PRISMA roadmap when applied to a large organization. Suggestions to refine the roadmap were delivered.

The results showed that Philips should continue with its rigorous practices of data selection and curation, as well as with its internal policies of cybersecurity and data privacy. Besides, the company should work further on ensuring that data from poor populations are included to train ML algorithms, and on finding diverse candidates to form the AI software development teams. It is also important that healthcare professionals (HCPs) and patients are included in the entire process of product development and not only in the initial phases of product design. They should also be trained to make adequate use of AI-enabled tools. Moreover, Philips should increase the number of activities of experimentation, such as hackathons or bootcamps to develop innovative solutions to deal with ethical problems. An agile-way of working in the Research and Development (R&D) department should also be encouraged to deal with unexpected problems. In telehealth applications, the company should include better interaction features to avoid the risk of social isolation of patients. Finally, Philips should start working on making the machine learning algorithms more transparent to increase the trust from HCPs and patients. These actions will help to enhance the social

acceptability of the AI products developed by the company and should be introduced under the name of CSR or business principles to improve their acceptability by people working in R&I.

This research represented the first step to implement RRI practices in healthcare companies working on AI. However, we focused our analysis on the specific case of Philips. Further research can be carried out in different companies to come up with common principles that contribute to the creation of a more comprehensive framework of responsible research and innovation for AI in healthcare. Besides, carrying out case studies in large corporations could lead to more insights to improve the applicability of RRI in the business strategy of big enterprises.

Keywords: Responsible research and innovation, artificial intelligence, machine learning, healthcare, Philips, PRISMA Project.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
CSR	Corporate Social Responsibility
DfX	Design for Excellence
DL	Deep Learning
EHRs	Electronic Health Records
FDA	Food and Drug Administration
GDPR	General Data Protection Regulation
HCPs	Healthcare Professionals
HIPAA	Health Insurance Portability and Accountability Act
HR	Human Resources
ICBE	Internal Committee for Biomedical Experiments
IGT	Image Guided Therapy
KPI	Key Performance Indicator
ML	Machine Learning
MNC	Multinational Corporation
MRI	Magnetic Resonance Imaging
Q&R	Quality and Regulatory
R&D	Research and Development
R&I	Research and Innovation
RRI	Responsible Research and Innovation
SMEs	Small and Medium Enterprises

1. Introduction

The deployment of electronic health record (EHR) systems has allowed the healthcare sector to collect a great volume of patient information. However, despite the huge amount of data that we have nowadays, only a small fraction is being used and translated into actionable outcomes (Noorbakhsh-Sabet, Zand, Zhang, & Abedi, 2019). The growing interest in Artificial Intelligence (AI) is closely related to this challenge as this technology will enable us to deal with large datasets, solving problems that used to require human intelligence (eHealth Initiative, 2018). AI is expected to create a paradigm shift to healthcare due to the numerous potential applications that it can cater to medical practice. The ability of AI to analyze patient data and generate intelligent suggestions will play a significant role in disease prediction and diagnosis, treatment effectiveness prediction, epidemic outbreak forecasting, and precision health, to name a few (Noorbakhsh-Sabet et al., 2019). Additionally, there is an important economic benefit associated with the development of AI tools for healthcare. For example, Accenture (2017) estimates that key clinical health AI applications can generate $150 billion in annual savings for the US healthcare economy by 2026 and a recent report by Global Market Insights (2019) expects the AI health market to reach $10 billion by 2024. The benefits that AI is presumed to create in medical and economic terms allow us to infer that the introduction of this technology is unavoidable and necessary. For this reason, it is paramount to take into account the potential consequences that its implementation will bring to the clinical field and to society as a whole.

Although the introduction of new technologies aims at improving human well-being and contributing to societal needs, there will always be risks that need to be estimated from the initial stages of development (Doorn, 2014). AI is not the exception and, therefore, the potential ethical issues of including intelligent systems in healthcare have to be fully assessed. The possibility of erroneous decisions, the accountability of decision-making, the transparency of the systems, the inherent bias that can be induced in the algorithms, the privacy and security considerations, and the difficulties to ensure public trust are just some of the ethical topics that have to be analyzed (Nuffield Council of Bioethics, 2018; Academy of Medical Royal Colleges, 2019). In order to address the social and ethical concerns of technology development, there is an approach that has been gaining traction in recent years: '*Responsible Research and Innovation*' (RRI). This concept has been defined as a way of "taking care of the future through collective stewardship of science and innovation in the present" (Stilgoe, Owen, & Macnaghten, 2013). In implementing this model, it is necessary to take into account dimensions of anticipation, reflexivity, inclusiveness and responsiveness, and to evaluate how these dimensions interact with each other (Stilgoe et al., 2013). By considering these dimensions, innovators and policymakers can have a notion of how responsibility is being implemented in the introduction of new technologies and how to tackle potential gaps. This approach allows enhancing the social value in product development to improve the ethical acceptability and societal desirability of a new technology (von

Schomberg, 2012).

This research discusses the implementation of RRI practices in the area of AI in healthcare. The chief practical objective is to introduce RRI practices to prevent the future risks that AI could pose (privacy breaches, discrimination, misdiagnosis, etc.). Specifically, this project is interested in seeing how private companies working in the field of healthcare technology are dealing with the ethical issues that emerge with the introduction of AI systems. This work also examines how responsible practices are steered with the use of an RRI framework. To carry out this research, our case study is Royal Philips, one of the largest companies in medical technology (Ellis, 2019). The framework proposed is based on the PRISMA Project, a research work that presents a roadmap to introduce RRI dimensions in industry (PRISMA Project, 2019).

Moreover, there is a chief scientific objective aimed at analyzing to what extent RRI can be a valid approach to steer responsibly the development of AI in healthcare. By triangulating the results from expert interviews and the literature, we investigate in what cases RRI represents an adequate methodology for dealing with the ethical risks that AI can bring (in healthcare applications) and in what cases the RRI methodology needs to be reframed or a different approach could be better. Recommendations are delivered to RRI researchers and policymakers based on the results. In addition, we discuss the suitability of the PRISMA roadmap when applied in large companies and offer suggestions for further improvement.

1.1 Problem Definition

There have been recent cases in which the viability of AI has been put into question. For example, an Uber self-driving SUV killed a woman on March 2018, because the engineers had shut down the pedestrian detection feature (Marshall & Davies, 2018). In another field, Amazon HR used a machine learning software between 2014 and 2017 to select personnel based on resumes and recommendations. This software ended up being gender-biased towards male applicants as it was trained with data from resumes sent to the company in the last decade when many more male candidates got selected (Dastin, 2018). These incidents had serious repercussions in the mentioned companies and have negatively affected the image and social acceptability of AI-enabled technologies. In the case of healthcare, IBM Watson for Oncology, a cancer-treatment recommendation system powered by AI, delivered wrong advice and put in danger the lives of several patients (Ross, 2018). This was just the first warning call for a sector in which human lives are at stake and if this kind of mistakes keep happening, we could be on our way to the new AI winter.

Misdiagnosis in healthcare or discrimination in recruitment processes are only some of the potential unintended consequences that the introduction of AI algorithms in our daily lives might have. For that reason, a good strategy to deal with those issues is to implement RRI practices to evaluate the potential risks and ethical concerns of the deployment of AI since the beginning of AI product design. This could be an approach to guide the

development of the technology and to improve its social acceptability in the long-term. For example, one of the most important works to address the ethical risks of AI are the "*Ethics Guidelines for Trustworthy AI*" compiled by the High-Level Group on Artificial Intelligence at the European Commission (European Commission, 2018). They designed a framework for trustworthy AI based on the principles of human autonomy, prevention of harm, fairness, and explicability. The document also includes a detailed assessment list that can be used by technology developers to evaluate the trustworthiness of their products. There is also RRI literature in different fields such as semi-autonomous driving (Baumann, Brändle, Coenen, & Zimmer-Merkle, 2019), lethal autonomous weapons (Santoni de Sio & van den Hoven, 2018) and healthcare robotics (Stahl & Coeckelbergh, 2016; Van Wynsberghe, 2015). Besides, Pacifico Silva, Lehoux, Miller and Denis (2018) developed a framework of responsible research and innovation for the entire health domain. Their work was mostly oriented to offer governance recommendations to policy-makers in the public sector (Pacifico Silva, et al., 2018). Similarly, Sun and Medaglia (2019) analyzed the challenges of AI introduction in healthcare for the case of China. Their research identified a broad set of challenges and offered policy recommendations to regulatory bodies and policy-makers in the Chinese government (Sun & Medaglia, 2019). However, there are no studies that provide recommendations to healthcare technology companies on how to deal with the potential risks of AI (see Chapter 2 for further explanation). Moreover, the concept of RRI has been criticized by the private sector for using terms, tools and methodologies that do not relate to industrial practices (Dreyer et al., 2017), and for being naive and idealistic (Sonck, Asveld, Landerweerd & Osseweijer, 2017), leading to a reduced interest from private companies to engage in RRI actions. In fact, Timmermans (2017) expressed that only 10% of people involved in RRI are affiliated to businesses and van de Poel, et al. (2017) stated that "the implementation of RRI in industry is still in its infancy".

This research aims at making a step to close the gap between academics and the private sector in the field of RRI. Specifically, for the responsible introduction of AI systems in healthcare. As said before, van Wynsberghe (2015) and Stahl & Coeckelbergh (2016) are already working on assessing the ethical concerns of the physical branch of AI (robotics). However, this research focuses on the virtual branch of AI in healthcare, which corresponds to machine learning software aimed at optimizing internal processes in hospitals and at offering decision-support to clinicians (Hamet & Tremblay, 2017). For this specific branch, ethical concerns have been identified but there is still no practical approach to deal with them. This situation intensifies in the private sector, where there is a reduced interest to engage in RRI frameworks (Timmermans, 2017). The practical objective then is to offer suggestions to healthcare technology companies, in this case Philips, on how to implement RRI actions to avoid potential risks and to improve the social acceptability of their AI products. Besides, the scientific objective is to analyze to what extent RRI represents an adequate approach for dealing with the ethical risks that AI may pose in healthcare, and to establish in what situations different approaches could be more useful. This leads to a set of recommendations to RRI researchers and policymakers on how to responsibly steer the path of innovation of AI in healthcare. Similarly, it is still to be seen whether the RRI roadmap

designed in PRISMA can be implemented or merged with the Corporate Social Responsibility (CSR) policies of large corporations.

1.1.1 Knowledge Gaps

Following from the problem definition, there are four knowledge gaps that have been identified:

- There is no compelling study that investigates how to implement RRI practices to address the ethical concerns of the introduction of AI (virtual branch) in healthcare.

- There is no study that analyzes whether RRI practices are effective to tackle the ethical risks that AI poses in healthcare or if for some particular situations a different approach could be better.

- Research on how to apply RRI practices in industry is still very limited. Especially in large corporations that have detailed CSR policies already in place. There is no clarity on how to incorporate or merge RRI within the CSR procedures developed by large companies.

- There is no study that evaluates if the PRISMA Project roadmap is suitable when applied in large corporations with established CSR practices.

1.1.2 Problem Statement

The definition of the problem and the knowledge gaps lead to the following problem statement:

There are several risks and ethical concerns, such as misdiagnosis, unintended discrimination or privacy breaches, which should be considered before fully deploying AI systems in healthcare. These risks can lead to situations that will adversely affect the social acceptability of AI. For example, sensitive data being exposed or populations being marginalized of high-quality healthcare. RRI frameworks have been developed to address these issues and guide the responsible introduction of new technologies. However, there is no clarity on how to implement RRI practices in the development of AI systems for healthcare. In addition, there is no certainty that RRI represents the best approach to deal with the ethical risks of AI. This issue intensifies due to the reduced interest from the private sector to engage in RRI practices. The PRISMA Project designed a roadmap to close the gap and incorporate RRI dimensions in industry, but it is still to be seen whether this framework applies for large corporations.

1.2 Research Objectives and Research Questions

1.2.1 Research Objectives

Scientific objectives

- To identify to what extent RRI practices are useful to address the ethical risks raised by AI in the field of healthcare in order to enhance the social acceptability of the technology. Recommendations are delivered to RRI researchers and policymakers based on the results.

- To evaluate whether the RRI roadmap developed in the PRISMA Project applies in large corporations and suggest improvements if necessary.

Practical objective

- The practical objective of this research is to offer recommendations to Philips on how to implement RRI practices to avoid potential risks and to improve the social acceptability of their AI products.

1.2.2 Main Research Question

Following the problem statement and the research objectives, the main question in this research is as follows:

How should RRI be framed and introduced to avoid the risks and enhance the social acceptability of AI-enabled products developed by large healthcare technology companies?

1.2.3 Research Sub-Questions

In order to answer the main research question, it is necessary to study the main ethical concerns related to the introduction of AI in healthcare. Additionally, it is paramount to identify how the healthcare industry is dealing with those concerns in order to verify if the RRI practices currently in place cover all the potential ethical risks that AI might pose or if for some situations a different strategy is needed.

The following set of sub-questions is established to help answering the main research question:

1. *Which are the chief ethical concerns that threaten the social acceptability of AI in healthcare?*

The introduction of AI in healthcare does not come without risks. There are several ethical concerns that need to be urgently addressed before the public acceptability of AI is severely affected. There are issues regarding accountability (who will be held responsible when AI goes wrong?), transparency (how to explain the outcomes of the machine-learning algorithms?), and inclusiveness (how to ensure that AI will not only benefit the wealthiest populations?) that have to be evaluated consciously. By answering this question, we plan to

bring to the table the main ethical issues that have to be taken into account when applying an RRI framework.

2. *How is Philips currently addressing the ethical issues raised by the introduction of AI systems?*

By studying the internal practices of Royal Philips, we plan to understand how the company is currently addressing the ethical concerns raised by their AI developments. This question is intended to study how ethical issues are approached in real practice and whether CSR policies are covering all the potential risks that AI might bring to the health sector.

3. *How should Philips introduce RRI practices to avoid the risks and increase the social acceptability of its AI-enabled products*?

This question is intended to offer a set of recommendations to Philips on how to implement RRI practices to improve the social acceptability of its products. The recommendations will be based on 1) strengthening the RRI practices already identified and 2) suggesting further RRI actions that have not been implemented yet in the process of technology development.

4. *What situations require a different strategy from RRI to address the risks of the implementation of AI in healthcare?*

This question aims at analyzing the specific situations in which RRI could fall short. By comparing the interviews' results with the literature, we discuss the limitations of RRI and offer different strategies to enhance the acceptability of AI in healthcare.

5. *How can the PRISMA roadmap be improved to strengthen its applicability in large corporations?*

The pilots of the PRISMA project have only been carried out at startups and SMEs. This work offers an opportunity to check whether the methodology established by PRISMA applies in large corporations. In case sections of the approach require fine-tuning, the corresponding improvement recommendations will be delivered.

1.3 Scientific and Practical Relevance

1.3.1 Scientific Relevance

The topic of responsible research and innovation in artificial intelligence has gained important traction in the last few years. There is a variety of topics that are being studied, such as AI and ethics at the Police (Dignum & Bieger, 2019), responsible innovation in semi-autonomous driving (Baumann, Brändle, Coenen, & Zimmer-Merkle, 2019), meaningful human control of lethal autonomous weapons (Santoni de Sio & van den Hoven, 2018) and ethics in healthcare robotics (Van Wynsberghe, 2015). However, there are also studies that have criticized the use of RRI in technology development. For example, Sonck, et al. (2017) claimed that RRI is naïve and idealistic. They argued that RRI assumes a transparent and smooth process of deliberation between stakeholders, which is not common in the complex

business world. Similarly, Dreyer, et al. (2017) expressed that there is not alignment between the terminology used in RRI and the concepts used in industry. While academics in the field of RRI formulate actions of anticipation, reflexivity, inclusiveness, and responsiveness to steer responsible technology development (Owen, Macnaghten, & Stilgoe, 2012; Pacifico Silva et al., 2018; Stilgoe et al., 2013), industry carry out similar practices under the names of Corporate Social Responsibility (CSR) or Creating Shared Value (CSV) (Dreyer et al., 2017). Moreover, Hoop, Pols and Romijn (2016) concluded that there are important limitations to implement RRI in innovation. They said that material barriers, power differences, strategic behavior, or lack of clear definition of responsibilities could hamper the process of responsible innovation.

This work aims at contributing to the research in the field of RRI in artificial intelligence by analyzing whether RRI can be used in the field of AI in healthcare or if a better approach might apply in certain cases. By triangulating the results from expert interviews and the literature, we plan to deliver recommendations to RRI researchers and to policymakers on how to define to what extent RRI can be effective for steering technology development in AI (healthcare applications), and in what cases it should be reframed or a different strategy should be used. Within that process, we also plan to identify whether RRI frameworks, such as the PRISMA roadmap, can be applied successfully in large companies. This roadmap has only been tested in startups and SMEs that tend to be more flexible and are more open to implement new practices (OECD, 2017). By carrying out a case study at Philips, we identify the potential shortcomings of the PRISMA roadmap and suggest recommendations for improvement.

1.3.2 Practical Relevance

Most of the research in RRI has been oriented to offer governance recommendations to policy-makers in the public sector (Pacifico Silva, et al., 2018; Sun & Medaglia, 2019). It is an important topic in the academic world and in policy circles, however, only very few companies are aware of the concept and have implemented ethical practices under the name of RRI (Dreyer et al., 2017; van de Poel et al., 2017). This research plans to give a more practical approach to the concept of RRI by coming up with recommendations on how to implement RRI practices in industry. The objective is to incorporate these actions into the CSR policies and business strategy of R&D teams working on the development of AI for healthcare applications within Philips.

1.3.3 Relevance for the Management of Technology Program

The MOT program aims at improving the quality of technology and innovation management by educating responsible decision-makers (TU Delft, 2019). In fact, one of the main goals of the program is to emphasize the importance of the connection between society and technology. Throughout the present work, we plan to analyze how the implementation of artificial intelligence in healthcare might affect society and how its development can be steered to ensure a responsible introduction. This approach extends from the usual management focus on economic benefits, to include societal and ethical considerations in the business approach. If these factors are considered early in the process of R&D, potential risks

could be avoided and the social acceptability of the technology will be strengthened (von Schomberg, 2012).

1.4 Research Design

For this research, an exploratory study has been proposed. According to Sekaran and Bougie (2013), exploratory research is appropriate when some initial facts are known, but more information is needed for developing a consistent theoretical framework. This situation applies to this case because there is an initial knowledge of the ethical concerns of the implementation of AI in healthcare, but there is no clarity on how to implement a framework of RRI to tackle those issues. In addition, there is no certainty whether RRI is useful to deal with all the potential ethical risks that AI might pose or if certain specific risks require a different approach.

Moreover, Verschuren & Dooreward (2010) argue that a qualitative case study makes a good fit with an exploratory approach. A case study allows for a comprehensive, holistic and in-depth research of a complex matter and focuses on gaining a proper understanding of the problem studied (Merriam, 2009). Additionally, the author completed a 6-month internship at Royal Philips, which represented an additional motivation to make a qualitative case study due to the opportunity to obtain useful information from interviews, observations, internal reports, and informal dialogues. These sources of information were valuable to get in-depth knowledge for this research. Consequently, the unit of study is Royal Philips, with an emphasis on the R&D department.

1.4.1 Data Collection Methods

The following methods were used to collect the data of this research:

Literature review: Information was gathered from articles, scientific books, thesis reports, etc. to get a better understanding of the main concepts of this study: Social acceptability, ethical concerns of AI in healthcare, RRI, and CSR . This initial step was relevant to get a glance of the state-of-the-art of the different fields in order to answer some of the sub-research questions.

Desk research: The objective of the desk research was to complement the information obtained in the literature review by searching in non-academic sources such as organizations' websites, business journals, and Philips' internal reports.

Expert Interviews: The primary data of this research was gathered by conducting semi-structured interviews with people working in Royal Philips. Due to the broad approach of this project, we interviewed employees working on AI and CSR teams. These interviews were structured by aligning a set of topics beforehand but aiming for a conversation more than a question-answer approach. Semi-structured interviews are the most common data collection method applied in exploratory research due to its flexibility and adaptability to change

(Sekaran & Bougie, 2013). Ten interviews were performed with a duration of 60 minutes each.

Results' feedback: Once the data obtained by using the collection methods was analyzed, two interviews with experts in the fields of AI and CSR within Philips were carried out to discuss the results and suggestions for improvement. Besides, the outcomes were socialized with members of the Design for Excellence (DfX) team within Philips to get some extra feedback regarding the methodology and the RRI actions proposed.

1.4.2 Research Phases

1) Literature review and desk research

In this initial phase, a comprehensive overview of the ethical issues for the acceptability of AI in healthcare is compiled. Besides, the concepts of social acceptability and RRI are defined and the RRI dimensions are introduced. The limitations of the RRI approach when applied in industrial settings are also presented. Moreover, the difference between RRI and CSR is explained. This information help us to formulate the questions for the semi-structured interviews.

2) Primary data gathering and analysis

Once the main concepts of the research have been understood, the next step is gathering primary data from semi-structured interviews in Royal Philips. People working on AI technologies and CSR are interviewed. These interviewees were selected because they were the experts dealing with the ethical issues of AI on a daily basis. On the one hand, the AI scientists have to think about the potential impacts that the products they create might have. On the other hand, the CSR officers have to make sure that the AI-enabled products developed in the company comply with the ethical and legal requirements established in external and internal regulations. The interviews are conducted to verify to what extent those practices are actually happening. In addition, the author had direct access to the names, roles, teams, and contact details of the people working in the company by using the internal network. This eased the process of selecting the experts and reaching out to them. The interviews were planned one month in advance to increase the probability of having a spot in the interviewee's schedules.

The qualitative data obtained from the interviews was organized, transcribed and analyzed. The interview transcripts were coded based on the methodology employed by Sun and Medaglia (2019) when studying the challenges for AI implementation in the public sector (see Section 3.3.1). The coding process was used to identify the main drivers and challenges for the introduction of AI in healthcare. Additionally, to identify the RRI actions already implemented by the company and the actions that still need to be developed. Finally, we established what AI-enabled products developed by Philips could be introduced in the market in the short-, medium-, and long-term. The detailed process can be found in Chapter 3, section 3.3.

3) Suggestions and validation

From the results of stage 2, we formulated potential actions to successfully implement the RRI dimensions into AI development for healthcare. The objective was to adapt the roadmap designed in the PRISMA Project to this specific case. This allowed us to propose a set of RRI actions to managers, scientists and other stakeholders. Subsequently, by triangulating the results of the interviews with the RRI literature, we identified the limitations of the RRI approach for the case of AI in healthcare, specifically when designing RRI actions to be implemented by companies. In the cases where we found that RRI could underperform, a different strategy was suggested.

Finally, we validated the recommendations. This validation was performed by conducting two interviews with experts working in AI and CSR, and by discussing the results with the members of the Design for Excellence (DfX) team within Philips. This process improved the objectivity of the research and provided further insights. The processed helped to identify the shortcomings of the PRISMA methodology when applied in large corporations.

The research's flow diagram is portrayed in Figure 2.

1.5 Thesis Overview

The thesis report is divided into 6 chapters. Chapter 1 introduced the research problem, objective and questions. Besides, the research design was briefly described. In Chapter 2, the literature review is presented. First, the concept of social acceptability is defined. Then, the ethical concerns of the introduction of AI in healthcare and the concept of RRI are explained. Moreover, the limitations of RRI when applied in industry are discussed. Chapter 3 describes the research methodology and introduces Royal Philips as our unit of study. Chapter 4 presents the results from the semi-structured interviews. Chapter 5 discusses the results obtained from the interviews and compare them with the literature. Finally, chapter 6 presents the conclusions of this thesis by answering the research questions, summarizing the contribution to research in the RRI field as well as stating the limitations and potential opportunities for future research.

Figure 1. Thesis Outline

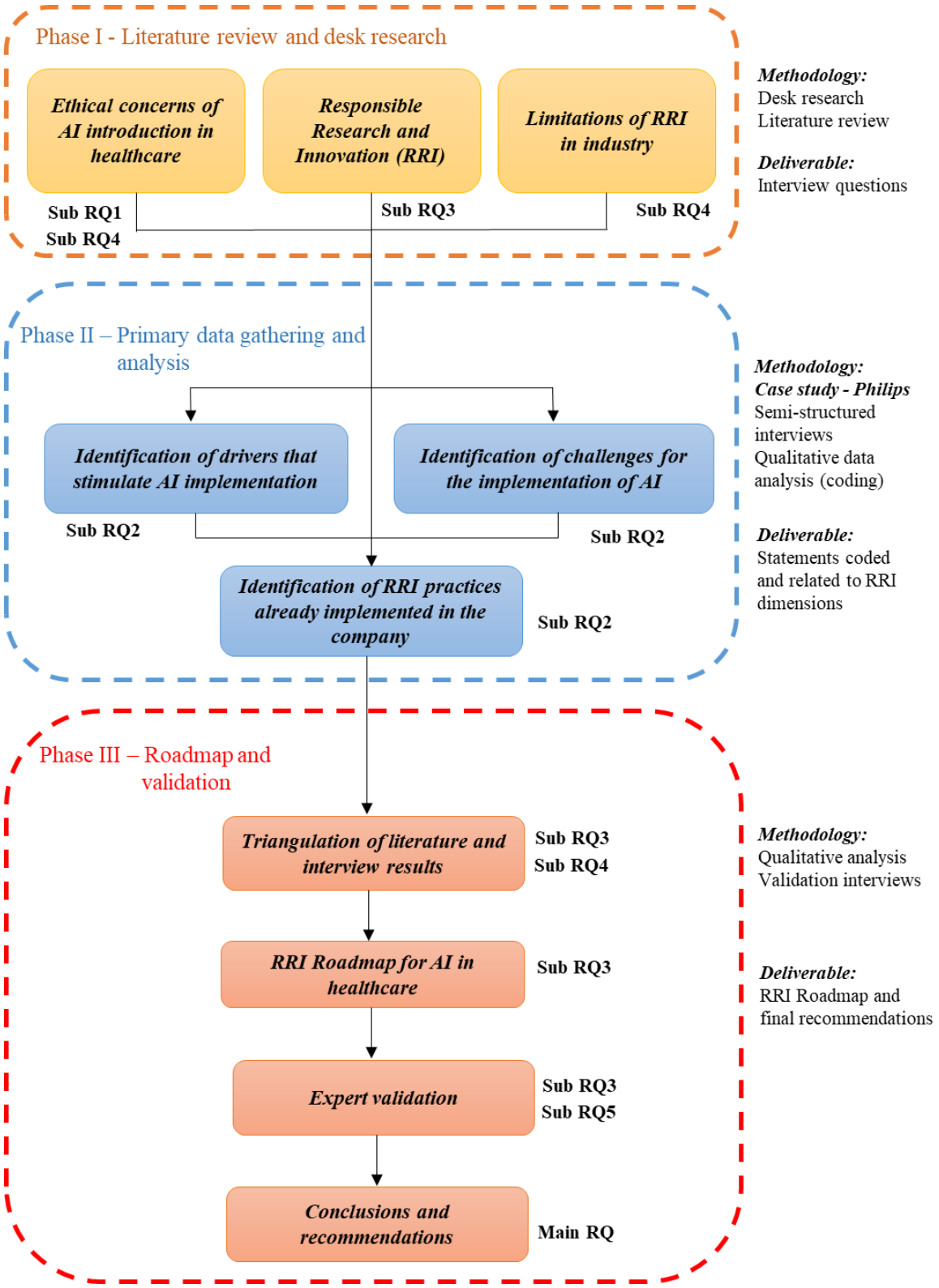

Figure 2. Research's flow diagram

2. Literature Review

The following sub-sections introduce the concept of social acceptability and the main ethical concerns of the application of AI in healthcare. Furthermore, the concept of Responsible Research and Innovation (RRI) is explained in more detail, as well as the limitations for the implementation of RRI in industry. Subsequently, the difference between RRI and Corporate Social Responsibility (CSR) is clarified. Finally, the RRI Roadmap developed by the PRISMA Project is introduced. The literature review is intended to study the state-of-the-art in the areas of RRI and AI ethics. The interview questions are formulated based on the outcomes of this section.

The literature review was done by searching key terms in three main databases: SpringerLink, ScienceDirect and Scopus. Once a relevant article was found (relevant if contained information related to the research questions of this research), the author proceeded with a process of backward snowballing by looking at the references of the paper at hand. These databases also offer a process of forward snowballing by suggesting related articles after one article has been downloaded. The articles suggested by the databases were also used for some parts of this research. When the search in databases was not satisfactory, we proceeded to do desk research starting by looking at the key term in search engines but with a focus on finding information in business journals and companies' reports. The process is explained in more detail in each subsection.

2.1 Defining Social Acceptability

One of the objectives of this research is to enhance the social acceptability of AI systems implemented in healthcare by using RRI. Therefore, it is important to define the concept of "social acceptability" before going further.

In order to define the concept of acceptability, the author started with a search in databases by the key term "Social acceptability". However, most of the results consisted of studies to improve the acceptability of a product or process without defining the concept first. The next step was to look for the term "Technological acceptability", which offered better results. After skimming several papers, most of them related to the work of Nielsen (1993), who established that a system or product has two kinds of acceptability: social and practical.

According to Nielsen (1993), social acceptability refers to the broader social context that surrounds the users. Culture, beliefs, interests, or even politics can influence the acceptability of a product. Individuals are challenged when their motivation to use a technology clashes with social restrictions. In that case, users decide whether to accept a technology by analyzing their surroundings and reflecting based on their existing knowledge (Nielsen, 1993). On the other hand, practical acceptability is related to the characteristics of the system

or product: usefulness, cost, reliability and compatibility (Nielsen, 1993). Nielsen finally defined acceptability as "whether the system or product is good enough to satisfy all the needs and requirements of the user". We modify this definition slightly to offer a more social meaning to the definition of a socially acceptable product:

"A socially acceptable product or system is the one that is good enough to satisfy all the needs and requirements of the user without affecting negatively the broader social context in which it is embedded".

2.2 Ethical Issues of Artificial Intelligence in Healthcare

In order to identify the main ethical issues that artificial intelligence might pose in the field of healthcare, the author developed a systematic search of key terms in ScienceDirect, Springer and Scopus. The key terms used were "AI in healthcare", "Challenges of AI in healthcare", "Ethical issues of AI", "Ethical issues of AI in healthcare", "Ethics in AI", "Risks of AI" and "Risks of AI in healthcare". Several papers were found directly and by checking the suggestions given by the databases. Additionally, these terms were also used to search in the web. Company's and associations' reports that described the ethical issues of AI in healthcare were also considered to complement the information obtained from the scientific articles. To select the most important challenges, we identified the ethical topics that were mentioned recurrently in the different articles or reports. Table 1 illustrates an overview of the literature used to define the ethical issues for this research.

Author	Title	Ethical issues mentioned
Academy of Medical Royal Colleges (2019)	*Artificial Intelligence in Healthcare*	• Patient safety • Doctor-patient relationship • Public acceptance and trust • Accountability • Bias, inequality and fairness • Data quality, consent and information governance • Impact on doctor's working lives • Impact on the wider healthcare system
Bartoletti (2019)	*AI in Healthcare: Ethical and Privacy Challenges*	• Privacy • Public trust • Transparency • Effect on healthcare professionals • Accountability
Becker (2019)	*Artificial intelligence in medicine: What is it doing for us today?*	• Privacy and security • Isolation of elderly patients • Bias and unfairness • Threat of doctor's replacement • Impact on doctor's working lives

Bhatnagar, et al. (2018)	*The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*	• Security threats • Privacy breaches
Bird, et al. (2016)	*Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI*	• Data bias, fairness and equity. • Privacy
Bostrom & Yudkowsky (2011)	*The Ethics of Artificial Intelligence*	• Unintended discrimination • Biases • Transparency • Accountability • Predictability • Robustness • Auditability • Incorruptibility
European Commission – High-level Expert Group on Artificial Intelligence (2018)	*Ethics Guidelines for Trustworthy AI*	• Human agency and oversight • Technical robustness and safety • Privacy and data governance • Transparency • Diversity, non-discrimination and fairness • Societal and environmental well-being • Accountability
Microsoft (2018)	*Healthcare, Artificial Intelligence, Data and Analytics*	• Fairness • Reliability and safety • Privacy and security • Inclusiveness • Transparency • Accountability
Mittelstadt, et al. (2016)	*The ethics of algorithms: Mapping the debate*	• Fairness, non-discrimination • Traceability • Opacity (“black-box”) • Transparency • Bias • Privacy • Autonomy • Moral responsibility
Nuffield Council of Bioethics (2018)	*AI in healthcare and research*	• Reliability and safety • Transparency and accountability • Data bias, fairness and equity • Consequences on patients • Consequences on healthcare professionals. • Trust • Privacy and security • Malicious use of AI

Table 1. Literature analyzed to define the ethical concerns of AI in healthcare

After analyzing the topics that appeared recurrently in the literature, we defined the following key ethical issues to be addressed in this work: 1) Accountability, 2) Data bias, fairness and equity, 3) Data privacy and security, 4) Effects on healthcare professionals (HCPs), 5) Effects on patients, 6) Reliability and safety and 7) Transparency and trust. It is important to note that in the literature shown in table 1, only the works from the European Commission, Bartoletti, and Bhatnagar offer recommendations to the private sector to mitigate the risks. The other articles and reports were focused on only mentioning the ethical concerns or on offering suggestions to policymakers. The *Ethics guidelines for trustworthy AI* (European Commission, 2018) presented an assessment list to evaluate if an AI product or system can be considered trustworthy. However, these guidelines are very general and do not provide clear actions to address the ethical concerns. For example, one section of the guidelines emphasize in the need of explicability of the machine learning algorithms, but it does not provide indications on how to actually achieve that explicability. The work of Bartoletti (2019) suggested the implementation of data privacy assessments, impact assessments and audit trails to build trust in the algorithms. Finally, the job of Bhatnagar, et al. (2018) recommended the implementation of strong cybersecurity algorithms that include machine learning technologies to avoid potential privacy breaches. Those articles offered valuable suggestions to deal with some of the ethical issues of AI but there is still no comprehensive framework aimed at addressing the wide-range of ethical concerns in healthcare with specific RRI actions.

There are many ethical and social concerns that appear when studying the application of AI in healthcare. Clinical practice normally requires complex judgments that demand human abilities unable to be replicated by AI algorithms, such as compassion or intuition (Parks, 2010). In fact, the claim that AI will be able to display 'autonomy' has been put into question by ethicists and social scientists stating that this is a human skill impossible to be reproduced by a machine (European Group on Ethics in Science and New Technologies, 2018). Besides, there are issues of privacy and safety that come into play with the vast amount of patient data required to train machine learning algorithms (Becker, 2019). For this reason, ethical challenges should be addressed jointly by all the relevant stakeholders to ensure the smooth and successful acceptability of AI (Bresnick, 2018). If technology developers fail to recognize the main values of the broad set of stakeholders, the social acceptability of the technology will be put in risk. The scope of the discussion of the possible implications of AI in healthcare is almost limitless. For that reason, we have bounded our research to the aspects that we considered more important. Below, the most salient ethical concerns are described:

1) Accountability

The main question here is: *Who should be held accountable if AI goes wrong?* This discussion is fundamental when reaching agreements between clinicians, healthcare organizations, policymakers and technology developers. AI is developing relentlessly and we will be faced with potential mistakes and unforeseen consequences (Academy of Medical Royal Colleges, 2019). Technology developers are currently focused on developing AI technologies that assist the clinicians but do not replace them, which implies that the

responsibility for decisions is still on the clinician (Hart, 2017a). However, doctors might fall in the 'automation bias', which is the human tendency to trust machines more than their own judgment, even if they are correct (Skitka, Mosier, & Burdick, 2000). In this case, complications can appear because clinicians may be justifying wrong diagnoses made by AI machines. These diagnoses might seem reliable in the first place and therefore doctors will not be willing to incur in the extra effort of double-checking the outcome. It would be difficult to hold somebody accountable under this kind of circumstances.

2) Data bias, fairness and equity

AI applications are designed with the aim of decreasing human biases and errors. However, if machine learning algorithms are trained with incomplete and unrepresentative data, AI could lead to discrimination based on race, gender, age, socioeconomic status, etc. (Bird, et al., 2016). For example, an algorithm called COMPAS was used in the US to help judges to decide sentences by predicting the risk that a suspect may re-offend. However, after analyzing the results it was found that the tool was biased against black individuals, classifying them as future criminals twice as much as for white individuals (Angwin, Larson, Mattu & Kirchner, 2016). Similarly, biases can be 'loaded' in the algorithms, reflecting the prejudices and beliefs of AI developers (Academy of Medical Royal Colleges, 2019). This could be challenged by ensuring diversity in the AI development teams (Crawford, 2016). Regarding equity challenges, AI health solutions will be more difficult to implement in places where data are limited or difficult to collect such as some African countries. Those populations will be underrepresented and therefore its medical concerns will not be addressed properly by AI developers (Hart, 2017b).

3) Data privacy and security

Medical data is considered to be sensitive and private (Nuffield Council of Bioethics, 2018). For that reason, proper measures have to be taken into account to ensure that AI systems cannot be hacked and that healthcare organizations or technology developers do not commercialize data. Furthermore, AI might be used to carry cyber-attacks in hospitals, causing huge harm at a minimal cost (Bhatnagar et al., 2018) and it is the responsibility of healthcare providers and technology developers to ensure that these risks are minimized. Besides, the discussion about ownership of data should be stressed in future regulation agendas (Academy of Medical Royal Colleges, 2019). Does healthcare data belong to the patient, the hospital, or the developer? Might ownership vary depending on the context? These questions should be answered to enhance the transparency of technology development in AI for healthcare.

4) Effects on healthcare professionals (HCPs)

Physicians may feel that their autonomy and expertise is threatened by AI systems (Hamid, 2016). Additionally, the skills needed to practice medicine will change significantly, leading to serious modifications in the training and education of clinicians. Especially in the field of radiology, where AI is making an important progress in image analysis that could generate a significant change in the role of the radiologist from a diagnosis-oriented approach

to a care-oriented approach (Academy of Medical Royal Colleges, 2019). Practitioners will be set free from routine tasks thanks to AI automation and they will be able to spend more time with the patients. However, this might also induce hospitals to employ less skilled staff as nurses supported by AI can replace some of the work currently done by the doctors (Nuffield Council of Bioethics, 2018). There is also a concern of doctors falling in the 'automation bias' explained before. Practitioners will be less likely to check the reliability of diagnosis from an AI algorithm that delivers 90+% of accuracy (Gretton, 2017).

5) Effects on patients

AI will lead to more autonomous healthcare in which apps could provide direct-to-patient advice. The development of telemedicine allows elderly patients to stay at home while following treatments that otherwise would have been done at care centers. However, this may lead to a loss of human contact that may cause psychological problems in an already vulnerable population (Sharkey & Sharkey, 2012). Besides, non-verbal cues that are commonly analyzed by doctors to investigate potential causes for a symptom will scarcely be interpreted by digital tools. For this reason, it is important to involve doctors in the process to ensure that values of quality, safety and patient support are maintained in AI apps (Academy of Medical Royal Colleges, 2019). Moreover, the autonomy of the patient must never be put at risk, including the right to make free and informed decisions about his own health. Apps that make it impossible for the user to verify whether he is dealing with a real person or with a machine should inform explicitly which the case is, and the patient should take the decision whether to follow the advice given or not (Mittelstadt, 2019).

6) Reliability and safety

Reliability and safety play an essential role when AI is introduced in treatment delivery, diagnosis support or prediction of diseases. This is due to the fact that AI could be wrong and harm patients across the healthcare system. For example, internal documents disclosed by IBM showed that its AI supercomputer IBM Watson for Oncology gave mislead treatment advice multiple times, putting in risk human lives (Ross, 2018). In a similar case, a clinical trial was organized to analyze the accuracy of an AI app able to predict what patients were likely to develop complications after suffering pneumonia (Caruana et al., 2015). The app indicated erroneously that patients with asthma should be sent home because their rate of survival was much higher than the rate of non-asthma patients. However, it did not take into account that patients with asthma were normally sent directly to the ICU and received better care than the other patients. This kind of context unawareness of the medical settings is one of the big limitations of the full deployment of AI tools for decision-support in healthcare. In fact, if AI algorithms are trained with insufficient or biased data and these issues are not detected, its introduction may cause harm at a great scale (Academy of Medical Royal Colleges, 2019). Finally, the performance of AI apps that offer health advice could be also questioned. For instance, these apps may be overly cautious, which could lead to an increase in the demand for health services that might be unnecessary (Frakt, 2016).

7) Transparency and trust

Machine learning algorithms are normally described as a 'black box'. The decisions and processes occur inside multiple layers of connections between 'neurons' that could be impossible to understand for the human brain (Anderson & Leigh, 2019). It makes it difficult to ensure the transparency of the system because the outputs and the logic behind them cannot be explained clearly. Then, it is difficult for physicians to justify the outcomes and for patients to trust the system. Besides, deep learning technologies are even more difficult to explain because they adapt and learn continuously, making generalization out of reach and complicating the identification of errors or biases. The EU General Data Protection Regulation (GDPR) states that individuals might be able to discredit decisions made only by artificial means and that in any case, the logic involved in data processing should be explained in detail (GDPR, 2018). However, this is still a grey area given that GDPR is a guideline that could be interpreted differently by any EU Country.

Additionally, there are uncertainties in the public about private companies accessing patient data. Some people argue that they do not trust AI technology developers to put the public interest over their financial rewards (Perrin & Mikhailov, 2017). It will just take a couple of news stories about failures of AI algorithms in healthcare to wipe out a reputation and trust built upon several years.

2.3 Responsible Research and Innovation (RRI)

To perform the literature review, we started by doing a search of the key terms "Responsible Research and Innovation" and "Responsible Innovation" in the databases. After skimming some of the papers that appeared in the search, we realized that all of them eventually referred to the works of Stilgoe, et al. (2013), von Schomberg (2012), and Owen et al. (2012) as the most important pieces of research in the field of RRI. For that reason, we based this section on the information from those articles.

Afterwards, the author used the key terms "Limitations of responsible innovation", "Responsible innovation in industry", "Responsible innovation in companies", and "Responsible innovation and CSR" to identify the state-of-the-art of the implementation of RRI in industry as well as its potential limitations. Eventually, we found the article of Martinuzzi, et al. (2018), who presented a very clear literature review of RRI in industry. The next step was to perform a process of backward snowballing based on the references mentioned in that article.

Finally, we searched the terms "Responsible innovation in healthcare", "Responsible innovation in health", "Responsible innovation in medicine", and "Responsible innovation of AI in healthcare" in the databases to evaluate whether RRI is being implemented in the field of healthcare. The results showed that some work has already been done in the physical branch of AI (robotics) (Stahl & Coeckelbergh, 2016; Van Wynsberghe, 2015) and that most

of the research has been oriented at offering recommendations to policymakers and healthcare providers on how to implement RRI (Batayeh, Artzberger, & Williams, 2018; Kerr, Hill, & Till, 2018; Pacifico Silva et al., 2018; Sun & Medaglia, 2019). Besides, the work of Auer and Jarmai (2017) was the only one that provided clear actions to companies in the medical field to implement RRI. They focused their investigation on SMEs and concluded that the companies analyzed should build upon existing CSR practices to further develop RRI in their business strategies.

However, after searching in the databases and doing a process of backward snowballing from the literature we could not found a comprehensive work that offered recommendations to large healthcare technology companies on how to implement RRI practices. This work is the first approximation to analyze how big corporations developing AI-enabled medical devices can implement RRI.

The following sub-sections explain what is meant by responsible research and innovation (RRI), its limitations, and its differentiation with the concept of corporate social responsibility (CSR). Besides, we present the PRISMA roadmap developed to introduce RRI actions in industry.

2.3.1 RRI Definition

The first significant definition of the concept of Responsible Research and Innovation (RRI) was given by von Schomberg (2012), who defined RRI as follows:

> "A transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view of the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)."

This definition is heavily oriented to ensure the societal acceptability and desirability of new products or technologies, which positively suits this research as our main goal is to embed RRI practices to achieve the social acceptability of AI in healthcare. However, reaching that objective also means making joint efforts to ensure the well-being of future generations that will be affected by AI. For that reason, the definition given by von Schomberg is complemented with the definition proposed by Stilgoe, et al. (2013):

> "Responsible Research and Innovation means taking care of the future through collective stewardship of science and innovation in the present".

This duty of care has to be collective in order to think what we want from innovation, and how to steer a responsible introduction faced by uncertainty. Responsible Research and Innovation (RRI) is about asking ourselves the kind of future that we want innovation to bring to the world (Owen et al., 2012). However, the introduction of new technologies frequently deals with the so-called 'Collingridge dilemma', which claims that the possibility to responsibly guide a technology is greatest in its early stages of development, when the

potential consequences are still unknown, whereas it becomes very difficult to steer once it is fully embedded in society and its effects have become manifest (Collingridge, 1980). For this reason, RRI approaches have to be well-timed in the sense that they have to be implemented early enough to be constructive but late enough to be meaningful (Stilgoe et al., 2013).

2.3.2 RRI Dimensions

In order to represent ethical and social concerns in research and innovation, Stilgoe et al. (2013) proposed a framework of RRI based on four dimensions:

1) Anticipation

The negative impacts of new technologies are often unforeseen and frequently estimations of harm have not worked properly to provide early alerts of potential adverse effects (Hoffmann-Riem & Wynne, 2002). Anticipation encourages scientists and innovators to ask "what if…" questions to analyze what is already known and what can we expect to happen in the future (Ravetz, 1997). However, anticipation is not only about the prediction of unintended consequences but also about steering desirable futures and organize the proper resources to achieve them (te Kulve & Rip, 2011). This process should be realistic to prevent overestimating the potential of the new technology by taking into account the complexity and uncertainty surrounding the co-evolution of science and society (Barben, Fisher, Selin, & Guston, 2008).

2) Reflexivity

Reflexivity means taking a closer look at each activity, commitment, and assumption to connect them with good practices of science and moral values. To succeed in this process, the researcher must be aware of his own knowledge limitations and that a particular framework might not be universally held (von Schomberg, 2012). An important characteristic of reflexivity is that it does not only refer to the self-critique of the researcher, but it is intended as an institutional practice or even a public matter (Wynne, 2011). By using mechanisms, such as codes of conduct, moratoriums or standards, connections between external values and scientific practice can be defined (Busch, 2011). In conclusion, reflexivity challenges scientists to go further than their role responsibilities and embed wider moral responsibilities in their daily job.

3) Inclusiveness

Including the public and diverse stakeholders in R&D processes since the very early stages of technology development to enable collective deliberation of the visions, purposes, questions, and impacts of innovation (von Schomberg, 2012). According to Callon et al. (2009), there are three criteria to assess the quality of the dialogue:

- Intensity – the way to interact with the members of the public.
- Openness – diversity in the group and fair representation.
- Quality – the depth and continuity of the discussion.

Furthermore, Grove-White et al. (2000) argue that dialogue should interrogate the 'social constitutions' embedded in technological products, such as the social, ethical and political consequences that innovation might bring.

4) Responsiveness

An ability to change the trajectory of technology development in response to stakeholder and public values as well as a changing environment (von Schomberg, 2012). It is important to note that shaping the course of action must be done while recognizing the lack of knowledge and control in innovation processes (Collingridge, 1980). According to Macnaghten and Chilvers (2014), there are mediating factors that can improve responsiveness:

- Deliberative science policy culture
- Emphasizing reflexive learning and responsiveness
- An open organizational culture
- Emphasizing innovation
- Creativity
- Interdisciplinarity
- Experimentation and risk-taking
- Top management leadership
- Commitment to public engagement and to take into account the public interest
- Commitment to openness and transparency

These four dimensions can be seen as the guiding principles that have to be analyzed and taken into account when developing new technologies. These principles are normally embedded by companies in professional codes of conduct and corporate social responsibility policies (Iatridis & Schroeder, 2015). However, these codes of ethics are frequently too general and weakly institutionalized which make them very difficult to implement in specific technologies or situations (van de Poel & Royakkers, 2011). Consequently, one of the aims of this research is to introduce the four dimensions of RRI in a specific framework for the concrete case of AI in healthcare.

2.3.3 Practical benefits of RRI

In a practical sense, the implementation of RRI practices in R&I can bring significant benefits for companies (TU Delft MOOC, 2019):

- Strengthening links with customers and end-users.
- Enhancing the company's reputation.
- Decreasing business risk and unintended consequences.
- Strengthening public trust in the safety of products.
- Increasing acceptability of products
- Adopting an environmentally friendly profile
- Enhancing the company's medium-term competitiveness/profitability

As can be seen, there are important reasons to establish RRI in industry. Taking into account responsible practices since the beginning of innovation will lead to competitive advantages in the medium and long-term. RRI enables companies to anticipate social and ethical issues and integrate them into the innovation and design processes and business strategy right from the start.

2.3.4 RRI Maturity level

A significant aspect of introducing an RRI strategy in a company is to assess the degree to which RRI practices are already implemented in CSR policies (van de Poel et al., 2017). For this reason, Yaghmaei (2016) proposed five successive stages of RRI implementation to evaluate how mature a company is in the development of responsible practices. The five stages are:

- ***Defensive*** – the company only reacts under critics of RRI aspects by the business environment in which it operates.
- ***Compliance*** – the company meets the legal requirements related to RRI, but does not extend its efforts from there.
- ***Managerial*** – RRI has been applied in different activities within the firm.
- ***Strategic*** – RRI is an important part of the business strategy of the company.
- ***Civil*** – the company converts in a role model that promotes RRI principles within the business environment and society.

This classification is useful to assess where a company stands and to reflect on what measures are necessary to go to the next stage.

2.3.5 Limitations of the RRI approach

The concept of RRI has been criticized by several authors due to its limitations when applied in industry. For example, Sonck, et al. (2017) argued that RRI is naïve, idealistic and unconcerned of the private sector characteristics. They claimed that RRI assumes a transparent and smooth engagement of stakeholders, where every member has the same information and the same decision power. However, the business world is characterized by taking risks based on power and information asymmetries. It makes it difficult to establish a fair ground for discussion (Sonck, et al, 2017). In a different study, de Hoop, et al. (2016) expressed that if not implemented adequately, RRI may end up becoming tool for 'greenwashing'. They argued that there are important limitations such as material barriers to innovation, unclear definition of responsibilities, power differences, strategic behavior, and conflicting interests that should be taken into account when introducing RRI practices. The authors claimed that not innovating could always be an outcome of RRI and proposed that that "*RI should be about innovating responsibly or not innovating at all*". Moreover, Dreyer, et al. (2017) expressed that there is not alignment between the terminology used in RRI and the concepts used in industry. While academics in the field of RRI formulate actions of anticipation, reflexivity, inclusiveness, and responsiveness to steer responsible technology development (Owen et al., 2012; Pacifico Silva et al., 2018; Stilgoe et al., 2013), industry carry out similar practices under the names of Corporate Social Responsibility (CSR),

Creating Shared Value (CSV), or ethical leadership (Dreyer et al., 2017). Similarly, Auer and Jarmai (2017) carried out interviews to analyze the drivers and barriers of introducing RRI in SMEs and concluded that despite all the interviewees were unaware of the concept of RRI, there were many RRI practices happening in their companies but under different names. This shows again that there is still a gap between industry and academy when defining the terminology, procedures, limitations and boundaries of RRI. In this work, we plan to analyze the effectiveness that RRI could have to enhance the social acceptability of AI-enabled products in healthcare. In the cases where RRI falls short, we will deliver suggestions to the RRI researchers to reframe the RRI approach in different terms or to use different strategies that could lead to better results.

2.3.6 Difference between CSR and RRI

Although Responsible Research and Innovation (RRI) has gained a lot of traction in the academic community and policy circles, most companies are unaware of the concept and therefore its implementation is still very limited (van de Poel et al., 2017). However, the practices encouraged by RRI are to some extent in industry under the name of Corporate Social Responsibility (CSR). For that reason, it is paramount to integrate RRI in CSR practices of companies to create a more robust approach for the social and ethical awareness of technology development (van de Poel et al., 2017).

The European Commission (2011) defined CSR as "the responsibility of enterprises for their impacts on society" and emphasized the need of CSR approaches to ensure stakeholder strategies aimed at addressing social, environmental and ethical issues in business strategies and operations. This concept has some similitudes with RRI but a broader scope. We can say that RRI is a sub-set under the bigger picture of CSR. Figure 3 shows that CSR applies for the entire business cycle of a product, while RRI is entirely based on the phase of research and innovation. CSR is intended to apply an ethical approach to company-wide business operations, while RRI is focused on the early phases of technology development.

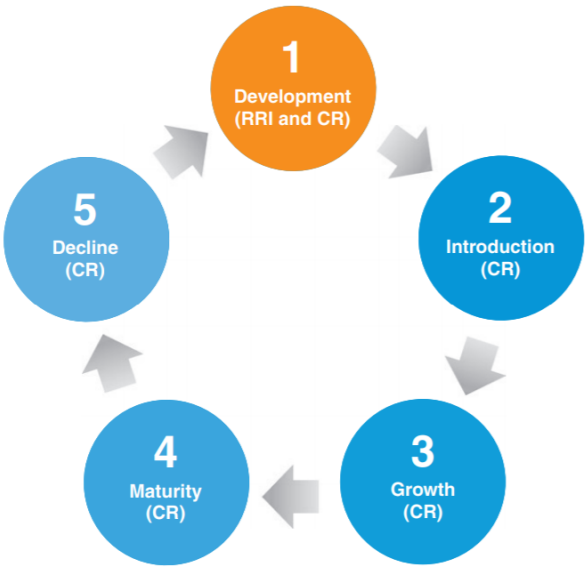

Figure 3. CSR and RRI. Involvement in the business cycle (Iatridis & Schroeder, 2015)

Iatridis and Schroeder (2015) have stated that a properly designed CSR strategy will represent an important competitive advantage for any company, as it will bring substantial business and social benefits. Figure 4 illustrates the expected benefits that CSR might provide.

Figure 4. Business and social benefits of CSR implementation (Iatridis & Schroeder, 2015)

However, despite the good intentions of CSR practices in companies. Research has shown that in most of the cases, industries only focus on regulatory compliance to develop their CSR strategy. They usually do not take a proactive role to enhance ethical and social development in the culture of the company (Groves, Frater, Lee, & Stokes, 2011). If RRI were to be included in CSR practices, companies would move from a reactive approach based on regulation towards a proactive approach in which ethical and moral issues would be discussed in more depth (Doorn & Nihlén Fahlquist, 2010).

CSR and RRI must be aligned with the corporate strategy of a firm to work, which sometimes represent a challenge as the common approach of corporations to reduce costs conflicts with the need for resources to design and deploy an RRI strategy. Promoters of RRI should look for areas where the company could be able to generate benefits for society while making profits (van de Poel et al., 2017). The PRISMA Project developed by a group of experts in RRI at the European Commission has established the first roadmap to introduce RRI practices in industry (PRISMA Project, 2019). In the following sub-section, the different parts of the roadmap are explained.

2.4 RRI Roadmap (PRISMA Project)

The main goal of the PRISMA Project is to support companies to introduce RRI practices in their R&I strategies to enhance the social acceptability of their products, thus creating a source of competitive advantage (PRISMA Project, 2019). This methodological approach is intended to be used in the context of disruptive technologies with the potential to transform society or change paradigms (van de Poel et al., 2017).

In the PRISMA Project, the RRI dimensions are embedded in the R&I strategies as a set of specific actions that will improve the acceptability of a product (Table 2).

Principles for RRI implementation	Action lines
Reflection & Anticipation	Integrate analysis of ethical, legal and social impacts since the early stages of product development
Inclusiveness	Perform stakeholder engagement to inform all phases of product development
Responsiveness	Integrate monitoring, learning and adaptive mechanisms to address public and social values and normative principles in product development

Table 2. Set of principles and actions for RRI Implementation (PRISMA Project, 2019)

The roadmap design includes a series of steps that are described below (PRISMA Project, 2019):

1) Top management commitment and leadership

Top management commitment is necessary (top-down approach), but it should be integrated with a bottom-up approach. Some of the activities are:

- Ensuring that the RRI roadmap, actions and vision are introduced and that they are compatible with the value of the different stakeholders.
- Ensuring that RRI practices are embedded in the company's operations and regulations.
- Ensuring that the resources to develop the RRI roadmap are available.
- Communicating the importance of RRI across the company.

This process leads to the setup of the initial vision for the RRI roadmap and the selection of potential R&I projects or products for which the roadmap will fit appropriately.

2) Context analysis

For an effective implementation of RRI practices, it is important to identify internal and external factors affecting the business operation:

- Ethical, social and legal impacts of the product to be developed. This helps to set the vision of the company towards RRI.
- The specific **technologies or products** in which the RRI roadmap will be focused (4th line of the roadmap).
- The expected time to market the product/technology (4th line of the roadmap).

- The stakeholders involved/affected by the development of the technology/product.

3) Materiality analysis

This step is essential to identify impacts in order to be able to make changes when necessary and ensure the optimization of benefits. The activities of this phase are:

- Identify **drivers** (creation of value, positive impacts) and **challenges** (organizational, technical, regulatory or ethical) to achieve the RRI vision. (1st line of the roadmap)
- Identify the **risk and barriers** (uncertainties) to consider in order to achieve the RRI vision. (2nd line of the roadmap)
- Select the **key stakeholders** within the innovation eco-system.
- Set an initial set of **RRI actions** to pursue. (3rd line of the roadmap)

A first complete version of the roadmap is prepared in this stage.

4) Experiment and engage

Stakeholder participation is key in for the effective development of RRI practices as it allows to validate the outcomes of the materiality analysis. In this phase, at least one inclusiveness action must be performed to evaluate the viability of the proposed roadmap.

5) Validation

In this stage, the company evaluates the validity and effectiveness of the RRI roadmap in terms of the impact on R&I processes of the organization. The feasibility of the roadmap is assessed and the potential refinements are put in place. The activities in this phase are:

- Identify what needs to be measured and monitored. Select criteria to perform an evaluation of the impacts of the RRI practices.
- Select the methods for measuring, monitoring and evaluating the impacts.
- Evaluate the impacts of the RRI actions.
- Identify until what extent the roadmap can be embedded in the current CSR policies of the company. Include KPIs to measure the impact of the RRI actions.

This step will lead to changes in some of the RRI actions proposed in the materiality analysis to ensure the alignment with the company's strategy and resources.

6) Roadmap design

The outcomes of the last five steps will lead to the final RRI roadmap, which is shown in Figure 5.

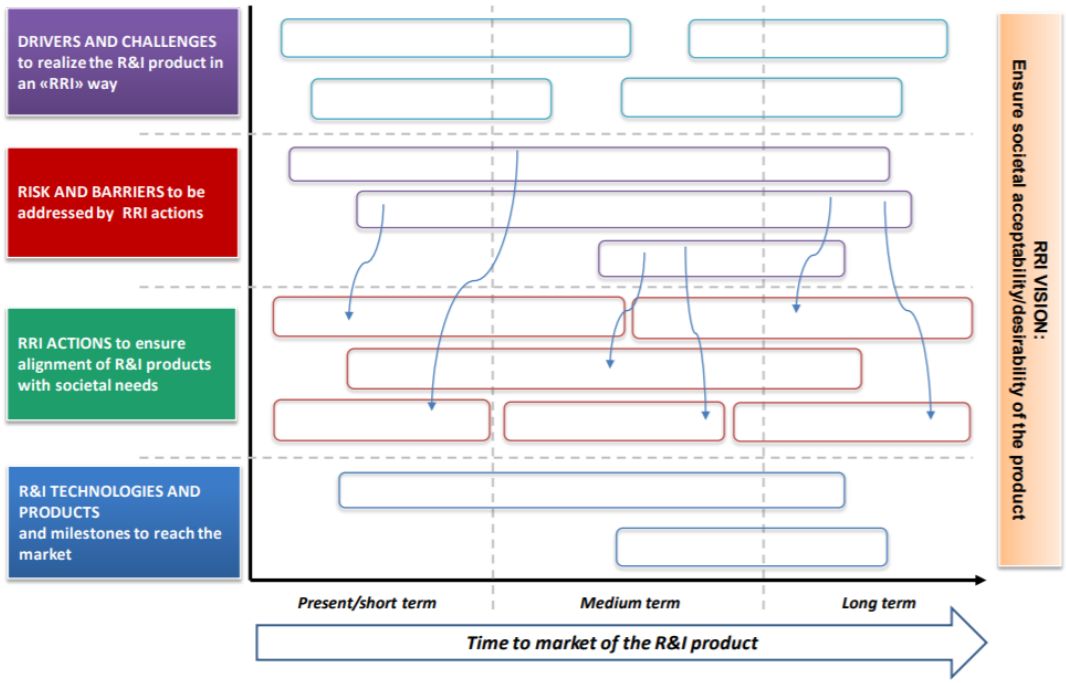

Figure 5. Visualization of the RRI roadmap (PRISMA Project, 2019)

Table 3 and Figure 6 summarize the steps to follow in the design of the RRI roadmap proposed by the PRISMA Project.

Step	Goal	Roadmap preparation
1. Top management commitment and leadership	Ensure endorsements of the organization toward RRI values and approach	Setting of the initial RRI vision, and selection of RRI product/technology candidates
2. Context analysis	Analyze the organization, the R&I products and technologies to focus on; identify ethical, social and legal impacts of the product and stakeholders of the product innovation eco-system	Compilation of the 4th line of the roadmap (R&I tech and products)
3. Materiality	Identify and prioritize: drivers and challenges to achieve the RRI vision; risks and barriers to overcome; stakeholders to work with; significant RRI actions to pursue	Compilation of the 1st and 2nd lines of the roadmap; refinement of the vision; first version of the RRI actions (3rd line). A first complete version of the roadmap is designed in this step.
4. Experiment and engage	Perform exploratory/pilot RRI actions, engaging with stakeholders to inform the RRI roadmap	Review of the overall roadmap with stakeholders
5. Validate	Evaluate the impact of the roadmap on both the product development and the organization (KPIs)	Review of RRI actions, in view of their technical, ethical, social, environmental, and economic impacts
6. Roadmap design	Consolidate and visualize the long-term RRI strategy	FINAL ROADMAP

Table 3. Methodological steps for an RRI roadmap design (PRISMA Project, 2019)

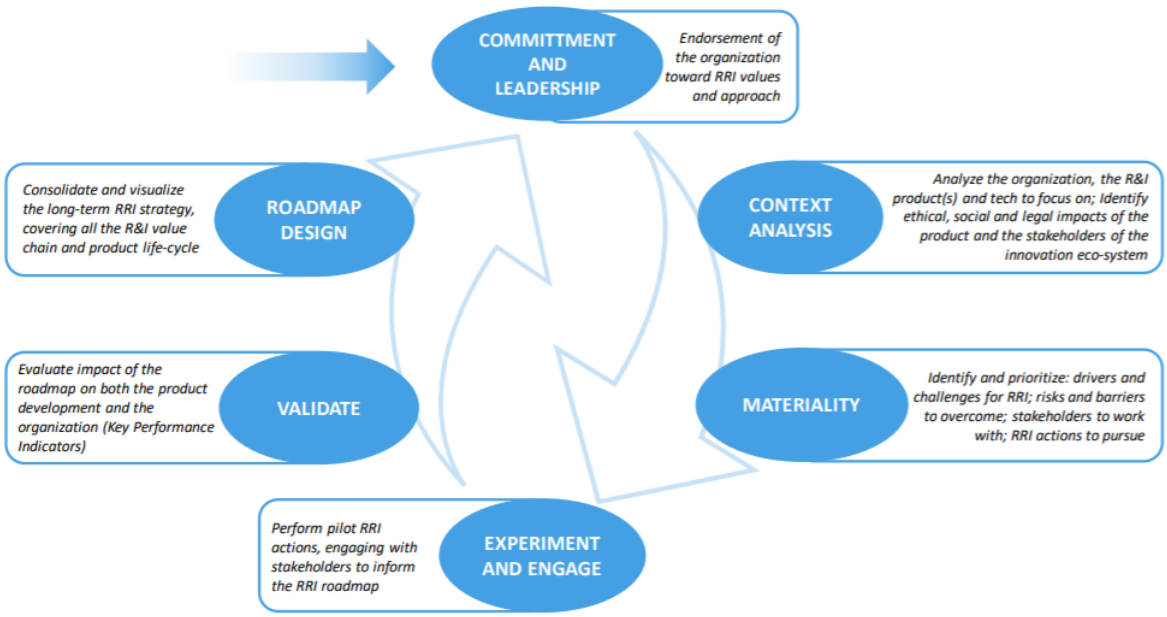

Figure 6. Steps to follow in the design of an RRI roadmap (PRISMA Project, 2019)

Note: Due to time and logistic constraints, this thesis only covers the first three steps of the methodological approach, which will lead to an initial set of RRI actions and an initial version of the RRI roadmap.

2.5 Literature Review Overview

There are important ethical concerns in the field of AI in healthcare. For example, there are risks of privacy breaches or unintended use of patient data, which is considered sensitive and private. Accountability is another issue as there is no clarity about who will be held responsible for an AI-induced clinical mistake. Furthermore, machine learning algorithms represent a 'black-box' impossible to be understood by the human brain, which may lead physicians and patients to distrust the outcomes of an AI decision that cannot be properly explained. Besides, if there is no quality and diversity in the data used to train the algorithms, unfair biases may appear that could lead to discrimination based on race, sex, age, etc. However, although these risks are continuously mentioned in the literature (Academy of Medical Royal Colleges, 2019; Nuffield Council of Bioethics, 2018), there is no study that illustrates how to address them in a practical manner in large companies. Most of the studies are concentrated in suggesting policy and regulatory solutions (Batayeh et al., 2018; Kerr et al., 2018; Pacifico Silva et al., 2018; Sun & Medaglia, 2019), or in offering recommendations to SMEs (Auer & Jarmai, 2017). However, there are no studies that offer suggestions to large technology companies working in the field of healthcare.

Responsible Research and Innovation (RRI) is an approach intended to address ethical issues since the first steps of technology development, specifically in the phase of R&I. The idea is to ensure that the views of the different stakeholders related or affected by the technology are included and that the potential impacts are predicted. This will lead to actions

aimed at steering the path of innovation in a responsible way to ensure the development of socially acceptable products (von Schomberg, 2012). According to Stilgoe, et al. (2013), the introduction of RRI practices should take into account four dimensions: Anticipation, Reflexivity, Inclusiveness, and Responsiveness. These dimensions represent the aspects in which the line of RRI actions must be based on. However, RRI has only gained traction in policy and academic circles, and its introduction in industry is still incipient (van de Poel et al., 2017). Some authors also argue that RRI has important limitations such as a lack of understanding of the private sector characteristics (de Hoop, et al., 2016; Sonck, at al., 2017), or misalignment between the terminology used in RRI and the concepts employed in industry (Dreyer et al., 2017).

The PRISMA Project came up with the first approach to support all types of technology companies to introduce RRI practices in their R&I strategies and enhance the social acceptability of their products. (PRISMA Project, 2019). By means of an RRI roadmap that analyzes impacts, challenges, risks, and barriers for the introduction of new technologies/products, RRI actions can be designed. These actions aim at covering the four dimensions of RRI and at ensuring the social acceptability of the innovative technology/product. This methodology has already been tested in eight pilot projects covering startups and SMEs, but there is still to be seen if it applies to large corporations with CSR practices already in place.

From the literature review, we can conclude that there is no comprehensive study that covers how to address the ethical risks of the implementation of AI in healthcare. Besides, there is no certainty of the suitability of RRI frameworks, such as the PRISMA roadmap, when applied in large companies. In the following chapters, we will go into these yet unexplored areas by carrying out an exploratory case study at Royal Philips to get an understanding on how to introduce RRI practices to enhance the social acceptability of AI in healthcare and to identify in what situations a different approach could be better. In the next chapter, the research methodology to achieve these goals will be described.

3. Research Methodology

As stated in Chapter 1, this research is an exploratory study based on a qualitative approach. According to Sekaran and Bougie (2013), exploratory research is undertaken when some initial facts of a specific issue are recognized but more information is needed to develop a consistent theoretical framework. In this case, we know the main ethical concerns of the implementation of AI in healthcare, but there is still no comprehensive study offering suggestions on how to deal with those issues. For that reason, this work aims at exploring how RRI practices can help to tackle ethical concerns and to validate whether the PRISMA roadmap represents a good strategy to introduce an RRI strategy in large companies.

Verschuren & Dooreward (2010) argue that a qualitative case study is appropriate when working on exploratory research. They specify that this kind of approach should concentrate on a small number of research units that allow focusing on deep intensive data generation. Similarly, Merriam (2009) states that a case study encourages the researcher to get a comprehensive, holistic and in-depth research of a complex situation. For these reasons, this research is focused on the case of Philips, with a specific emphasis on the R&D department. Philips is one of the world leaders in healthcare medical devices (Ellis, 2019) and the author completed a 6-month internship in the company. This situation offered an opportunity for the author to get insights from interviews, informal dialogues, observations, internal reports, and day-to-day activities in the organization.

The following sub-sections present a description of the unit of study for this project (Royal Philips), the data collection methods formulated for the research and the data analysis process established after gathering the qualitative data.

3.1 Unit of Study – Royal Philips

Royal Philips is a leading health technology multinational company established in Eindhoven in 1891 by Gerard and Frederik Philips (Philips, 2019a). The company started in the lightning business but has shifted its focus to concentrate on healthcare. Philips' *mission* is 'to improve people's lives through meaningful innovation', with the *vision* of 'improving the lives of 3 billion people a year by 2030' (Philips, 2019b). In 2018, the company sold EUR 18.1 billion and it currently has around 77,000 employees working in more than 100 countries (Philips, 2019c). The headquarters are located in Amsterdam and Philips is a leader in diagnostic imaging, image-guided therapy, patient monitoring, healthcare informatics, and personal health (Philips, 2019c). Figure 7 shows the four segments in which the company operates. The strategic focus is based on generating constant innovations to deliver on the Quadruple Aim of value-based healthcare: improved patient experience, better health outcomes, improved staff experience, and lower cost of care (Philips, 2019c).

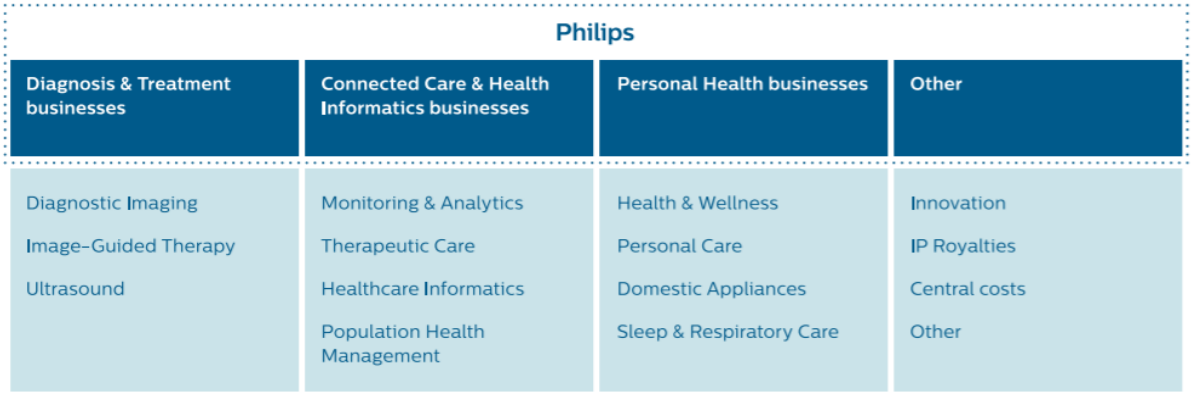

Figure 7. Philips business segments (Philips, 2019c)

Philips sees an important value in the concept of 'integrated healthcare', applying techniques of data analytics and artificial intelligence to optimize health delivery across the health continuum (Philips, 2019c). However, the company prefers to talk about "adaptive intelligence", which refers to the use of AI to help analyzing large quantities of data to generate outcomes that support and empower people. Adaptive intelligence combines the power of AI with the contextual knowledge of Philips in the clinical domain (Philips, 2018).

The interest of Philips in the field of AI has increased considerably in the last few years. In fact, Philips is the second patent filer in AI for healthcare in the world and 36% of all of its researchers worldwide are working in data science and AI (Philips, 2019d). Below, the most important AI Philips products are presented:

- ***Philips IntelliSpace Console***

This system is a decision support dashboard that uses adaptive intelligence to extract data from electronic health records and patient devices, to give clinicians a dynamic display's of a patient's condition, using organ body system view (Philips, 2019e).

- ***Philips Intellispace PACS with Illumeo***

Intellispace Portal is an advanced visualization and analysis software solution that gives radiologists access to images and information anywhere they need it, including access to prior studies of patients within the network. This version uses machine learning to gather data from the clinician's workflow to predict use patterns and pre-process patient data before the information or images are opened. It provides the radiologist with the most relevant case-related information and toolsets so that they can pinpoint quickly the regions of interest and critical findings (Philips, 2019e).

- ***Philips HealthSuite Insights***

HealthSuite Insights is a platform that gives data scientists, software developers, physicians and healthcare institutions access to advanced analytic tools to curate and analyze patient data. It offers them the option to build, maintain, deploy and scale AI-based solutions (Philips, 2019e).

- ***PerformanceBridge***

This tool is a management software that provides healthcare professionals with real-time data on department performance through an interactive dashboard. It aims at improving productivity, patient experience, and value-based care. PerformanceBridge helps radiology departments to prioritize resources, peer-to-peer collaboration and to optimize administrative practices (Philips, 2019e).

- ***CareSage***

CareSage is a predictive analytics engine which helps to provide a view into the home by collecting and analyzing data from multiple sources. Hospitals can monitor elderly or chronic patients and assess the risks remotely. It can avoid unnecessary hospitalizations and offer cheaper treatment at home (Philips, 2019e).

- ***Wellcentive***

Adaptive intelligence is used to enable clinicians to discover patterns in public health and predict care needs for entire populations. Wellcentive is a cloud-based platform that analyzes clinical data from a particular place and offer suggestions based on the conditions identified for the population of that place. For example, if almost everybody in that place has developed diabetes, the system indicates that there is a high risk of this disease for the new generations (Philips, 2019e).

- ***Intellispace Precision Medicine***

This decision-support system brings patient and medical data together, to provide a clear and comprehensive view of patient status that facilitates clinical decisions. The data comes from EHRs, lab systems, pathology, and genomics. There is an initial version in the market, but the system is being improved continuously (Philips, 2019e).

- ***Philips Ingenia MRI machines***

Philips Ingenia is a series of MRI machines produced by Philips. Some of them offer a faster image acquisition time by using ML algorithms that require fewer data to generate MRI scans. The upcoming versions are expected to reduce the acquisition time from approx. 20 minutes to 5 minutes (Philips, 2019e)

- ***Shaver Series 7000***

This shaver offers real-time guidance and coaching for optimal shaving results. It is connected to an app that gathers data from your pattern of shaving, skin type and shaving frequency so that you can avoid harming your skin (Philips, 2019f).

- ***Sonicare DiamondClean Smart***

This toothbrush is connected to an app. By means of a sensor that detects the cleanliness of your teeth, it gives real-time feedback to help you reach the areas that are not yet clean. It also monitors your improvement in oral healthcare day by day (Philips, 2019f).

- ***Pregnancy+***

This is an app that allows mothers to check the progress of their pregnancy and that shows 3D fetal images of the development of the baby. It helps future mothers to track their health during the pregnancy period and offer suggestions on how to deal with potential problems (Philips, 2019f).

- ***SmartSleep Analyzer***

This is an app that tracks your sleep pattern and provides personalized feedback to improve your sleep. This is the only clinically proven sleep app (Philips, 2019f).

- ***Digital Twin concept***

Digital twins are detailed models of anatomy, physiology, and pathology. This technology makes it possible to get a virtual representation of the different organs of your body. In that case, clinicians could run simulations of the effects of medication in your virtual body and gather conclusion from the results. For example, instead of an ultrasound of your heart, physicians will have a 3D model of your heart that can help them to come up with potential treatments or to have more clarity on how to proceed with surgery (Miskinis, 2018). This technology is still a concept in Philips.

As can be seen, Philips is investing heavily in AI development for healthcare and covering several business areas. However, the company is still struggling to address the ethical concerns of the implementation of these intelligent systems. There are ethical and data management codes that have to be developed for the field of AI in order to gain the trust of the public (Philips, 2019d). In that sense, this research offers a practical approach to deal with those topics by means of an RRI roadmap that can be used by the company to analyze the viability of their future AI developments.

3.2 Data Collection Methods

Qualitative research is a very open and unstructured methodology that usually consists of interviews, literature research and surveys (Smith, 2015). For that reason, and given the exploratory approach of this work, this research consists of an extensive literature review and desk research (Chapter 2), followed by a set of semi-structured interviews with experts in AI and CSR in healthcare from Philips. We developed a process of methodological triangulation (Denzin, 1973) to improve the reliability of the results. In Chapter 5, the outcomes from the semi-structured interviews are compared to the results of the literature to establish if there is consistency in some of the topics or if there are important differences that require further reflection. This process is also aimed at mitigating the "selection bias" that could be present in the research due to the fact that all the interviewees were employees of Philips.

3.2.1 Literature Review and Desk Research

The literature review and desk research were intended to get secondary data information to identify the key ethical concerns of AI in healthcare and to define the concepts of social acceptability and RRI. The objective of this phase was to analyze the current state-of-the-art of RRI in order to formulate the questions for the semi-structured interviews of the case study. Information was collected from scientific articles, scientific books, organization's websites and business journals by using key terms in academic databases and by following processes of backward- and forward-snowballing (see Chapter 2 for further details).

3.2.2 Expert Interviews

The primary data for this work was gathered by conducting semi-structured interviews with people working at Philips. Due to the exploratory nature of this research, the best option was to carry out interviews without a rigid structure, as this approach allows for flexibility and adaptability to change (Sekaran & Bougie, 2013). In this case, a set of topics were aligned beforehand to discuss in a fluid conversation instead of a fixed question-answer approach. This was done because according to Schutt and Chambliss (2013), spontaneous statements are easier to obtain under these conditions and normally they reflect what the interviewee really thinks or believes. The topics to be addressed varied depending on the path of the conversation and the interests and expertise of the interviewee.

There were three main stages during the interview. First, we asked the interviewee about the main drivers and challenges for the development of AI in healthcare. Besides, we asked them what AI technologies could be implemented in the short- and medium-term without generating social and ethical risks and what technologies still require a long time to appear in the market. Second, we talked about the main ethical concerns of AI in healthcare and how Philips is dealing with them. Third, we framed some questions to evaluate until what extent Philips has introduced practices of responsible research and innovation in its operations. The detailed interview protocol can be found in Appendix 1.

The experts came from the research teams of AI in Philips and from the ICBE, which stands for Internal Committee for Biomedical Experiments. This committee is in charge of ensuring that every product developed in Philips Research complies with the ethical, legal and privacy policies of the company. This is a regulatory board that ensures that all the risks are covered before bringing a product to the market. To ensure diversity in the answers, we made sure that the experts were working in different projects and departments within the organization. Besides, we contacted people in all the position-levels. From entry-level employees to senior managers. Table 4 illustrates the position, years of experience and expertise of the interviewees.

Interviewee	Position	Years of Experience	Expertise
A	Chair of the Internal Committee for Biomedical Experiments (ICBE)	26	CSR
B	Senior AI Scientist at Philips/ Professor at TUe	10	AI
C	Senior Director, Program Lead AI for IGT (Image Guided Therapy)	23	AI
D	Project Manager, Chief Architecture Office/ Surgeon	5	CSR and AI
E	Research Scientist in Artificial Intelligence	5	AI
F	Head of the Center of Expertise in Data Science and AI	28	AI
G	Head of Data Strategy and Artificial Intelligence	32	CSR and AI
H	Research Scientist in Artificial Intelligence	4	AI
I	Principal AI Scientist at Philips/ Professor at TUe	34	AI
J	Senior Scientist Workflow Innovation for Healthcare	31	CSR and AI

Table 4. List of Interviewees

Each interview lasted approximately one hour and all of them were conducted face-to-face. The experts were asked to consent the recording of the interview in audio files and all of them approved. Finally, the interviews were transcribed and analyzed. Section 3.3 presents the data analysis process in detail.

3.2.3 Results' Feedback

As a final step of this research. We conducted two validation interviews to verify the quality of the outcomes of this project. The goal was to validate two points. First, we aimed at validating the suggestions made to implement practices of RRI for the development of AI products within Philips. Second, we wanted to validate the suitability of the PRISMA roadmap when applied in large corporations. These interviews were based on a conversation in which the methodology and main results were exposed and the expert gave constant feedback for improvement. We wanted to have recommendations from two different standpoints. For that reason, one expert was the Head of the Center of Expertise in Data Science and AI at Philips, and the other expert was the Chairman of the Internal Committee of Biomedical Experiments in the company. While the first interviewee is a scientist in the field of AI, the second interviewee is the head of the team checking the ethical, legal and privacy compliance of Philips products. The conversations took approximately one hour and the suggestions were used to improve the RRI actions established for AI in healthcare. Furthermore, we also got recommendations to refine the PRISMA roadmap. Similarly, the outcomes were discussed with members of the DfX team within Philips, who offered additional comments to improve the quality of the research.

3.3 Data Analysis

The qualitative data obtained from the interviews was transcribed, coded, organized, and analyzed. There was a risk that the interviewees came out with socially desirable answers and that they were afraid to criticize the practices carried out by Philips. For those reasons, we emphasized at the beginning of the interview that their names will not be disclosed and that it would be useful if they could take a critical approach. Besides, some questions were intended to get suggestions for improvement in Philips practices in order to find the RRI areas in

which the company was underperforming. However, there is no fully-proved strategy to verify whether the interviewee is giving socially desirable answers or not. For that reason, the results of the expert interviews are triangulated with the outcomes of the literature in Chapter 5. That process aims at increasing the reliability of the results. In the following sub-sections, we explain the methodology used to code the interviews, how we built the RRI roadmap and how the conclusions were drew.

3.3.1 Coding of interviews

In order to code the results from the interviews, we employed a similar methodology as the one used by Sun and Medaglia (2019). They divided the process of coding in two rounds. In the first round, they summarized the statements made by the experts in a few words. For example, if the interviewee talked during one minute about how difficult was to develop AI research with the restrictions of the GDPR, the first round of coding summarized it as 'Regulation is very strict and perceived as too stringent'. In the second round, the first order codes were re-grouped into codes that are more general. For instance, following the same example of 'Regulation is very strict and perceived as too stringent', the second-order coding summarizes that as a 'Regulatory challenge'. The second-order coding differed for the drivers, challenges, RRI actions, and future technologies/products. Below, we explain the process for each part:

- ***Drivers:*** the codes of the first round were re-grouped into technical, organizational, social or economic drivers depending on the nature of the statement (see Table 5).

Interviewee	Empirical data	First-order coding	Second-order coding
A	"Within Philips there is a driver to understand what patient data is telling us in order to make more efficient diagnosis"	Improve technological efficiency	Technical driver

Table 5. Example of coding for drivers

- ***Challenges:*** the coding process was divided into non-ethical challenges and ethical challenges. For the non-ethical challenges, the second round categorized them as regulatory, technical or organizational challenges (see Table 6). On the other hand, the ethical challenges were coded in the second round according to the ethical principle that they might affect (see Table 7):
    - Accountability
    - Data bias, fairness and equity
    - Data privacy and security
    - Effects on healthcare professionals (HCPs)
    - Effects on patients
    - Reliability and safety
    - Transparency and trust

Interviewee	Empirical data	First-order coding	Second-order coding
A	"In a big company, it is always hard to change direction. There are many people working on staff and to coordinate them is not an easy task"	Lack of flexibility	Organizational challenge

Table 6. Example of coding for non-ethical challenges

Interviewee	Empirical data	First-order coding	Second-order coding
D	"The biggest challenge is to understand what the bias is of the AI we develop. It is really difficult to implement contextual information in the algorithm"	Presence of biased data	Data bias, fairness and equity

Table 7. Example of coding for ethical challenges

- ***RRI actions*:** from the interviews, it was possible to get answers about the RRI actions already developed in Philips, and actions to be done or reinforced. In the first round of coding, those actions were summarized, and in the second round, we established to which RRI dimension the action corresponded: anticipation & reflexivity, inclusiveness or responsiveness (see Table 8).

Interviewee	Empirical data	First-order coding	Second-order coding
D	"Philips has a lot of co-creation. We put together hospitals, patients, former companies and other device companies together in a room and we ask the following question: What kind of product would be nice for you?"	Set co-creation exercises with a wide range of stakeholders	Inclusiveness

Table 8. Example of coding for RRI actions

- ***R&I Technologies/products*:** For this part, we identified the type of technology/product in the first round and then we proceeded to identify if the responsible introduction of the technology will take place in the short-, medium-, or long-term (see Table 9).

Interviewee	Empirical data	First-order coding	Second-order coding
B	"If we talk about consumer applications, such as baby monitors, we are talking about short-term introduction as you can quickly put it in the market"	Consumer applications - Monitoring devices	Short-term

Table 9. Example of coding for technologies/products

Note: The complete coding tables for each part of the RRI roadmap can be found in Appendix 2.

3.3.2 Building the RRI roadmap

As established in the literature review, this thesis only covers the first three steps of the PRISMA Project approach: Commitment and leadership, context analysis, and materiality. In these stages, the drivers and challenges for the implementation of the new technology are identified, the risk and barriers to overcome are indicated, a first set of RRI actions to ensure the social acceptability of the technology are designed, and the main products that could be

introduced following the RRI actions are defined. This leads to an initial version of the RRI roadmap.

The presentation of the roadmap is divided into two parts: Case description and RRI roadmap. The contents of each part are described below:

- Case description
  - The company
  - RRI commitment
  - Context
  - Materiality & experimentation

- RRI roadmap
  - RRI vision
  - R&I Technologies and products
  - Drivers and challenges for RRI
  - Risk and barriers to be addressed by RRI actions
  - RRI actions
  - Roadmap design

For the RRI roadmap design, the following activities were developed in each part:

- ***Drivers***
  The main drivers for the implementation of AI in healthcare were defined from the coding of the interviews. They were divided into social, technical, organizational, and economic drivers. Additionally, the specific drivers exclusive to Philips were identified (i.e. reach 3 billion people a year by 2030).

- ***Challenges***
  From the coding of the interviews, we identified ethical and non-ethical challenges. There are organizational, regulatory, technical and ethical issues that are further discussed in Chapter 4.

- ***Risks***
  Once the ethical challenges for the implementation of AI were defined, this information was complemented with content from the literature review to identify the risks that the challenges represent. For example, the presence of biased data leads to a risk of unintended discrimination. The risks were divided into technical, ethical and organizational.

- ***Barriers***
  From the coding procedure to identify organizational, regulatory and technical challenges, we could define the barriers for the introduction of AI. For example, strict regulation or lack of flexibility in a large company such as Philips.

- ***RRI actions***
  From the interviews, we identified the RRI practices that are already being developed by Philips and the actions that need further attention. There are many responsible practices within the CSR policies of the company. For that reason, we asked the interviewees to select from those practices the ones that are more critical to enhance the social acceptability of AI products. Each RRI action mentions the potential benefits, the corporate functions involved to make it happen, and the stakeholders that should be part of the action.

- ***R&I technologies/products***
  The interviews gave us information about the type of technologies and products that can be introduced in the short-, medium-, and long-term. We will complement that information by classifying the different products developed by Philips in those categories.

Finally, we can proceed to design the initial version of the RRI roadmap for the implementation of AI Philips products in healthcare.

3.3.3 Drawing conclusions

The final step of the data analysis process is to draw conclusions based on the data gathered from the semi-structured interviews and the literature research. This is a crucial part of qualitative research because by triangulating the results from the different sources the author can establish the consistency and potential differences, reflect upon them, and deliver recommendations based on the analysis (Denzin, 1973). We answer whether the entire process for RRI implementation based on the PRISMA Project works for the case of AI in healthcare in order to enhance the social acceptability of AI products. The RRI actions are validated with experts from Philips and the potential recommendations for improving the methodology are mentioned. Besides, we identify the contribution to the scientific and practical fields that can be derived from the research. Furthermore, the limitations of the project, the recommendations for future work, and the personal reflection of the author are presented.

Chapter 4 shows the results obtained after following the research methodology described in this chapter.

4. Results

This chapter illustrates the results obtained and derived from the coding of the expert interviews (see Appendix 2). In the following sub-sections, the results for the main components of the PRISMA roadmap are presented: Drivers, challenges, risks, and barriers for the implementation of AI in healthcare; RRI actions to address the main issues in a responsible way; and the technologies/products whose development will be influenced by the introduction of RRI practices.

4.1 Drivers

The drivers were divided into economic, organizational, social, and technical categories (see Figure 8). The criterion to include a driver is that it had to be mentioned at least for one interviewee. Next to the drivers, the number of interviewees that mentioned them is presented in brackets.

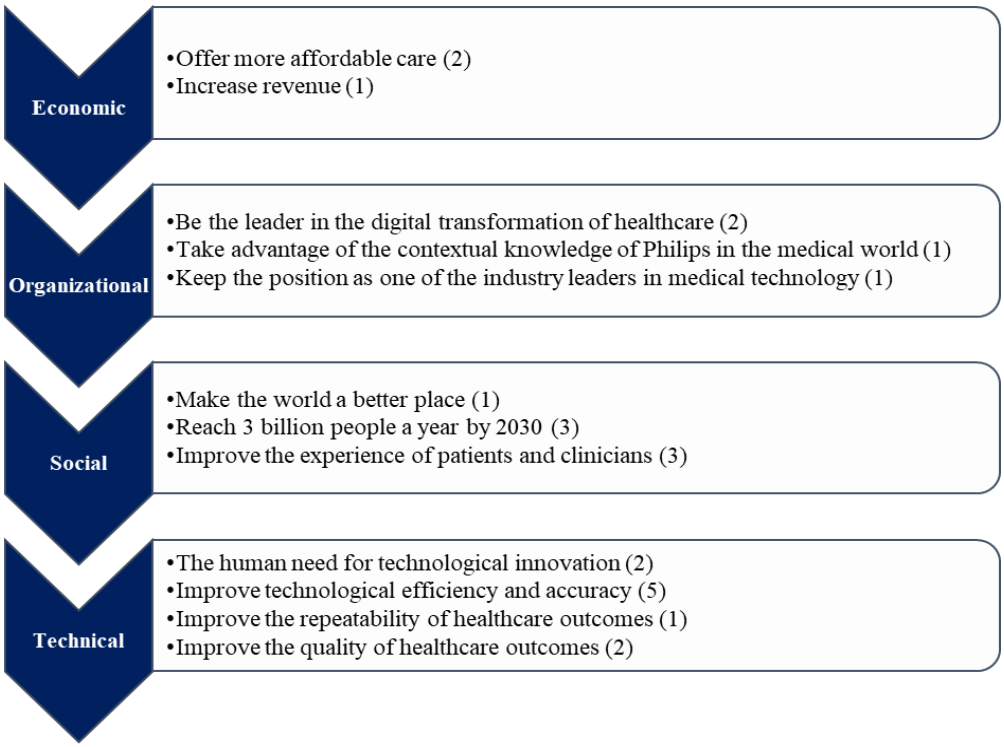

Figure 8. Drivers for the implementation of AI in healthcare

4.1.1 Economic drivers

- ***Offer more affordable care***
  Healthcare is expensive and represents an important expenditure for governments. By introducing AI in healthcare providers operations, the costs can be downsized. This can be translated into less cost for patients.

- ***Increase revenue***
  By reaching more people with AI, Philips expects to increase its revenue.

4.1.2 Organizational drivers

- ***Be the leader in the digital transformation of healthcare***
  Philips wants to disrupt the healthcare industry and to be the leader in the digital transformation of healthcare.

- ***Take advantage of the contextual knowledge of Philips in the medical world***
  Philips is excellent at processing data and at combining it with the contextual knowledge from the clinical field. This important competitive advantage can be further exploited by using AI.

- ***Keep the position as one of the industry leaders in medical technology***
  Philips has maintained a leading position in the medical device industry for decades. Investment has to be made in new technologies, such as AI, to keep that position.

4.1.3 Social drivers

- ***Make the world a better place***
  This is an overarching driver to develop AI in healthcare. The world can be made a better place by adding value to global health.

- ***Reach 3 billion people a year by 2030***
  The vision of Philips is to reach 3 billion people a year by 2030. By using AI technologies, millions of patient data can be analyzed in seconds, therefore, reaching more people than ever.

- ***Improve the experience of patients and healthcare professionals (HCPs)***
  By reducing the quantity of routine work, AI will allow HCPs to provide better care to patients.

4.1.4 Technical drivers

- ***The human need for technological innovation***
  There is an on-going need for continuous technological innovations. Humans always want to explore new territories.

- ***Improve technological efficiency and accuracy***
  By using AI, large volumes of data can be analyzed in a fraction of the time, making processes more efficient. Besides, it can support clinicians to make accurate diagnoses by comparing symptoms to large databases.

- ***Improve the repeatability of healthcare outcomes***
  Good practices and outcomes can be standardized. Tasks performed by humans can be reproduced by machines.

- ***Improve the quality of healthcare outcomes***
  By combining patient data with state-of-the-art literature, AI can help clinicians in the diagnosis and treatment of diseases.

4.2 Challenges

To analyze the challenges for the implementation of AI systems in healthcare, we divided them into non-ethical and ethical challenges. The non-ethical challenges were classified in organizational, regulatory, and technical categories (see Figure 9). On the other hand, the ethical challenges were divided according to the key ethical issues related to them: accountability; data bias, fairness, and equity; data privacy and security; effects on healthcare professionals; effects on patients; reliability and safety; and transparency and trust (see Figure 10). The criterion to include a challenge (ethical and non-ethical) is that it had to be mentioned at least for one interviewee. Next to the challenges, the number of interviewees that mentioned them is presented in brackets.

4.2.1 Non-ethical challenges

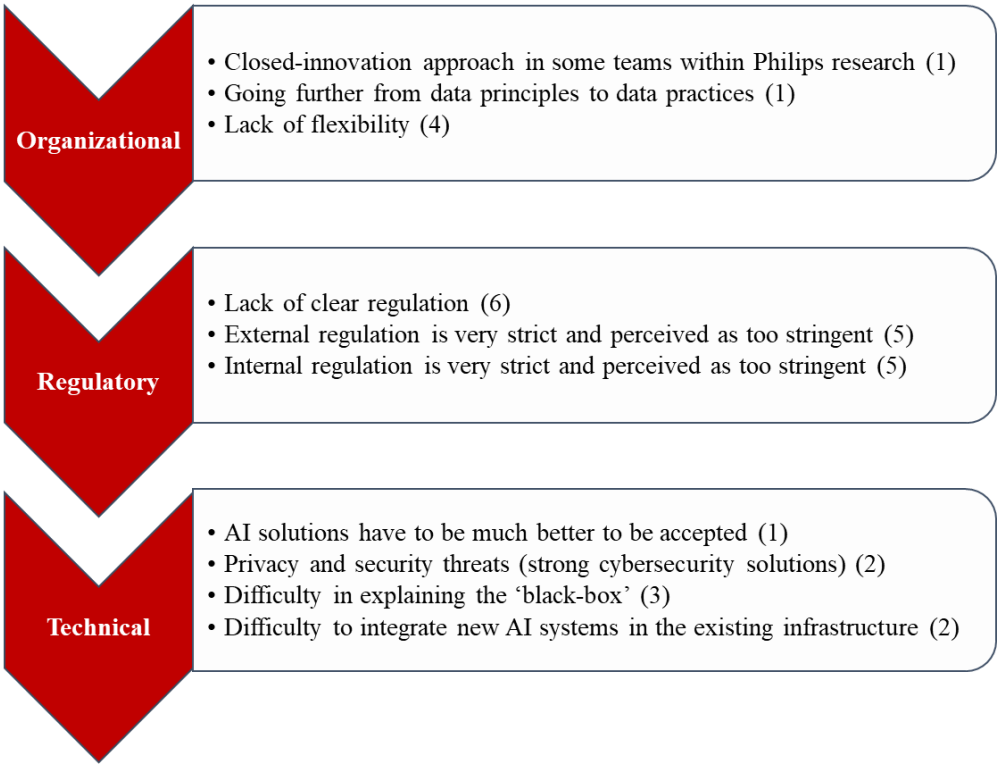

Figure 9. Non-ethical challenges for the implementation of AI in healthcare

4.2.1.1 Organizational challenges

- ***Closed-innovation approach in some teams within Philips research***
  Some researchers expressed that in their teams innovation is quite closed. Publishing internally is more appreciated than publishing externally.

- ***Going further from data principles to data practices***
  Philips recently developed a code of data principles. However, the translation of those principles into practices can take some years.

- ***Lack of flexibility***
  As a big company, R&D teams within Philips are restricted by many internal regulations. Once a path has been designed, it is really difficult to change direction.

4.2.1.2 Regulatory challenges

- ***Lack of clear regulation***
  There is a significant lack of clear regulation in several aspects of AI. First, there is no regulation to establish the accuracy needed to change from a decision-support system to a decision-making system. Second, each of the 28 countries of the EU interprets the GDPR differently. Third, there is no clear regulation regarding who has the ownership of patient data. Finally, there is still no regulation for dynamic algorithms in which the system keeps learning from new patient data.

- ***External regulation is very strict and perceived as too stringent***
  European legislation is very strict around privacy and data management. It intensifies with the stringent interpretation of the GDPR in the Netherlands.

- ***Internal regulation is very strict and perceived as too stringent***
  Philips is a risk-averse company and has very strict rules regarding data privacy. It slows down the process of innovation.

4.2.1.3 Technical challenges

- ***AI solutions have to be much better to be accepted***
  AI solutions have to be significantly better than the solutions that are already in the market to overcome privacy concerns.

- ***Privacy and security threats (strong cybersecurity solutions)***
  Hackers and digital criminals are constantly developing new techniques to steal data or cause harm. The cybersecurity teams in companies have to be one step ahead of criminals.

- ***Difficulty in explaining the 'black-box'***
  Making AI systems more transparent and explainable is a huge challenge. Especially in deep learning applications, where there are millions of neurons interacting and making correlations in a way that is impossible to understand for the human brain.

- ***Difficulty to integrate new AI systems in the existing infrastructure***
  Medical devices are connected to an entire ecosystem. New AI systems have to fit in the pre-existing digital infrastructures.

4.2.2 Ethical challenges

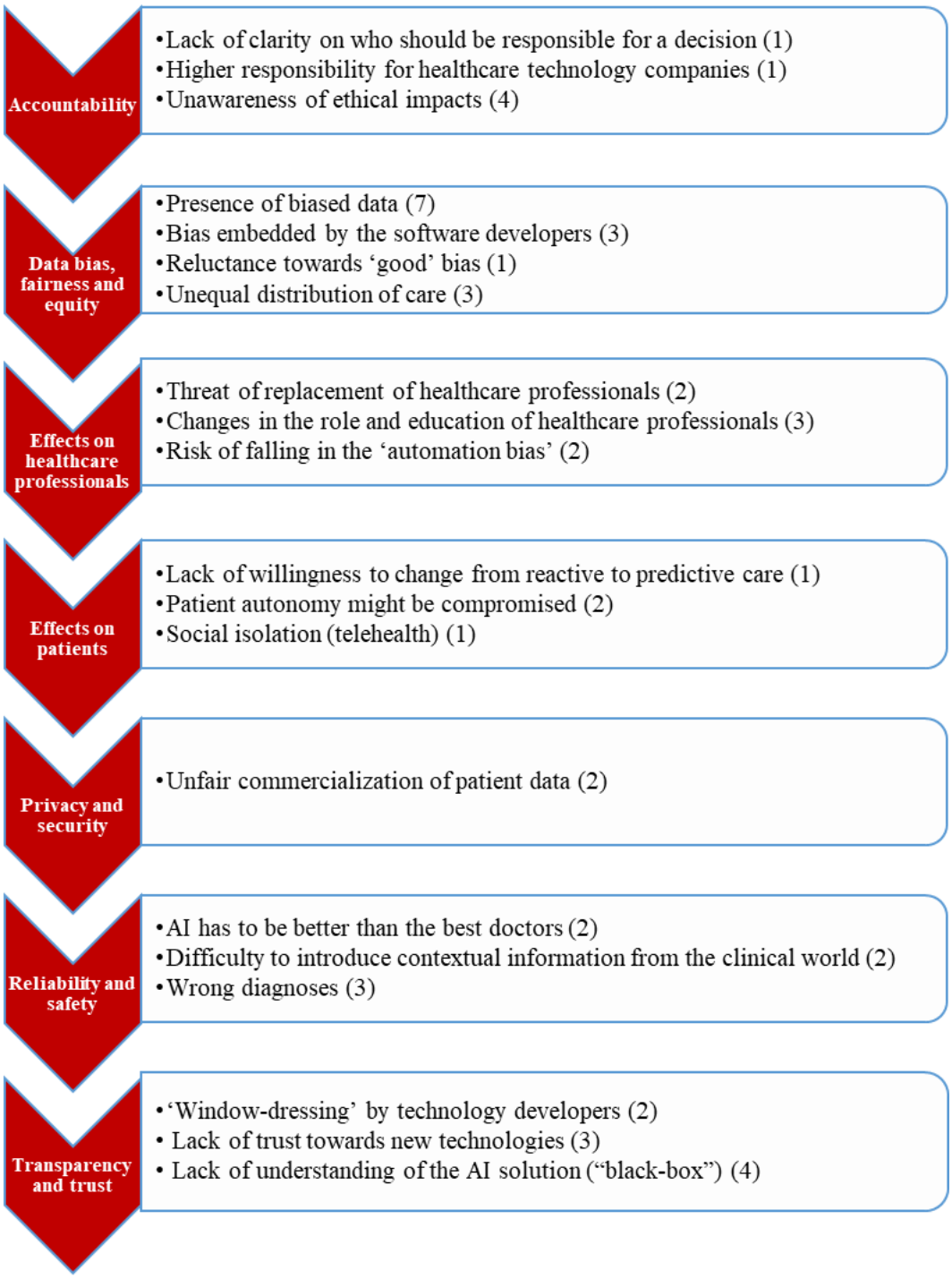

Figure 10. Ethical challenges for the implementation of AI in healthcare

4.2.2.1 Accountability

- ***Higher responsibility for healthcare technology companies***
  If a healthcare technology company wants to have a higher responsibility because there is a business opportunity, then it has to be held accountable for the outcomes. If the company gets the economic benefits, then it should also get the responsibility.

- ***Lack of clarity on who should be responsible for a decision***
  There is no clarity about who would be liable if AI goes wrong. The scientist, the doctor, the insurance company, the technology company, the hospital?

- ***Unawareness of ethical impacts***
  Doctors, scientists, and customers can express unawareness of ethical impacts. In that case, it is difficult to establish responsibilities. Some doctors are not interested in the process inside the "black-box", only in the outcomes. Some scientists see ethical reflection on their work more as a luxury than as a duty. They tend to see the moral part as a hindrance for technology development. Finally, if the end-customers do not see the benefits of responsible practices, then the efforts to implement RRI actions will be lost.

4.2.2.2 Data bias, fairness, and equity

- ***Presence of biased data***
  Sometimes, datasets only apply to certain contexts and cannot be applied in a wider scope. Therefore, it is not possible to have a full representation of an issue and there are mismatches with reality. These mismatches can lead to potential discrimination of populations or regions.

- ***Bias embedded by the software developers***
  There is no fully proved way against having a software developer that is deliberatively or inadvertently biased. Besides, it is hard to establish a diverse group of computer scientists that represent different countries, races, cultures, income levels, etc.

- ***Reluctance towards 'good' bias***
  Even if you have a bias that affects certain populations positively, people might not accept it. Such a bias can be seen as discriminative.

- ***Unequal distribution of care***
  Ensuring that AI products will not only reach wealthy populations is a big challenge. For example, people in poor regions cannot afford smartphones or wearables to monitor their health.

4.2.2.3 Effects on healthcare professionals (HCPs)

- ***Threat of replacement of healthcare professionals***
  AI can be seen as the super expert that will take over the doctors' expertise.

- ***Changes in the role and education of healthcare professionals***
  Doctors will have to change their role from focusing on diagnosis to focusing on providing care and assist patients during their treatments. AI will be in charge of the diagnosis. Additionally, the education of clinicians has to change to make them more used to deal with AI-enabled support systems.

- ***Risk of falling in the 'automation bias'***
  If AI systems were correct 98% of the time, doctors would behave less critically. They would tend to believe what the AI is telling them.

4.2.2.4 Effects on patients

- ***Lack of willingness to change from reactive to predictive care***
  People would not be motivated to take an active approach to their health.

- ***Patient autonomy might be compromised***
  People should have the right to decide whether they use a new technology or not. In telehealth, for example, people should be able to decide whether staying at home and being monitored remotely or going to the hospital and talk to the clinician. Even if it is cheaper for the healthcare provider, telehealth must not be imposed.

- ***Social isolation (telehealth)***
  Telehealth can create a sense of social isolation for patients who live alone. This intensifies in the case of elderly patients.

4.2.2.5 Privacy and security

- ***Unfair commercialization of patient data***
  Patient data can be used for unintended purposes.

4.2.2.6 Reliability and safety

- ***AI has to be better than the best doctors***
  AI has to outperform the best doctors to be accepted and to overcome privacy concerns. Humans tend to forgive human errors but humans do not forgive machine errors. It intensifies in healthcare, where lives can be at risk.

- ***Difficulty to introduce contextual information from the clinical world***
  Data analysis has to be combined with contextual information from the clinical world to provide robust hypotheses. If there is no contextual information, cause and effect analysis will not reflect the reality and AI solutions will not work.

- ***Wrong diagnoses***
  If data is not representative enough or the AI algorithm has some errors, wrong diagnoses might appear. It will seriously affect the trust towards the technology.

4.2.2.7 Transparency and trust

- ***'Window-dressing' by technology developers***
  Scientists tend to resort to "window-dressing" to keep high expectations towards a technology and to secure constant investment for technology development. The expectations are going faster than what scientists can deliver in a reliable manner.
- ***Lack of trust towards new technologies***
  Some clinicians and patients are very conservative and do not trust new technologies. Besides, users might not be informed that they are interacting with an AI application. This undermines transparency.

- ***Lack of understanding of the AI solution***
  This challenge is related to the control and explicability of the AI application. Technology companies might not understand entirely why an AI system is generating certain outcomes but still, they will put it in the market. Besides, doctors are not AI experts and some of them will trust the AI solution. It becomes risky if the algorithm provides erroneous results.

4.3 Risks

The statements of the interviewees were analyzed to establish the potential risks that might threaten the social acceptability of AI in healthcare. The risks were divided into ethical, technical, and organizational categories (see Figure 11). Technical risks correspond to threats for which a technical solution might be available. For example, privacy breaches can be avoided by implementing strong cybersecurity systems. Ethical risks are threats for which there is no clear solution at hand. For example, there is no established solution for avoiding the unintended discrimination of people. Many actions can be made to improve the diversity of data and of software developers, but these measures do not ensure that discrimination will be fully avoided. Finally, organizational risks correspond to unfortunate situations that Philips could experience if their AI solutions go wrong.

Figure 11. Risks for the social acceptability of AI in healthcare

4.3.1 Ethical risks

- ***Unfair commercialization of data***
  There is a risk that private companies can use patient data for purposes different from providing care, such as personalized marketing or monetization by reselling it.

- ***Reinforcement of inequality in healthcare***
  Poor regions where there are no electronic health records cannot provide the data to train AI systems for healthcare. Additionally, low-income populations cannot afford AI-enabled consumer products such as wearables or health apps for smartphones. These factors can stretch the existing gap between poor and rich for access and quality of healthcare.

- ***Social isolation***
  Telehealth can lead to the social isolation of people. Elderly patients might feel alone fighting against their health problems. Besides, clinicians are properly trained to identify non-verbal cues from patients. However, it would be quite difficult for telehealth systems to identify those cues, especially when dealing with psychological issues.

- ***Loss of patient autonomy***
  Hospitals would be able to send patients back home and treat them via telehealth to avoid costs. This can threat the autonomy of patients that want to be treated in the hospital in a traditional way. Moreover, health apps that do not inform the users that they are interacting with AI can mislead the patients to take certain decisions regarding their own health.

- ***Unintended discrimination***
  Some datasets used to train machine learning algorithms might not represent the diversity of a population. When applied at a wide scope, the results can only work for certain groups of people, which will cause unintended discrimination. Moreover, the software developers can unconsciously embed biases in the algorithms that they create. They can introduce biases from their culture, education or uprising without even realizing it.

4.3.2 Technical risks

- ***Misdiagnosis***
  AI decision-support systems trained with biased data or with data that does not take into account contextual information from the medical world might lead to wrong outcomes and put patients at risk. Besides, doctors might fall in the "automation bias", which is the human tendency to trust machines more than their own judgment. In that case, doctors will be approving wrong diagnoses made by AI.

- ***Privacy breaches***
  Patient data is considered sensitive and private. Hackers could be able to steal this data to de-anonymize it and blackmail people. They can also resell the data to make some profits. In a more extreme case, criminals could infiltrate hospital systems to create cyber-attacks and cause a huge harm for a minimal cost.

4.3.3 Organizational risks

- ***Loss of reputation***
  Philips has built a high quality reputation for more than a century. If the company creates a biased AI solution or violates the GDPR, the positive reputation will be put in danger. This will be extremely detrimental for Philips.

- ***Loss of trust from doctors, patients, and/or public***
  It is difficult to ensure trust in a technology that represents a "black-box". It affects the ethical principle of explicability. Doctors will be unable to explain the logic behind the results of AI-enabled decision-support systems. Besides, patients will be hesitant to believe in the outcomes of a process that cannot be explained. Moreover, the trust in AI can be at risk if technology developers resort to practices of "window-dressing" or "green-washing" to inflate the expectations.

- ***Opposition from the medical community***
  Healthcare professionals might feel threatened by AI diagnosis-support systems. They can have the feeling that they are going to be replaced. Besides, the training and education of clinicians will change radically with the introduction of AI in healthcare, which could bring more opposition.

4.4 Barriers

The statements of the interviews were analyzed to identify the most salient barriers that scientists have found when working on AI for healthcare applications. These barriers were classified in organizational, regulatory, and technical categories (see Figure 12).

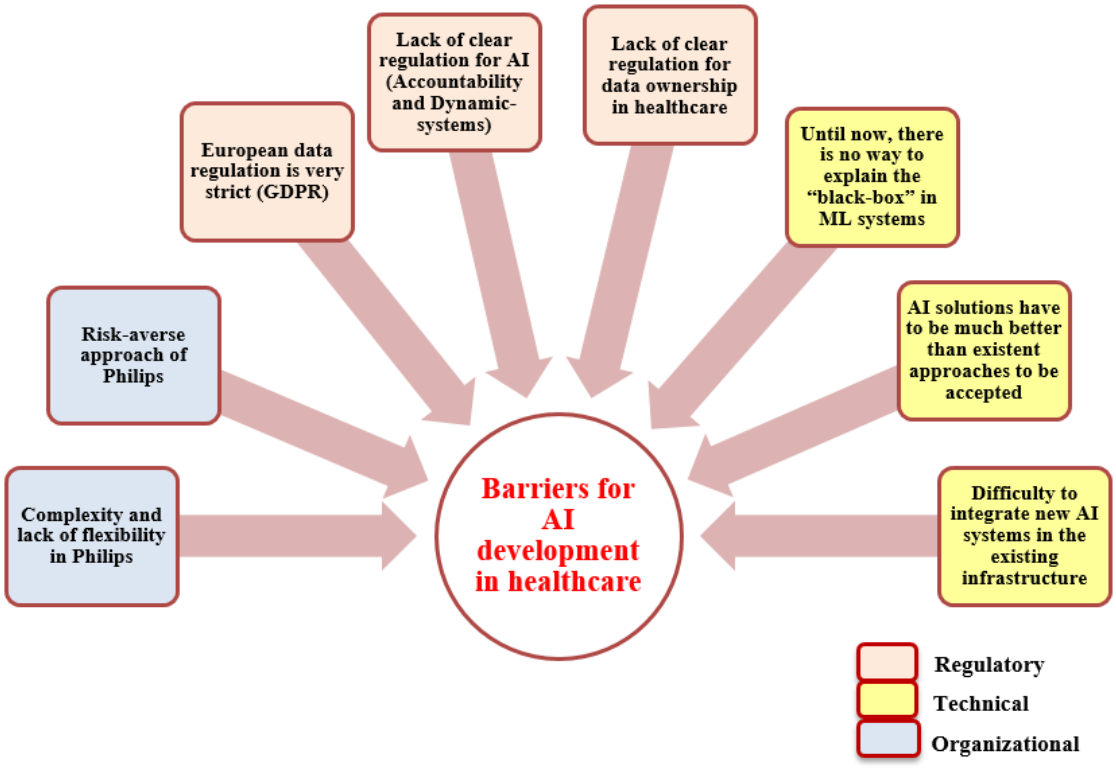

Figure 12. Barriers for AI development in healthcare

4.4.1 Regulatory barriers

- ***European data regulation is very strict (GDPR)***
  Health data is not easily accessible and European legislation is quite stringent around privacy and consent. Additionally, the Netherlands has adopted a very strict interpretation of the GDPR. For these reasons, scientists feel constrained when trying to access data to train their ML algorithms.

- ***Lack of clear regulation for AI (Accountability and dynamic-systems)***
  There is still no regulation for dynamic algorithms in which the system keeps learning in the field when new data is introduced. The FDA is working on it but developers are still unable to start clinical trials with that kind of algorithms. Furthermore, there is still no clarity regarding who should be held accountable for an AI mistake in healthcare.

- ***Lack of clear regulation for data ownership in healthcare***
  There is still no regulation that dictates who owns patient data or if patients should be compensated for it. Currently, the hospitals are considered the owners but this topic is still a grey area.

4.4.2 Technical barriers

- ***There is no way to explain the "black-box" in ML systems***
  The technology has not been developed yet to understand how a deep learning algorithm reaches an outcome. Scientists are attempting to make the "black-box" a "grey-box" by trying to identify the most important features that lead to a result. However, this approach is still incipient.

- ***AI solutions have to be much better than existent approaches to be accepted***
  In healthcare, AI systems have to be much better than current solutions to overcome privacy concerns. If the comparison is with a human doctor, AI has to be better than the best specialists. People would be willing to forgive human errors but never machine errors.

- ***Difficulty to integrate new AI systems in the existing infrastructure***
  Integration is a big barrier because new technologies are constrained to what already exists. AI systems have to fit in the existing IT infrastructure of hospitals. This situation stifles innovation.

4.4.3 Organizational barriers

- ***Complexity and lack of flexibility in Philips***
  Philips is a large company in which procedures are clearly established and written in stone. Changes can take a long time to be implemented, as so many different departments have to be coordinated.

- ***Risk-averse approach of Philips***
  Philips is very protective regarding data and consent. The interpretation of the GDPR inside the company is highly stringent. It makes AI development slower.

4.5 RRI Actions

After coding the expert interviews (see Appendix 2), we could identify the RRI actions that are already being developed by Philips. Those actions are divided into the different RRI dimensions: Anticipation & Reflection (Table 10), Inclusiveness (Table 11), and Responsiveness (Table 12). Each RRI action mentions the potential benefits of its implementation, the corporate functions that should be involved to make it happen, and the stakeholders that will be taking part in the action. The actions that were described as ***critical*** for the social acceptability of AI are highlighted in red. Philips is already implementing some of these actions, but the effort should be sustained to ensure acceptability in the long-term.

4.5.1 Anticipation & Reflection actions

Anticipation and reflection actions are intended to evaluate the potential impacts of AI development since the early stages of technology development (PRISMA Project, 2019). By reflecting on the potential ethical, social, and legal consequences of the introduction of a product, risks can be avoided and social acceptability can be strengthened.

Anticipation & Reflection Integrate analysis of ethical, legal and social impacts since the early stages of product development				
#	**Action**	**Benefits**	**Corporate functions involved**	**Stakeholders involved**
1	Analyze the potential impacts of AI systems since the initial phases of development	Prevention and mitigation of risks	CSR, R&D, legal	AI scientists, ICBE, privacy officers, data officers, project leaders, end-users
2	Develop and implement data and AI principles (codes of conduct)	Compliance with EU regulations/Creation of clear guidelines for technology development	Management, CSR, legal, Q&R	Managers, legal officers, privacy officers
3	Educate PHDs within the company in the ethical aspects of research	Increase awareness of ethical concerns	CSR, R&D, HR	AI scientists, CSR officers, PHD students
4	Ensure that AI-enabled medical devices are clinically validated	Compliance with EU regulations/Increase product reliability	Legal, Q&R, R&D	Legal officers, AI scientists, project leaders, external regulatory bodies, clinicians
5	Design rigorous procedures to ensure that data has a high-quality, it is well-curated and representative of the population	Introduction of contextual knowledge/Diminish the risk of discrimination/Avoid misdiagnosis	R&D, CSR, Q&R	ICBE, AI scientists, data officers
6	Ensure that the vision of the company with AI is aimed at supporting the doctors, not replacing them	Securing the support of clinicians/Preserve the principle of human autonomy/ Increase social acceptability/ Build corporate image and reputation	Strategy, management, marketing	Managers, strategists, project leaders, marketing managers

7	Establish an ethical, social and legal monitoring board	Prevention and mitigation of risks/Increase awareness of ethical concerns/Compliance with EU regulations	Management, R&D, CSR, Q&R, legal	ICBE, AI scientists, data officers, legal officers
8	Establish clear and rigorous practices of data management regarding privacy and consent	Prevention and mitigation of risks/Compliance with GDPR/Prevent privacy breaches	R&D, CSR, legal, Cybersecurity	Legal officers, data officers, privacy officers, cybersecurity experts, end-users
9	Every AI application developed has to comply with the ethical principles of AI established by the European Commission	Compliance with EU regulations/Build corporate image and reputation	Legal, CSR, R&D	Legal officers, AI scientists, external regulatory bodies, ICBE
10	Guarantee that patient data is anonymized	Prevention of privacy breaches/Compliance with GDPR	R&D, Cybersecurity, legal	AI scientists, legal officers, cybersecurity experts, data officers, privacy officers
11	Organize workshops with internal stakeholders to reflect on the challenges of AI	Prevention and mitigation of risks/Increase awareness of ethical concerns/Compliance with EU regulations	CSR, R&D	AI scientists, project leaders, CSR officers, privacy officers, legal officers, ICBE
12	Train employees in ethical and privacy principles	Prevention of privacy breaches/Increase awareness of ethical concerns	CSR, HR, R&D, legal	All employees within the company
13	Understand and implement the legislation early on the technology development process	Compliance with EU regulations/GDPR	Legal, R&D	AI scientists, legal officers

Table 10. RRI actions for Anticipation & Reflection

4.5.2 Inclusiveness actions

Inclusiveness aims at taking into account the view of all the potential stakeholders involved or affected by technology development (Stilgoe et al., 2013). This RRI dimension is very important because considering the values and expectations of the stakeholders early in product development could be translated into a huge increase in social acceptability once the product is launched.

Inclusiveness Perform stakeholder engagement to inform all phases of product development				
#	**Action**	**Benefits**	**Corporate functions involved**	**Stakeholders involved**
14	Design cross-population studies to validate AI applications	Prevention of biases/Diminish the risk of discrimination	R&D, CSR	AI scientists, clinicians, patients, public, data officers
15	Make sure that health care professionals are included in the process of selecting and curating the data	Introduction of contextual knowledge/ Diminish the risk of discrimination/Avoid misdiagnosis	R&D, CSR, Q&R	Clinicians, nurses, healthcare administrative personnel, AI scientists, privacy officers, data officers, project leaders, ICBE
16	Enhance the collaboration with academy for AI related purposes	Introducing open-innovation approaches/ Validation of AI developments	R&D, legal	Universities, AI scientists, legal officers, privacy officers
17	Ensure a high level of diversity in AI software development teams	Prevention of biases/Diminish the risk of discrimination	HR, CSR	Recruiters, CSR officers
18	Establish AI research centers in different income-level countries	Avoid inequality in healthcare/Prevention of biases/Diminish the risk of discrimination	Management, strategy, HR, R&D	Managers, strategists, recruiters, AI scientists
19	Find a balance in the teams between innovativeness and experience	Prevention and mitigation of risks/ Increase product reliability	HR, R&D	Recruiters, project leaders
20	Include healthcare professionals (HCPs) and patients in the process of technology development	Introduction of contextual knowledge/Validation of AI developments	R&D, CSR, Q&R, legal	Clinicians, nurses, healthcare administrative personnel, AI scientists, privacy officers, data officers, project leaders
21	Involve patients when developing AI regulations for healthcare	Preserve the principle of human autonomy/ Increase social acceptability/Anticipate regulatory changes	Q&R, legal	Patients, legal officers, data officers, external regulatory bodies
22	Set co-creation exercises with a wide range of stakeholders	Introduction of contextual knowledge/Introducing open-innovation approaches/Validation of AI developments	R&D, CSR, Q&R, procurement, marketing	Clinicians, nurses, healthcare administrative personnel, patients, public, suppliers, universities, AI scientists, privacy officers, data officers
23	Work together with companies that have expertise in other fields	Introducing open-innovation approaches/Validation of AI developments	Management, strategy, R&D, legal	Digital-first companies, startups, AI scientists, legal officers, privacy officers, strategists, managers

Table 11. RRI actions for Inclusiveness

4.5.3 Responsiveness actions

Responsiveness refers to the ability of the company to change the direction of innovation according to stakeholders' and public' requests (PRISMA Project, 2019). Besides, the speed of reaction when new regulations appear, when there is a change in the business environment, or when unintended consequences occur (Stilgoe et al., 2013). If the company is fast enough to deal with unexpected problems and adjust to stakeholders' requirements, the social acceptability of AI will be enhanced.

Responsiveness Integrate monitoring, learning and adaptive mechanisms to address public and social values and normative principles in product development				
#	**Action**	**Benefits**	**Corporate functions involved**	**Stakeholders involved**
24	Implement internal policies that allow AI scientists to make quick changes once a project has started	More flexibility to deal with unexpected problems	Management, R&D	AI scientists, project leaders, managers
25	Always inform the patients or customers that they are interacting with AI and not a real person	Preserve the principle of human autonomy	R&D, marketing	AI scientists, project leaders, marketing managers, patients
26	Communicate that consumer-monitoring applications should only be used for pre-screening of diseases	Prevention and mitigation of risks	R&D, marketing	AI scientists, project leaders, marketing managers, users
27	Communicate to the clinicians the limitations of the AI solution	Securing the support of clinicians/Prevention of "automation bias"	R&D, marketing, sales, Q&R	AI scientists, project leaders, marketing managers, users, clinicians, vendors, quality inspectors
28	Develop activities of experimentation and innovation without regulatory constraints (bootcamps, hackathons, etc.)	Introducing open-innovation approaches/ Validation of AI developments/Boost innovation	R&D	AI scientists, project leaders
29	Develop and implement AI solutions to detect and avoid biases from software developers	Prevention of biases/Diminish the risk of discrimination	R&D, CSR	AI scientists, project leaders, CSR officers
30	Develop AI solutions that can be easily integrated in the existing systems	Enhancing the support of clinicians, nurses and healthcare administrative personnel	R&D	Healthcare administrative personnel, doctors, nurses, AI scientists
31	Encourage an agile-way of working	More flexibility to deal with unexpected problems /Boost innovation	R&D	Project managers, AI scientists
32	Ensure that AI applications do not compromise patient autonomy	Preserve the principle of human autonomy/Increase social acceptability	R&D, CSR	AI scientists, CSR officers

33	Ensure that data scientists work in teams and that there are peer-review discussions going on	Prevention of biases/Diminish the risk of discrimination/ Validation of AI developments	CSR, R&D	CSR officers, AI scientists, project leaders
34	Ensure that your AI application is only employed for its intended used	Prevention and mitigation of risks/ Increase product reliability	R&D, marketing	AI scientists, project leaders, marketing managers, clinicians, patients
35	Establish a strong cybersecurity team	Prevention and mitigation of risks/ Prevention of privacy breaches	HR, Cybersecurity	Recruiters, cybersecurity experts
36	Establish periodic data security risk assessments	Prevention and mitigation of risks/ Prevention of privacy breaches	Cybersecurity, legal	Cybersecurity experts, data officers, privacy officers
37	Educate the HCPs on how to use AI-enabled systems and on how to interpret the “black-box”	Prevention of "automation bias"/ Increase product reliability	Sales, R&D	Clinicians, AI scientists, vendors
38	Only put in the market AI applications that are at least as good as the best doctor or the best solution already existent	Increase product reliability/Prevention and mitigation of risks/ Avoid misdiagnosis	R&D, marketing	AI scientists, project leaders, marketing managers, clinicians
39	Implement interaction features in telehealth applications	Prevention of social isolation/ Improve communication with patients	R&D	AI scientists, project leaders, clinicians, nurses, patients, users
40	Involve privacy officers during all phases of technology development	Compliance with GDPR/Prevent privacy breaches	R&D, Cybersecurity, legal	AI scientists, legal officers, cybersecurity experts, data officers, privacy officers, ICBE
41	Iterate based on the feedback from users	Increase social acceptability/ Validation of AI developments	R&D	AI scientists, project leaders, clinicians, nurses, patients, users
42	Never draw conclusions based on small datasets	Prevention of biases/ Diminish the risk of discrimination	R&D, CSR	AI scientists, data officers, ICBE
43	Validate every document and every phase of AI development	Prevention of biases/Validation of AI developments/Compliance with GDPR	R&D, CSR, legal, Q&R	ICBE, AI scientists, legal officers, privacy officers, quality inspectors
44	Work closely with legislators on addressing legal challenges	Compliance with EU regulations/GDPR	Legal, R&D	Project leaders, legal officers, external regulatory bodies
45	Work on new ML technologies to try to make the "black-box" more transparent	Improve the explicability of outcomes/Improve social acceptability	R&D	AI scientists
46	Always include the percentage of accuracy of an AI solution	Prevention of "automation bias"/ Avoid misdiagnosis	R&D	AI scientists

47	Design mechanism that allow the user to shut down the AI system in the case of misbehavior or under cyberattacks	Prevention and mitigation of risks/ Prevention of privacy breaches	R&D, Cybersecurity	AI scientists, cybersecurity experts
48	Publish the research advances in AI within Philips in peer-reviewed journals	Validation of AI developments	R&D	Universities, AI scientists

Table 12. RRI actions for Responsiveness

4.6 Technologies and products

From the statements of the interviews, we were able to identify the AI technologies/products that have already been introduced or that will be implemented in healthcare. We classified them in short-, medium-, and long-term introduction. For each different application, we identified the products and solutions that have been or are being developed by Philips. The strategy of the company in the long-term is to change from a product-oriented approach to a solution-oriented approach. In the latter, several products play a role within a big healthcare solution (Philips, 2019c). For example, the AI-enabled Philips toothbrushes can measure how healthy your oral health is and relate it with your patterns of sleep detected by Philips SmartSleep Analyzer. By introducing all these data in ML algorithms, an app would be able to indicate you if your health is deteriorating, and provide nutrition or exercise recommendations.

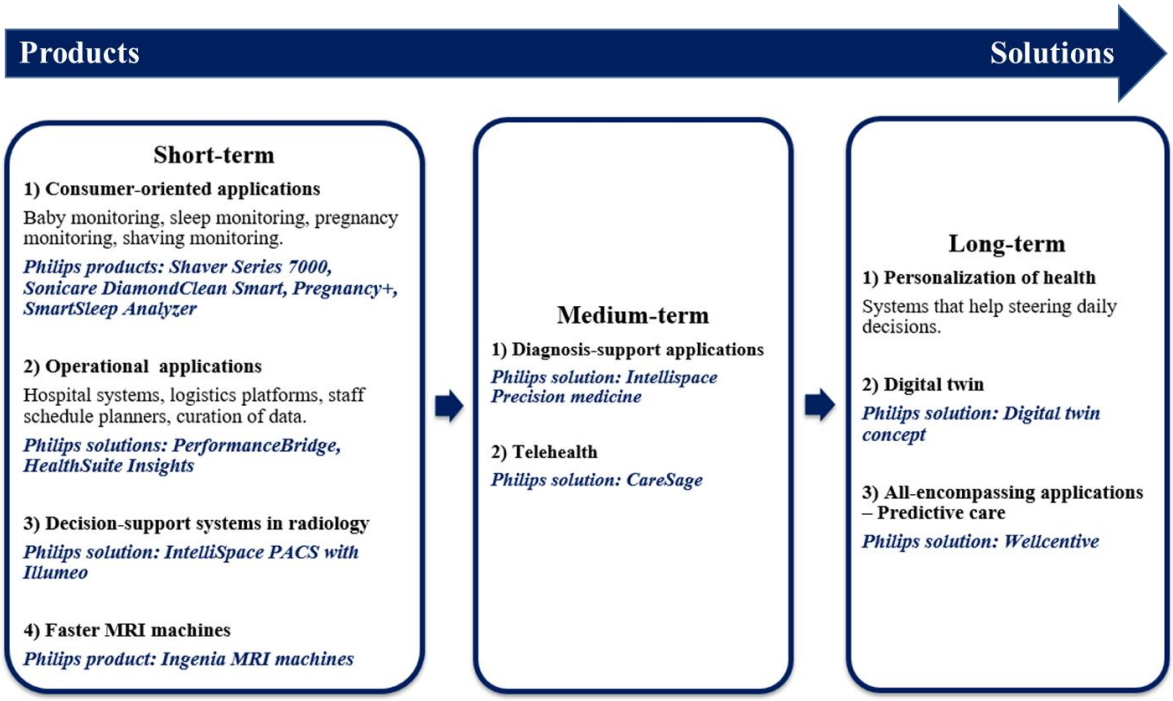

Figure 13. AI-enabled technologies/products in healthcare

4.6.1 Short-term introduction

Short-term products correspond to AI technologies that are already in the market and for which there are not many potential ethical risks. These kinds of products are monitoring devices, operational systems aimed at improving efficiency, decision-support systems in radiology, and AI-enabled MRI machines.

- ***Consumer-oriented applications***
  These products are intended to be used directly by consumers. Baby, sleep, pregnancy, and shaving monitoring devices can be included in this category.

  Philips products: *Shaver Series 7000, Sonicare DiamondClean Smart, Pregnancy+, SmartSleep Analyzer*

- ***Operational applications***
  AI systems aimed at improving the operational side of healthcare. These applications make the processes faster, more efficient and less wasteful. Logistics, procurement, and curation of data are some of the processes that will be improved.

  Philips products: *PerformanceBridge, HealthSuite Insights*

- ***Decision-support systems in radiology***
  Radiology is the field of medicine that has had the greatest impact from AI technologies. There are already systems that show the radiologist the area to focus on, making their work more efficient.

  Philips products: *IntelliSpace PACS with Illumeo*

- ***Faster MRI machines***
  New generation MRI machines use machine learning to analyze data faster and generate scans significantly faster than before.

  Philips products: *Ingenia MRI machines*

4.6.2 Medium-term introduction

Medium-term products correspond to applications in which there are some ethical risks to be considered. For example, diagnosis-support systems that will change the nature of clinicians' work or that could induce "automation bias". Besides, telehealth applications can generate a sense of social isolation in some patients. These products also require a longer time to be developed because they have to be certified as medical devices by the FDA.

- ***Diagnosis-support applications***
  Systems that will augment the capabilities of healthcare professionals by assisting them in diagnosis and treatment activities.

  Philips products: *IntelliSpace Precision Medicine*

- ***Telehealth***
  There is a problem of an aging population and lack of healthcare professionals. Telehealth allows patients to be treated remotely by installing monitoring devices at their homes. These devices check the patients' health constantly and send alerts to the doctors in case of health deterioration.

  Philips products: *CareSage*

4.6.3 Long-term introduction

Long-term products are still concepts of how healthcare will be provided in the future. There are still studies to be done and impacts to be assessed to ensure that these technologies will be accepted by healthcare professionals, patients, and society.

- ***Personalization of health***
  Applications that will suggest you the healthiest path of actions in your daily life. A device or app will tell you the pros and cons of your decisions, so that you can better steer your actions to avoid risks for your health. However, this concept challenges the principle of human autonomy.

- ***Digital twin***
  Detailed models of anatomy, physiology, and pathology. This technology makes it possible to get a virtual representation of the different organs of your body. Clinicians could run simulations in your virtual body to assess the efficacy of different treatments. However, this concept challenges the privacy of the patients. Do people really want to have an exact copy of themselves online?

  Philips products: *Digital twin concept*

- ***All-encompassing applications – Predictive care***
  AI systems that connect society, healthcare systems, hospitals, and caregivers in an organized way to identify in advance potential health risks for individuals or populations. The risk here is that such systems can be seen as surveillance instrumetns.

  Philips products: *Wellcentive*

In this chapter, the main drivers, challenges, risks, barriers and RRI actions to implement AI in healthcare were identified from the coding process of the interviews. Besides, Philips products were divided according to their time to market: short-, medium-, or long-term. Appendix 3 shows the RRI roadmap for the case of AI development in Philips. In the next chapter, we will discuss the results and relate them with the insights obtained from the literature.

5. Analysis and Discussion

This chapter focuses on discussing the results obtained from the interviews and combining them with the information gathered in the literature review. We analyze the outcomes obtained for the drivers, challenges, risks, and barriers for the implementation of AI in healthcare. In addition, the RRI actions that are already being developed by Philips and the actions that require further work are discussed. Finally, we reflect in what cases RRI practices represent a good strategy to deal with ethical concerns and increase the social acceptability of IA in healthcare and in what cases other approaches are more suitable.

5.1 Drivers

The first interesting insight that we got from the interviews is that AI plays a strong role in Philips business strategy for technology development. The vision of the company is to reach 3 billion people a year from 2030 (Philips, 2019b). Therefore, this vision was mentioned as one of the main drivers for Philips to develop AI solutions, as this technology will help the company to reach more people than ever. This driver matches the economic driver of increasing revenue, because AI can help Philips to extend their customer base. There are also drivers related to keeping the position as one of the industry leaders in medical devices, or becoming the leader in the digital transformation of healthcare. According to the interviewees, the company is in an excellent position to achieve those goals due to the contextual knowledge of Philips in the medical world. They expressed that digital-first companies can be one step ahead in the development of AI but that Philips is one step ahead in making sense of clinical data and developing solutions that apply to real healthcare settings. This expertise puts the company in par with tech giants such as IBM or Microsoft in the field of AI for healthcare. To illustrate this, we refer to an example from the literature review. Caruana et al. (2015) analyzed the case of an AI application designed to evaluate the risk of developing health complications after suffering pneumonia. The app erroneously diagnosed that patients with asthma should be sent home as they have a higher rate of survival. However, it did not take into account the fact that asthma patients are usually sent directly to the ICU and provided with better care, which increases their chances to survive (Caruana et al., 2015). Philips is aware of this kind of issues, because the company has developed a long-term supportive relationship with healthcare professionals. This represents an important competitive advantage that should be fully exploited by designing context-aware AI solutions.

Besides, there are social drivers that apply to the entire healthcare industry. For example, making the world a better place and improving the experience of patients and clinicians were some of the answers given by the interviewees. These are overarching drivers that should govern AI development. Finally, the interviewees mentioned technical drivers such as

improving efficiency, accuracy, repeatability, and quality of healthcare outcomes. Despite their technical nature, these drivers are the base to reach the overarching social drivers described before. The study of Auer and Jarmai (2017) also provided some drivers for innovation in SMEs that could apply for this case and that were not mentioned by the interviewees. For example, they expressed that an important driver is to get customer knowledge. By getting feedback on how users react to products, developers can create better solutions (Auer & Jarmai, 2017). They also claimed that computer-driven technologies, such as AI, offer an important pool of external knowledge that can be introduced in the organization by collaborations and networks. In the case of Philips, the company is already working together with Microsoft and some medical device startups to create innovative products (Philips, 2019c). Finally, Auer and Jarmai (2017) mentioned that research in innovative technologies was useful to get public funding in case that financial resources were limited. We believe that this driver does not apply to the same extent to startups or SMEs than to multinational companies (MNCs). MNCs usually have the resources to engage in innovation and although public funding could give an extra boost, this is not a critical driver for a large technology corporation with a significant budget assigned to R&D.

5.2 Non-ethical challenges

All the interviewees agreed that the most important non-ethical challenges for AI development in healthcare are related to regulation. First, they argued that regulation about data and privacy is very strict in Europe. They think that this stringent approach is seriously stifling innovation in AI. Additionally, they think that it intensifies due to the lack of clarity in different aspects. For example, the GDPR is only a guideline that can be interpreted differently by the 28 countries of the European Union, making it difficult to ensure fairness and equity in the process of technology development. For example, interviewee F said "the *GDPR has a very lax interpretation in some countries and a very strict interpretation in other countries, such as the Netherlands. Then, we are at a disadvantage when competing with companies that operate in lax regulatory environments, in which you can get more data easily*". Besides, there is still no clear regulation regarding data ownership and regarding the introduction of open algorithms in healthcare, in which the systems keeps learning in the field when new data is fed. The interviewees expressed that it made research difficult because they were not sure whether the systems that they were developing would be allowed or not. Interviewee H said "*We can spend three years working in a very helpful open algorithm and then maybe we would not be allowed to put it in the market due to regulation. It is a big concern that we have because those are three years lost*". These results are quite interesting because technology developers complain about the strictness of the regulation already in place but want more regulation for some other topics. It allow us to infer that scientist do not want to have empty spots in the regulation for AI but want it to be lax.

Regarding the stringency of the policies, the author believes that there is not much that can be done because the main requests of the interviewees were to have easier access to data and to avoid the need of explicability of the AI outcomes. First, the GDPR is relatively new,

as it was implemented on May 25th, 2018 (GDPR, 2018), and important changes are not expected in the short-term regarding data access. Second, the "Ethics guidelines for trustworthy AI" designed by the High-Level Expert Group on Artificial Intelligence (European Commission, 2018) gave a strong emphasis to the protection of data and the need for explicability. It does not seem that AI regulation is going towards a more relaxed approach. Besides, the avoidance of explicability could lead to a situation in which the technology developers could evade their responsibility. Floridi, et al. (2018) argued that explicability is crucial for implementing "accountability" in AI. They claimed that in order to hold the developers accountable for a negative outcome, it is necessary to have at least a minimum understanding of how the outcome arose. On the other hand, the request for clear regulation regarding open algorithms is justified because this technology could bring important benefits to the medical field, such as more accurate decision-making support or more efficient processes (Academy of Medical Royal Colleges, 2019). The FDA is already working on the regulation for this type of AI (FDA, 2019). However, the introduction of the regulation will take some time due to the complexity of the technology (FDA, 2019).

When talking about the organizational challenges within Philips, the answers varied significantly depending on the age and experience of the interviewees. Younger scientists tend to classify Philips as a very closed and inflexible company, where testing new ideas or making changes is difficult. For example, interviewee E expressed "*Internal policies in Philips have a big impact. Sometimes, I just want to test an idea with regular data, not patient data. Then, I need to make a project proposal to make an experiment that will take an afternoon. I am not going to work on a project proposal for a week for something that I can easily test in one afternoon. Then, I prefer not to do the experiment*". In addition, interviewee H said "*Despite being a global company, Philips adopts a western perspective in AI and it is very protective regarding data. It makes the process a lot longer*". On the other hand, experienced scientists in Philips argued that internal regulations are necessary to avoid mistakes that could damage the reputation of the company. Interviewee J said, "*If something goes wrong with AI and there are big headlines in the newspapers, it can be very detrimental for Philips*". The literature supports the statements of the younger scientists. For example, Kirsner (2018) established that large companies have important barriers that stifle their process of innovation. He mentioned that there are important factors such as politics, inability to act, lack of executive support or inability to catch signals from the business environment that seem to be common in big corporations. For the case of RRI implementation, some studies mentioned that there are barriers such as unawareness of RRI in the business world, power differences, higher costs to deploy RRI and lack of clarity regarding the potential benefits of RRI, which may hamper the adoption of responsible practices (Auer & Jarmai, 2017; de Hoop, Pols, & Romijn, 2016). We conclude that there is still more job to be done here. Scientists should organize meetings and come up with solutions that allow for flexibility in AI development without risking the company's reputation.

The interviewees also mentioned some technical challenges. These include the development of stronger cybersecurity practices, the lack of technical solutions to explain the

"black-box", the difficulties to ensure data quality and quantity, and the challenge of implementing new AI solutions in the existing clinical infrastructure. However, all of the interviewees agreed that overcoming these challenges was just a matter of time and that the real effort should be put in dealing with the ethical challenges. Nevertheless, we believe that they overlook the "black-box" problem. According to Bathaee (2018), there are two types of black boxes: weak and strong. Weak black boxes are opaque to humans but allow for reverse engineering to determine the most important variables that AI takes into account in the decision-process. On the other hand, strong black boxes are entirely opaque to humans and there is no way to determine how the system reached an outcome. Mittelstadt, et al. (2016) argue that strong black boxes seriously affect the principles of transparency and explicability in AI. This bring us back to the topic of regulation because open algorithms are considered strong black boxes (FDA, 2018) and, therefore, regulation must take into account how to deal with the opacity of such systems.

Finally, regarding privacy and security threats within Philips. The experts agreed that the company has very strong cybersecurity practices and an entire cybersecurity team making the systems more difficult to access every day. In addition, they expressed that the interpretation that Philips gives to the GDPR is extremely stringent, which gives almost no chance for privacy and safety risks, such as the de-anonymization of patient data.

5.3 Ethical challenges

The statements related to ethical challenges were classified according to the key ethical categories described in the literature review: accountability; data bias, fairness, and equity; effects on healthcare professionals; effects on patients; privacy and security; reliability and safety; and transparency and trust.

In accountability, the interviewees agreed that the most difficult challenge is to establish who should be responsible when AI goes wrong. There is no clarity on how to design the regulation for this issue. Until now, the doctor is still responsible as he is the one that takes the final decision (Hart, 2017a). However, doctors might fall in the 'automation bias', which is the human tendency to trust machines more than their own judgment, even if they are correct (Skitka et al., 2000). In that case, responsibility becomes a 'grey-area' because doctors might end-up justifying wrong decisions supported by AI-systems. In situations like that, interviewees recognized that it is time for healthcare technology companies to start having a higher responsibility. For example, interviewee E expressed that "*If companies want to take advantage of the business opportunities that AI offers, they should also take part of the responsibility when it goes wrong*".

Data bias was another important challenge brought to the table by the interviewees. They argued that Philips does an excellent job of selecting and curating the data used to train the algorithms. Then, they did not see a problem with data quality. However, they see a problem in ensuring that data scientists working on the algorithms are not embedding biases. Although

the company tries to ensure diversity in the software development teams, sometimes it is not an easy task. For example, interviewee F said "*it is hard to find as many computer scientists, female, from emerging countries in comparison to computer scientists, male, from the US or Europe.* Macnish (2012) expressed that the values of the developers are always frozen and institutionalized in the algorithms. Similarly, Johnson (2006) claimed that technology development is never neutral or linear; the developers will always have to make choices based on their knowledge and values. This could lead to biases being 'loaded' unconsciously in the algorithms, reflecting the prejudices and beliefs of AI developers (Academy of Medical Royal Colleges, 2019). The experts also mentioned the reluctance of individuals to accept "good bias" when it is seen as discriminatory. Interviewee E gave the example of a drug discovered using AI that only worked for black people. Although it worked really well in that population, it was discarded because the users did not want to be seen as a different group. This is corroborated by Mittelstadt, et al. (2016), who expressed that an action can be found discriminatory, even if it is based on conclusive, scrutable, and well-founded evidence.

The interviewees also discussed the unequal distribution of care. They said that the status of AI in healthcare is very elitist. First, because the data used to train the algorithms comes from regions where there are electronic health records in place, which is already leaving aside countries where medical notes are still taken with pen and paper. This issue also leads to bias, because the solution will only work for the populations whose data were used to train the algorithms (Bird, et al., 2016). Second, AI healthcare apps that run in smartphones and wearables are inaccessible for poor people. The interviewees expressed that healthcare companies should extend their efforts to make sure that AI will also benefit these people.

The clinicians interviewed were very positive about the development of AI in healthcare. They understand the value that this technology could bring to their profession and they think that the fears are unjustified. For example, interviewee D said, "*some doctors are scared that they will be replaced by AI, but only because they don't know what AI is*". However, the literature indicates that there is a serious threat for physicians. According to Hamid (2016), physicians may feel that their autonomy and expertise will be challenged by AI systems. The concerns intensify in the field of radiology, in which AI has done an important progress in image analysis (Platania et al., 2017; Ranschaert et al., 2019). In fact, some studies have anticipated that the first radiologists will be replaced by computers within the next 4-5 years and that this medical specialization could disappear in the next 10 years (Chockley & Emanuel, 2016; Obermeyer & Emanuel, 2016). The interviewees also expressed that the education and roles of physicians would change in the future, from a diagnostic-based approach to a care-based approach. We believe in that last affirmation, however, we conclude that there is a serious threat for the field of radiology and that actions should be developed to ensure that radiologists do not become redundant. Instead of replacing them, AI could enhance them and ease their work. If replacement is unavoidable, their role as the point of contact between the machine and the patient should never be substituted.

Interviewee B expressed that "*under no circumstance, we can threat the autonomy of the patient*", which is one of the principles of medical ethics (AMA, 2019). He said that in telehealth, people should also have the option of going to the hospital and talking to healthcare professionals. Besides, Mittelstadt (2019) argues that healthcare apps should always inform the user that they are interacting with AI, leaving the autonomy to the patient to decide whether to use it or not. This observations align with the "*Principle of respect of human autonomy*" presented in the Ethics Guidelines for Trustworthy AI (European Commission, 2018). The High-level Expert Group on Artificial Intelligence argued that AI systems "should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans" (European Commission, 2018). The author agrees with interviewee B and the literature in the fact that autonomy should never be compromised and that proper measures should be developed to ensure that.

The next challenge identified by the experts was the fact that there is a double standard when judging machine and human errors. Interviewee D expressed *"As a doctor, I'm allowed to make mistakes. If I give a proper explanation of my reasoning, the mistake might be admissible but people will never accept it when it comes from a machine";* and Interviewee F said, "*Humans tend to forgive human errors but humans do not tend to forgive machine errors. In healthcare, machines have to be much better than humans to be accepted and to replace even the simplest human tasks*". This is a big challenge because AI systems have to be not just better but far better than humans to be accepted, which might not be attainable. Bostrom and Yudkowsky (2011) argued that AI falls short of human intelligence. They said that although AI can perform specific activities quite well, it stills lack generality. That means that an AI system cannot be used for any activity, only for the specific one for which it was designed (Bostrom, N., & Yudkowsky, 2011). We believe that this has implications regarding the design of AI systems for healthcare because if the AI tool is confronted with a situation that differs from the specific activity to which it was designed, the lack of generality could lead to erroneous outcomes. Those results will undeniable affect the trust in AI.

Finally, the experts indicated that trust from doctors, patients and society might be put in danger if scientists resort to "window-dressing" practices to inflate the expectations of AI in healthcare. In fact, interviewee F expressed "*AI expectations are going faster than what you can deliver in a reliable manner. If this trend continues, the trust in the technology will be lost*". They suggest that to avoid this situation, scientists should always inform the limitations of their models and the scope of the solution. According to interviewee C, "*if people know clearly what they are going to get when using AI systems, they will be more willing to accept them*". However, this trend is difficult to avoid as it corresponds to the "Gartner Hype Cycle" usually experienced by digital technologies (Gartner, 2019). A technology normally starts with an innovation trigger, followed by a peak of inflated expectations. Then, there is a period of disillusionment in which the technology fails to deliver. This stage is followed by a period of enlightenment in which the technology is better understood. Finally, a plateau of productivity is reached when the technology becomes widely accepted (Gartner, 2019). In fact, there is already a hype cycle for the field of AI (Figure 14).

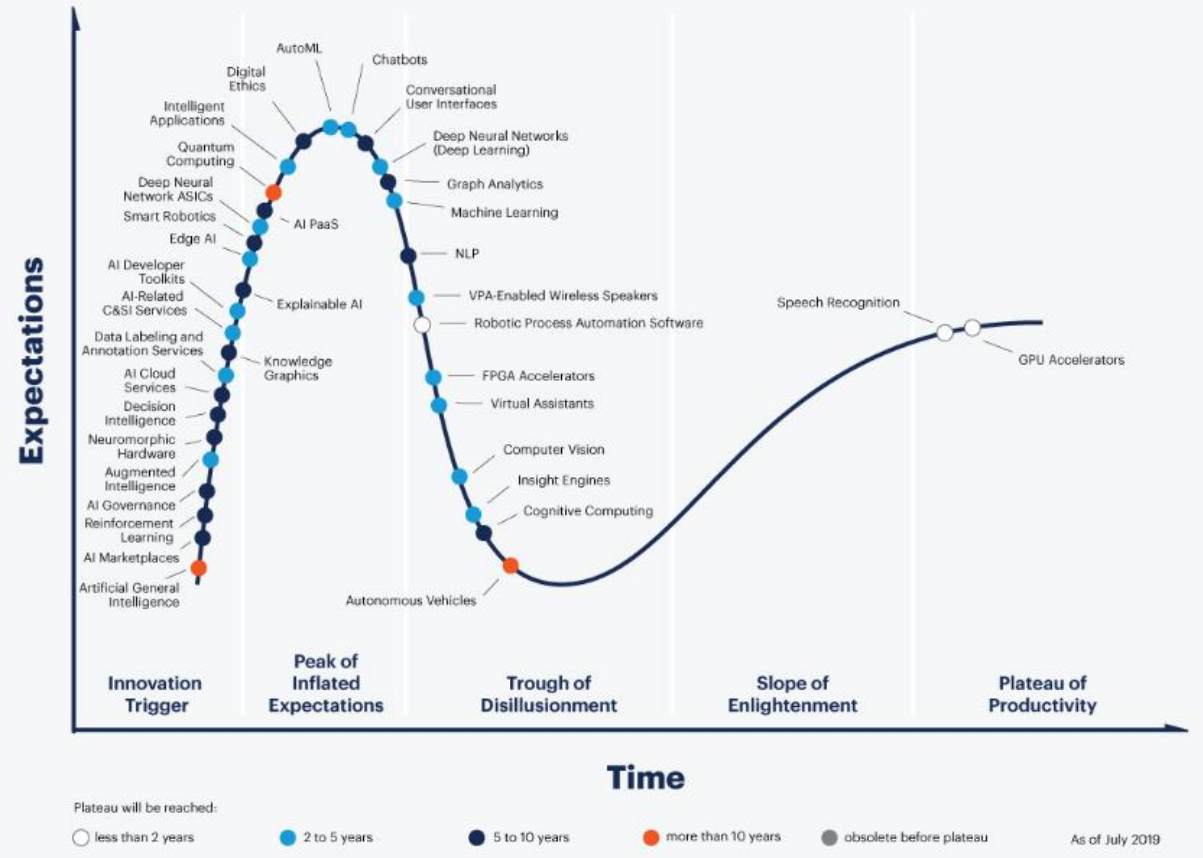

Figure 14. Hype Cycle for AI (Gartner, 2019)

In figure 14, we can see that the two technologies that are considered the base for the development of AI in healthcare (Deep Neural Networks and Machine Learning) are at the end of the period of inflated expectations and on their way to the trough of disillusionment. In the past, purely hopeful expectations have represented a big problem for AI, as they led to the first AI winter (McDermott, 1976). If AI follows that path again, the concerns of interviewee F could materialize and the trust in the technology could be lost for a long time. We suggest that Philips should work together with the other companies developing AI-systems for healthcare to clarify the limitations that the technology has in order to create expectations that can eventually be fulfilled. Although it is difficult to avoid a period of mistrust in the technology, offering achievable objectives could help to decrease the length of the disillusionment stage.

5.4 Risks

The risks were identified by analyzing the challenges for the introduction of AI in healthcare. These risks were divided into technical, ethical and organizational.

The two main technical risks are misdiagnosis and privacy breaches. The main sources for them to happen are biased data and lack of proper cybersecurity measures. However, the experts expressed that these risks are not likely to happen in Philips for two reasons. First,

because the company has a very strict procedure of data selection and curation, which diminishes the chance of introducing biases. Second, because the interpretation of the GDPR used by the company is quite stringent and the cybersecurity measures are very strong, minimizing the risk of privacy breaches. Nevertheless, some studies have shown that although the biases present in the data can be minimized, there is no fully proved way to avoid the biases that could be embedded by software developers (Johnson, 2012; Macnish, 2012). Those 'loaded biases' could lead to misdiagnosis. Similarly, unfair commercialization of patient data was also not considered a big risk for Philips. This is due to the strict measures of privacy and security enforced by the company. There is no way to extract information without the company knowing it (USBs and document-sharing platforms are not allowed in Philips' computers).

The interviewees expressed that more effort should be put into dealing with the ethical risks. First, the lack of diversity in software development teams could lead to the risk of unintended discrimination. The experts established that diversity should be ensured, even if more budget has to be allowed to recruitment teams. Besides, they think that the risk of inequality in healthcare could be even reinforced with the introduction of AI. They believe that it is time for healthcare technology companies to start working together with governments from poor countries to digitize health care records, and to introduce those records in the training of AI algorithms.

There are also ethical risks of social isolation and loss of patient autonomy related to telehealth. The experts indicated that autonomy should never be compromised and that telehealth treatments should never be imposed against the will of a patient. If people feel that a telehealth treatment will socially isolate them, they should be able to take the decision to decline it and get a traditional treatment in the hospital. The vision of the experts agrees with the "*Principle of respect of human autonomy*" that makes part of the Ethics Guidelines for Trustworthy AI (European Commission, 2018).

Finally, the interviewees indicated three main organizational risks: Opposition from the medical community; loss of trust from HCPs, patients and/or public; and loss of reputation. The opposition from the medical community was not considered a risk for most of the interviewees. The experts said that AI technology should be framed as augmenting the capabilities of the clinicians instead of replacing them to avoid issues of acceptability. In addition, they argued that AI technology should be properly explained to healthcare professionals, including its limitations. They said that it will increase the acceptability because HCPs would be aware of what to expect. However, these result might be influenced by the selection bias from the choice of the interviewees. All of them are Philips employees and there is a possibility that their answers could be aligned with what the company wants the public to believe. Besides, new technologies frequently deal with the so-called 'Collingridge dilemma' which claims that the possibility to responsibly guide a technology is greatest in its early stages of development, when the potential consequences are still unknown, whereas it becomes very difficult to steer once it is fully embedded in society and its effects have

become manifest (Collingridge, 1980). In this case, AI is a novel technology and many of its effects are still unknown. Therefore, explaining the limitations to HCPs could only help to address the risks already identified, but not the risks that might arise in the future. Some negative effects could become manifested but there could be already too late to fix them, which might hamper the acceptability of the technology.

The second organizational risk is the possibility of losing the trust of doctors, patients, and society in case that AI does not deliver as expected. To avoid that, the experts recommended that "window-dressing" practices should be avoided at all costs. According to them, inflated expectations only lead to incomplete solutions. However, in sub-section 5.3 we explained that a period of mistrust in AI is difficult to avoid as it corresponds to the "hype cycle" of technology development (Gartner, 2019). In addition, the two technologies that are considered the base for the development of AI in healthcare (Deep Neural Networks and Machine Learning) are already close to the period of disillusionment. For that reason, we suggest that instead of avoiding that period at all costs, Philips should aim at diminishing the length of that stage. By defining together with other companies the expectations that could eventually be fulfilled, the arrival to the period of enlightenment can be accelerated.

Finally, the risk of losing reputation could be considered as an overarching consequence of any misbehavior of AI or any unethical behavior from Philips employees. Interviewee C stated, "*we have something to lose and we are a trusted brand. If you lose it once, you lose it forever*". There have been big headlines of privacy scandals on Facebook (Forrest, 2019) and Novartis (Nowatzke, 2019). These scandals are very detrimental to the image and reputation of any company.

5.5 Barriers

As discussed in the non-ethical challenges, all the experts agreed that the main barriers to AI development in healthcare are regulatory. The lack of clear policies regarding patient data ownership, open algorithms, and GDPR interpretation creates an uneven innovation environment. According to the experts interviewed, some companies and countries are more willing to take risks and exploit the gaps in regulation. However, European companies prefer to play on the safe side. In fact, the former Google executive Kai-Fu Lee has stated that Europe has no chance in the AI race due to the stringent regulations (Minsky, 2018). However, as discussed in section 5.2, there is no indication that that European regulations will be soften in the near future and the scientists should adapt to them (when developing AI systems in Europe). To tackle regulatory barriers, Philips has established AI research centers in the US and India, where regulations are more lenient. However, the internal policies of the company are still very risk-averse and the company prefers not to risk patient's data to achieve faster results. The risk-averse attitude and the lack of flexibility of Philips are seen by the interviewees as two important organizational barriers that affect AI development.

In addition, the experts identified technical barriers such as the lack of methods to understand the "black-box" in machine learning applications, the difficulty to integrate new AI systems in the existing infrastructure, and the fact that AI systems have to be much better than humans are. The interviewees again emphasized that technology development never stops and that these barriers will be overcome sooner than later. However, the literature indicates there is still a long way before some of those challenges can be solved. On the one hand, there is no way to open strong black boxes such as the open algorithms for decision-support in healthcare (Bathaee, 2018). On the other hand, the lack of generality in artificial intelligence could create risks when AI systems are confronted to unexpected situations (Bostrom, N., & Yudkowsky, 2011). These barriers have to be taken into account by developers when creating expectations for AI systems in healthcare.

5.6 RRI actions

Philips has multiple RRI practices already in place. However, the RRI actions are implemented in the company under different names, such as CSR and business principles. The fact that companies are unaware of the RRI concept but have RRI practices already embedded in technology development is common in industry according to the literature (Auer & Jarmai, 2017; Dreyer et al., 2017). One of those actions, was the establishment of the Internal Committee for Biomedical Experiments (ICBE), which evaluates that every product developed in Philips Research complies with ethical, legal and privacy policies before putting it in the market. The committee started as an ethical board to supervise clinical trials, but extended from there to evaluate the ethical, social and legal compliance of all the products developed in Philips Research. It includes legal representatives, privacy officers, data officers, IP officers, scientists, clinicians, CSR officers, quality officers, and members of the management of the company. This committee has an influential role in the development of AI within the company. According to interviewee A, "*In 2018, the ICBE committee reviewed around 234 studies and 40% of those were datasets studies, which were primarily linked to AI. That gives an indication of how much work is done. For this year, the number is even higher; the percentage for AI is the largest component between all the studies in Philips Research. A lot is being done across Philips Research globally*". In addition, the company is developing the first version of the code of ethics for data management and AI. Different groups within Philips are evaluating this code and the results of this thesis will be taken into account for further refinement of the code.

In addition, the company has an inclusive vision of AI by defining it as "adaptive intelligence" instead of artificial intelligence. Philips aims at generating AI outcomes that support and empower healthcare professionals, but the vision of replacing them has never been considered. According to the doctors working in the company, the idea of substituting the doctor is just impossible. Interviewee D said "*there is no way you can take away the interaction between the doctor and the patient in diagnosing treatments. You will never have a screen telling you that you have 6 months to live by our estimations. This is a complete dehumanization of care. Society will never allow that to happen".*

Regarding the introduction of unintended biases, Philips has a dedicated procedure of data selection and curation in which doctors and nurses play an important role. They work together with the scientists in assessing the data that should be used to train decision-support algorithms. This slow but rigorous process allows Philips to take into account contextual information from the medical world. According to several interviewees, many companies working on AI for healthcare still fail to deliver in this aspect. However, the experts agreed that this work of selection and curation could be wasted if the software developers introduce their own biases in the algorithms. The company is trying to ensure diversity in the AI development teams, but there is still no effective strategy to ensure that developers do not introduce biases.

Philips also aims at having an inclusive culture of technology development by setting co-creation exercises with several stakeholders. Interviewee D said "*we put together hospitals, patients, former companies and other device companies together in a room and we ask the following question: what kind of product would be nice for you?*" and interviewee B stated, "*We do a lot of co-creation sessions with physicians and patients. We talk to each other to avoid surprises. They contribute to the ideation of products*". These exercises are very important because if a wide range of stakeholders are taken into account and their values introduced in the phase of product design, the acceptability of the solution will be greatly enhanced (Owen et al., 2012).

Regarding responsiveness, the company has a skilled cybersecurity team that is continuously developing new measures to avoid potential attacks. Besides, privacy officers are involved in the entire process of technology development, from design until release to the market, ensuring that privacy is always secured. Finally, the company takes into account the feedback from end-users (doctors and patients) very seriously and make modifications in the new versions of their products. However, scientists feel that making changes once a project has started is quite difficult and that they have to adjust to the initial design, even if they find better options along the way. Philips can make some improvements in this aspect.

The interviewees mentioned 48 RRI actions that should be deployed for the responsible introduction of AI in healthcare (see Section 4.5). The experts indicated that from those 48, there are 15 actions considered as the most important to ensure social acceptability. The author agrees with the actions formulated by the experts and introduces two more actions based on the triangulation with the literature review. First, keeping the role of radiologists as the point of contact between the machine and the patient, and second, working together with other companies to set achievable goals and reduce the length of the disillusionment period of the Gartner hype cycle. The 17 actions are described below.

Anticipation & Reflection

1. ***Design procedures to ensure that data has a high-quality, it is well-curated and representative of the population***

   This action has been implemented quite well by Philips. Healthcare professionals and scientists work together curating patient data and then combining it with medical data from publications. Subsequently, hypotheses are formulated and validated. Despite being a slow approach, it brings important benefits such as the introduction of contextual knowledge, minimizing the risk of unintended discrimination and avoiding misdiagnosis. Data officers and the ICBE must ensure that this activity remains as a core process in AI technology development.

2. ***Establish clear and rigorous practices of data management regarding privacy and consent***

   According to interviewee A, "*if people feel that a new technology threatens their privacy, they will be less willing to accept the new technology*". In the case of AI for healthcare, this issue plays even a bigger role because medical data is considered to be sensitive and private (Nuffield Council of Bioethics, 2018). Clear policies have to be introduced to ensure that privacy is never compromised and that patients consent the use of their data. Philips is very strict in this area. For example, AI scientists have to explain to the ICBE why they need the data and what they are going to do with it. If the ICBE considers that the proposition is not strong enough, the committee will not grant access to the data. Besides, once a project has started, privacy and data officers are involved during the entire development process verifying that patient data is being used only for the initial intended purpose.

Inclusiveness

3. ***Design cross-population studies to validate AI applications***

   AI decision-support systems in healthcare are normally trained with data that is representative of some specific populations. This increases the risk of discrimination and unequal distribution of care because the products will only work for some people. There is also a chance of misdiagnosis because if a product is trained with data from a region and deployed in a different region the outcomes could be completely different. Currently, Philips performs cross-population studies to validate AI applications in some countries. However, the company still has to work harder to ensure that their AI products could also be introduced in countries where data has still not been digitized.

4. ***Ensure a high level of diversity in AI software development teams***

   Unintended biases could be embedded in AI algorithms if diversity is not ensured in software development teams. This might lead to unintended discrimination because developers could inadvertently introduce biases based on their prejudices and beliefs (Academy of Medical Royal Colleges, 2019). Finding equilibrium in gender, race, nations, and religions within the development teams is not an easy task. However, if biases are avoided, the social acceptability of AI could be greatly enhanced. For this

reason, the hiring process should be aimed not only at finding the best talents, but also at ensuring diversity.

5. ***Involve clinicians and patients in the process of technology development***
   The values and points of view of doctors and patients must be taken into account. In the end, they are the end-users who decide whether technology is embraced or not. Philips normally sets co-creation exercises where different stakeholders help the scientists in ideating products. However, the involvement of the clinicians and patients should not only happen at the beginning of the design process. They should be an important part of the entire process of technology development. This will help to identify potential risks on time.

6. ***Make sure that healthcare professionals are included in the process of selecting and curating the data***
   Philips performs an excellent job in this action. Doctors, nurses and administrative medical personnel are always included in the process of data selection and curation. This helps the company to introduce contextual knowledge from the medical world in their solutions. This activity should continue in the long-term.

Responsiveness

7. ***Implement internal policies that allow AI scientists to make quick changes once a project has started***
   Most of the interviewees defined Philips as a complex and inflexible company, where making changes is quite difficult. They argued that sometimes they identified that a product needed some improvements, but that making the changes once a design was established was so tedious that they preferred to wait and implement the changes in the new version. This approach makes it difficult to deal with unexpected problems or changes in regulation because there is no flexibility to set a new direction. This intensifies for the case of new technologies such as AI, where problems and new regulations appear frequently. The company has to develop internal policies to increase flexibility in product design and development.

8. ***Develop activities of innovation and experimentation without regulatory constraints (bootcamps, hackathons…)***
   Scientists need spaces to develop and test new ideas without all the regulatory constraints from the medical world. Bootcamps and hackathons are a good way to stimulate experimentation and innovation by allowing the developers to create solutions for specific challenges. These solutions do not have to take into account GDPR or FDA compliance; they only have to tackle the proposed challenge. Interesting ideas can come out from these exercises and then, they can be refined in the R&D department.

9. ***Develop and implement AI solutions to detect and avoid biases from software developers***

   Ensuring diversity in software development teams and curating the data are not enough measures to prevent biases in AI. IBM is already developing tools to detect and mitigate biases in machine learning algorithms (Fay, 2019). They argue that there are biases that even the most trained human cannot detect, but that can be easily spotted by AI. It would be interesting for Philips to start working on such technologies to improve its bias-avoidance measures.

10. ***Encourage an agile-way of working***

    Agile practices should be encouraged in the different teams and departments within Philips Research. By working agile, quick changes can be made to steer the responsible development of AI solutions. Scientists from different departments can work together in some projects in order to share the knowledge of agile practices across the organization.

11. ***Ensure that data scientists work in teams and that there are peer-review discussions going on***

    This action also aims at diminishing biases in AI algorithms. Besides, it allows avoiding misdiagnosis by constantly checking the validity of the outcomes with other scientists. Scientists that come from different backgrounds and cultures could interpret results in a different way. For that reason, peer-review discussions can reduce the risk of unintended discrimination by including different angles when analyzing the potential of a solution.

12. ***Educate the HCPs on how to use the AI-enabled systems and on how to interpret the "black-box"***

    This action aims at avoiding the risk of 'automation bias'. Healthcare professionals should be properly trained to use diagnosis-support systems. Videos, classes, and workshops can be organized to explain what are the inputs that the ML algorithms require, how the variables inside the "black-box" might affect the outcome and how to interpret the results.

13. ***Implement interaction features in telehealth applications***

    Most of the telehealth applications currently in the market, including Philips CareSage, focus on monitoring patient health. When unusual values are identified, a warning message is sent to the clinician to take further action. However, the use of this technology can cause a sense of social isolation for some people, especially elderly patients who do not have relatives taking care of them (Sharkey & Sharkey, 2012). Some of them may prefer to stay in the hospital to have some social contact, but cannot afford the treatment there (Sharkey & Sharkey, 2012). Telehealth applications are still lagging behind in the implementation of interactive features for their patients. Interviewee A stated "*we have underestimated the value of social*

networks. A social network where telehealth patients can interact could be a nice way to avoid social isolation of patients" and interviewee B expressed "*communication in telehealth can be improved with screens, virtual reality or other technologies so that isolated patients can communicate with doctors and nurses at a different location. This can enhance engagement from the patients*". Philips should start working on communication features for its telehealth applications to avoid the risk of social isolation of patients. Constant interaction with healthcare professionals and other remote patients can enhance the acceptability of telehealth.

14. ***Work on new ML technologies to make the "black-box more transparent"***

    All the experts interviewed agreed that opening the "black-box" in machine learning algorithms is extremely difficult and that we should accept the results without understanding the process behind them. However, Harvard and IBM already started working on systems intended to open the "black-box" (Dickson, 2018). So far, they managed to understand the sequence followed by translation apps powered by AI. In the field of healthcare, opening the "black-box" will represent a major breakthrough because patients will have an understanding of the reasoning behind their diagnoses. This will allow them to take autonomous informed decisions related to their health, which complies with the autonomy principle of medical ethics (AMA, 2019). Making the "black-box" transparent will take a long time, but Philips should be one of the companies working on this because the company visualizes AI as the future of healthcare (Philips, 2019e). Interviewee F said, "*In 20 years we expect that all of our products have a component of AI*". If the strategy of the company is to focus on AI, then the explicability of their algorithms should be a priority.

15. ***Work closely with legislators on addressing legal challenges***

    All the experts agreed that regulation for data management and AI still requires a lot of improvement. Either because there is a lack of regulation in some issues or because the regulation already implemented is too stringent. Philips makes a great job of working together with the legislators to address the challenges of policies for AI in healthcare. For example, interviewee C expressed "*we are collaborating with the FDA and other companies to establish the regulation for dynamic algorithms that keep learning in the field when new data is fed*". This collaboration should be paramount because companies can help to steer the regulation to avoid unnecessary measures that could stifle innovation. Besides, this collective work allows for a better understanding of the regulations as the scientists are directly involved in interpreting the policies.

16. ***Ensure that radiologists are the point of contact between the machine and the patient***

    In case that the replacement of radiologists in the activity of image analysis becomes unavoidable, these professionals should keep their role as the point of contact between the AI system and the patient. The technology developers must ensure that their

products are easily understandable by the radiologists and that they enhance the abilities of the clinicians to deliver care and guide the patients in their treatments.

17. ***Work together with other companies to set achievable expectations for AI***
    Philips should work together with other companies developing AI technologies for healthcare to set realistic expectations of technology development. If the limitations are clarified and the objectives seem achievable, the length of the disillusionment period of the Gartner hype cycle could be reduced.

5.7 RRI limitations and different approaches

The first limitation that we found in the interviews was that the terminology used in the field of RRI is unknown for people working in industry. We asked the interviewees if they had implemented practices of RRI in the company and nobody was aware of the term. Besides, we asked the experts what were the activities developed by Philips in the dimensions of RRI (Stilgoe et al., 2013): anticipation, reflexivity, inclusiveness, and responsiveness. However, there was no clarity of the concepts and we had to explain and give examples of practices for every dimension. Once the examples were given, the interviewees understood the meaning of the terms but still referred to the practices developed by Philips as risk assessments, participation exercises, CSR, ethical behaviour, agile-way of working, and business principles. Dreyer, et al. (2017), identified the same problem and proposed that the RRI methodology should be better aligned with industry practices like Design Thinking, Business Innovation Canvas, Risk Management, and Innovation Project Management. Auer and Jarmai (2017) also discovered that SMEs working in the medical device industry were also unaware of the RRI concept. They proposed that companies should build upon their existing CSR practices to further develop RRI practices. Interviewee G also mentioned that introducing a set of responsibility actions under a different name from CSR and business principles could cause confusion in the research community, affecting its acceptance within the company. For that reason and according to the literature results we advise that RRI actions should be reframed in companies under the name of CSR practices instead of RRI.

The second limitation is described by Sonck, et al. (2017). They argue that RRI assumes a transparent and smooth engagement of stakeholders, where every member has the same information and the same decision power. However, the business world is characterized by taking risks based on power and information asymmetries. This is also mentioned by de Hoop, et al. (2016), who identified that there are power differences and dependencies as well as irreconcilable interests that are overlooked by the RRI academics. Interviewee D expressed a related concern. He claimed that the process of defining the regulation in AI and healthcare is highly political and that some stakeholders have much more power than others. For example, he defined that institutional bodies and companies have a big number of representatives when defining regulation but that it is very rare to see a patient or non-experts (public) taking part of the discussions. Our suggestion for this issue is based on the work of Brower, et al. (2013) who concluded that low-power stakeholders should broaden their power

base by connecting and engaging with other stakeholders. He expressed that sometimes stakeholders should support non-beneficial actions in order to gain allies for the future. This strategy clearly differs from the transparent and smooth process of RRI, where all the stakeholders express their true intentions. We propose that this strategic behavior can be useful to introduce the interests of patients and public in the development of regulations for AI in healthcare.

For the specific case of the PRISMA roadmap, the interviewees offered two main recommendations. First, the PRISMA roadmap should include a section in which the user can analyze which risks are worth taking. According to interviewees A, D, E, and H, the roadmap focuses on designing RRI actions to avoid the risks that technology development might bring, but actually, some risks are justified. According to them, it will make more sense from a business and social perspective to balance goals and risks instead of only focusing on avoiding risks. For example, interviewee E said that "*the goal of creating a huge improvement in global health by using AI could justify the risk of potential privacy breaches*". In a different example, interviewee D expressed that "*even if we do not understand the 'black-box', the results obtained are very valuable. There is no need for explicability when there is a potential to save human lives*". However, these results show an important gap between the principles of RRI and the risks that some people at Philips consider acceptable. The Ethics Guidelines for Trustworthy AI (European Commission, 2018) indicate that privacy and explicability are key principles for the acceptability of AI and hence, the assurance of these principles must not be jeopardized. The idea of balancing risks and benefits seems interesting for further versions of the roadmap, but it must be highlighted that this step does not compromise key ethical principles of the technology studied.

The second recommendation given to the roadmap applies mainly to big companies with strict quality and regulatory practices in place. Van de Poel, et al. (2017) suggested that in order to review the process of RRI implementation it was useful to implement Key Performance Indicators (KPIs) to track the introduction of RRI practices. However, most of the interviewees established that Philips products already have to fulfil hundreds and in some cases 1000s of KPIs in terms of cost, intellectual property (IP), design, manufacturability, quality, procurement, sustainability, marketing, etc. For that reason, they concluded that adding more KPIs would only slow the process of innovation. Besides, they said that it could lead to "greenwashing" because the scientists will just be checking the boxes for the RRI KPIs without really given them the adequate importance. The interviewees recommended that is better to have continuous meetings with the ethical board of the company (ICBE in the case of Philips) to check if the RRI practices are being developed in the R&D processes. They argue that this interaction could generate more awareness in RRI than checking a few of the thousands of KPIs that a product might have.

6. Conclusions

This research provided a first step towards the study of how responsible research and innovation (RRI) practices can be implemented in the field of AI for healthcare. We analyzed the case of Philips and identified the RRI actions that are already being developed and the actions that should be further incorporated to improve the social acceptability of its AI-enabled products. In addition, we analyzed in which cases the RRI tools could fall short and suggested alternative approaches. This study helped us to realize that AI will bring an important paradigm shift to the field of healthcare. The benefits are numerous and undeniable but there are also key ethical risks that should be avoided. For that reason, we encourage companies working on AI technologies to adopt the RRI actions presented in this work to prevent potential risks and to steer the technology development towards a good end.

This section starts with the answers to the research questions. Then, we present the academic and practical contributions of the research. The next sub-sections consists of the limitations and the future research recommendations of this study. Finally, the author describes the fit of the research in the Management of Technology Program and elaborates a personal reflection of his work.

6.1 Answering the research questions

Sub RQ1. Which are the chief ethical concerns that threaten the social acceptability of AI in healthcare?

In the literature review, we identified the chief ethical concerns that threaten the social acceptability of AI in healthcare: accountability; data bias, fairness, and equity; effects on healthcare professionals; effects on patients; privacy and security; reliability and safety; and transparency and trust.

In ***accountability***, the main concern is to determine who should be held responsible if AI goes wrong. According to the current regulation, clinicians are responsible as they take the final decisions regarding diagnoses and treatments (Hart, 2017a). However, if an AI decision-support system is correct 98% of the time, doctors might fall in the ‘automation bias’, which is the human tendency to trust machines more than their own judgment (Skitka et al., 2000). In this case, complications can appear because clinicians may be justifying wrong diagnoses made by AI machines.

In ***data bias, fairness and equity***, the main risk is that ML algorithms could be trained with incomplete and unrepresentative data. This could lead to discrimination based on race, gender, age, socioeconomic status, etc. (Bird, et al., 2016). Similarly, software developers can

'load' unintended biases in the algorithms, reflecting their beliefs and prejudices (Academy of Medical Royal Colleges, 2019). Furthermore, regions of the world where there are still no electronic health records could be left aside from the benefits of AI in healthcare as their data cannot be used to train the AI systems (Hart, 2017b).

The implementation of AI will undoubtedly bring ***effects on healthcare professionals***. The medical profession is expected to change from a diagnostic-oriented approach to a care-oriented approach (Academy of Medical Royal Colleges, 2019). This concerns physicians because they may feel that their autonomy and expertise could be threatened by AI systems (Hamid, 2016). Besides, the automation of some medical procedures might induce hospitals to employ less skilled staff to perform some of the work currently done by clinicians (Nuffield Council of Bioethics, 2018).

The ***effects of AI on patients*** are mostly related to the development of telehealth. Some patients might feel socially isolated and left at home dealing with devices instead of people (Sharkey & Sharkey, 2012). Moreover, the autonomy of the patient can be put at risk if telehealth is enforced to cut costs at hospitals. People should always have the option of going to the hospital and talking to healthcare professionals if they prefer that alternative.

In ***privacy and security***, the chief concerns are the risk of data to be stolen and the potential to suffer cyberattacks. ML algorithms in healthcare need a huge quantity of patient data to be trained. These data are considered sensitive and private (Nuffield Council of Bioethics, 2018). For that reason, hackers find it valuable to steal this information in order to resell it or to blackmail people (Academy of Medical Royal Colleges, 2019). Besides, AI might be used to carry cyber-attacks in hospitals, causing huge harm at a minimal cost (Bhatnagar et al., 2018).

The ***reliability and safety*** of healthcare outcomes is another concern when introducing AI in medical settings. In some cases, AI can be wrong and cause big harm across the healthcare system (Ross, 2018; Caruana et al., 2015). There is a huge risk if ML algorithms do not take into account the contextual information of the medical world and deliver wrong advice (Academy of Medical Royal Colleges, 2019).

Finally, software developers still have a lot of work to do to ensure ***transparency*** in AI. Machine learning algorithms are usually described as a "black-box" because until now there is no way to explain the logic behind the algorithms to reach an outcome (Anderson & Leigh, 2019). Then, it makes it difficult for physicians to justify the outcomes and for patients to ***trust*** the system. In addition, there are uncertainties about private companies accessing patient data. Some people argue that they do not trust AI technology developers to put the public interest over financial rewards (Perrin & Mikhailov, 2017).

Sub RQ2. How is Philips currently addressing the ethical issues raised by the introduction of AI systems?

Philips has already developed significant practices to deal with the ethical issues raised by the introduction of AI in healthcare. There are important RRI actions already in place but they are framed under the names of CSR and business principles.

A very important step to ensure that ethical issues are taken into account was the creation of the Internal Committee for Biomedical Experiments (ICBE). This committee ensures that every single product developed in Philips Research complies with ethical, legal and privacy policies before launching it to the market. This committee has an influential role in the development of AI within the company. In addition, the head of strategy for data science and AI took the decision of compiling a code of ethics for data management and AI. The idea is to have an internal guideline with clear principles that could be used by scientists to steer the development of their AI-enabled products. This code will allow scientists to have a clear manual to base their decisions instead of trying to interpret European regulations on their own. The code is based on the "Ethics Guidelines for Trustworthy AI" (European Commission, 2018) and covers the same ethical principles: "1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination and fairness, 6) environmental and societal well-being and 7) accountability". Until now, the code is focused at establishing the principles for the ethical development of AI in Philips. The results of this research will be used to give a step further and go from principles to practices.

Philips has also established a rigorous procedure of data selection and curation. This process aims at avoiding biases from datasets. Healthcare professionals and scientists work together curating the patient data. Then, this information is combined with medical data from publications. Subsequently, hypotheses are formulated and validated. This slow but steady approach generates important benefits such as the introduction of contextual knowledge from the medical world, the minimization of unintended discrimination and the avoidance of misdiagnosis.

The company also aims at having an inclusive culture of technology development by setting co-creation exercises with several stakeholders. Usually, doctors, patients, and potential customers are taken into account to contribute to the ideation of new products. In addition, Philips considers the feedback from end-users (doctors and patients) very seriously and makes modifications in the new versions of their products. Moreover, the company has alliances with companies that have expertise in other fields in order to accelerate the process of innovation. For example, Philips is working together with Microsoft to develop augmented reality tools that could help physicians to perform surgeries (Philips, 2019g).

Finally, Philips has a skilled cybersecurity team ensuring the privacy and security of patient data. Data risk assessments are performed periodically to prevent breaches in Philips

databases. Besides, privacy officers are involved in the entire process of technology development, from design until release to the market, ensuring that scientists respect the principles of data privacy established by the company. Moreover, data transfer tools such as USB sticks or online transfer platforms do not work on the company's computers.

Sub RQ3. How should Philips introduce RRI practices to avoid the risks and increase the social acceptability of its AI-enabled products?

The following RRI actions are recommended to avoid the potential risks of AI systems in healthcare and to increase their social acceptability:

1. Philips should continue its efforts in ensuring that the patient data used to train the AI algorithms has a high quality, is well-curated and represents the diversity of the population. Doctors, nurses and administrative medical personnel should always be included in the process. This action brings important benefits such as the introduction of contextual knowledge from the medical world, the minimization of unintended discrimination and the avoidance of misdiagnosis.

2. Clear policies have to be introduced to ensure that privacy is never compromised and that patients consent the use of their data. Philips should continue investing in its cybersecurity team to avoid potential attacks. Besides, privacy officers should verify periodically that AI scientists do not use patient data for unintended purposes.

3. The company should keep working closely with legislators to shape the regulation of AI in healthcare. Currently, European regulation regarding data and privacy (GDPR) is very strict. These policies are affecting the innovativeness and competitiveness of European companies in the healthcare space. Philips can join other companies working on AI in healthcare to lobby for less stringent regulations.

4. Cross-population studies to validate AI applications should be performed. It decreases the chances of misdiagnosis by ensuring that AI systems are only deployed in populations that represent the data used to train the algorithms.

5. Philips should work together with healthcare providers in poor regions of the world to help them digitize their medical records. By doing that, data of poor populations can be included in training the ML algorithms. This decreases the risks of discrimination and inequality that could be introduced by AI tools.

6. Recruiters and project leaders must work together in ensuring diversity in AI software development teams. This will help to avoid unconscious biases and unintended discrimination. Besides, developers should never work alone in a solution and there must be frequent peer-review discussions going on to validate the outcomes of the algorithms.

7. Philips should work in technological solutions to avoid biases from data or from software developers. Some companies are already working on tools that detect and mitigate biases in machine learning algorithms and Philips should not be the exception.

8. Health care professionals and patients should not only be involved in the process of product ideation, but also during the entire process of product development. This will help to identify potential risks on time and increase the acceptability of AI technology. In the end, they are the end-users who decide whether technology is embraced or not. Radiologists should play an important role in this activity because they will be the most affected by the introduction of AI in healthcare.

9. Philips should implement practices that allow AI scientists to make quick changes once a project has started. Besides, bootcamps and hackathons to solve challenges without regulatory constraints must be organized frequently. These actions can help the company to develop strategies to deal with unexpected problems and to set new directions for innovation.

10. The company should encourage an agile-way of working in all the teams within Philips Research. Scientists from teams or departments recognized for being agile can be mixed in some projects with scientists from departments where agility has to be improved. In that way, agile-practices can be shared across the organization.

11. Future telehealth systems developed by Philips should include better interaction features to avoid the social isolation of users. Internal social networks, screens, or virtual reality technologies could be used to enhance the communication and engagement of patients.

12. Philips should work on the education of healthcare professionals to ensure that they understand properly how to use diagnosis-support systems and how to interpret to some extent the logic behind the "black-box". Workshops, training, or continuous support on-site could be some of the alternatives to educate the HCPs in the use of AI technologies. The limitations of the AI-systems should also be highlighted to avoid potential 'automation biases'.

13. According to the autonomy principle of medical ethics, patients should be able to make autonomous and informed decisions regarding their health (AMA, 2019). Because of this, Philips should start working on making the "black-box" of ML technologies more transparent. Opening the "black-box" will represent a major breakthrough because patients will have an understanding of the reasoning behind their diagnoses.

14. Philips should work together with other companies developing AI technologies for healthcare to set realistic expectations of technology development. If the limitations

are clarified and the objectives seem achievable, the length of the disillusionment period of the Gartner hype cycle could be reduced.

Sub RQ4. What situations require a different strategy from RRI to address the risks of the implementation of AI in healthcare?

From the analysis carried out in sub-section 5.7, we identified two situations in which a strategy that differs from RRI offers a better approach to enhance the introduction of responsible practices in AI for healthcare.

1. In the case of companies, it is recommended to reframe RRI concepts and use terms more related to the private sector. This terms could be CSR, business principles or business ethics. In addition, Auer and Jarmai (2017) propose that companies should build upon their existing CSR practices to further develop RRI practices. Interviewee G also expressed that the use of unknown terms could affect the acceptability of responsible practices in the company.

2. Low-power stakeholders in the process of regulation design for AI in healthcare, such as patients or the public, should broaden their power base by connecting and engaging with other stakeholders (Brouwer, Hiemstra, van der Vugt, & Walters, 2013). This strategic behavior allows them to counterbalance the power and information asymmetries present in a deliberation process. These stakeholders have a better chance to express their values and interest if they are part of a group with high influence.

Sub RQ5. How can the PRISMA roadmap be improved to strengthen its applicability in large corporations?

The interviewees offered two main recommendations for improvement of the PRISMA roadmap:

1. The risks can be divided into the risks that are worth taking and the risks that should be avoided at all costs. It helps to balance business goals and challenges. However, it must be stressed that this step should not compromise the key ethical principles of the technology studied.

2. Meetings with the ethical board are considered a better strategy than KPIs to assess RRI implementation in large companies. These corporations have already thousands of KPIs defined for a single product and there is a risk that the RRI KPIs could end up being used only for “greenwashing”. The interviews claimed that frequent discussions with the ethical board of the company could generate better awareness of responsible practices.

Main RQ. How should RRI be framed and introduced to avoid the risks and enhance the social acceptability of AI-enabled products developed by large healthcare technology companies?

According to this research and to the results of the study of Auer and Jarmai (2017), SMEs and large companies are unaware of the concept of RRI. However, these enterprises have RRI practices already in place under different names. We suggest that RRI activities should be framed as CSR practices or business principles and that further RRI development should build upon these concepts.

For the case of AI in healthcare, there are different RRI activities that can be framed under the name of CSR to avoid potential risks such as unintended discrimination, unequal distribution of care, social isolation or privacy breaches. For example, the process of selection and curation of data to train AI algorithms must be rigorous and inclusive. Healthcare practitioners must play an important role in this action because they have the contextual knowledge of the medical world. Besides, capable teams of cybersecurity and data privacy should be formed to prevent attacks and privacy breaches. Companies should work on ensuring inclusiveness in two ways. First, by including the data from poor populations and regions to train ML algorithms and second, by ensuring diversity in the software development teams. Besides, clinicians and patients should be included in the entire process of technology development and not only in the initial phases of product design. They are the end-users of the AI products and they take the decision on what products they want. Moreover, better interaction features should be included in telehealth systems to reduce the risk of social isolation and enhance the engagement of patients. Finally, companies should start working on making the "black-box" of machine learning algorithms more transparent. This will help patients to take informed decisions about their health, making the technology more acceptable.

For the specific case of large companies, more flexibility should be incorporated in the operations of the R&D departments. This will take some time as it requires changing the way of working of the companies but the benefits are worth it. The responsiveness of the companies towards unexpected problems or regulations can be improved if there are more activities of experimentation and innovation. Hackathons and bootcamps can be used to solve challenges without restraining the AI scientist to the internal regulatory boundaries of the organization.

Another action is to encourage patients and the general public to broaden their power base by connecting and engaging with other stakeholders in order to counterbalance the power asymmetries when developing regulations for AI in healthcare. Finally, we recommend that the assessment of RRI compliance should be based in meetings with the ethical boards instead of the evaluation of KPIs. KPIs can work in startups and SMEs where processes are more flexible and there are not as many stakeholders and departments involved in product development as in large companies. However, the products in big corporations

already have to comply with thousands of KPIs and implementing even more KPIs could have a detrimental impact in innovation. In addition, if the RRI KPIs are not given the right importance, RRI could turn into a tool for "greenwashing".

6.2 Academic contribution

There are studies that have criticized the use of RRI in technology development. For example, Sonck, et al. (2017) claimed that RRI is naïve and idealistic. Dreyer, et al. (2017) expressed that there is not alignment between the terminology used in RRI and the concepts used in industry. Similarly, Hoop, Pols and Romijn (2016) concluded that there are important limitations to implement RRI in innovation such as power and information asymmetries, and material barriers. Besides, Auer and Jarmai (2017) analyzed the implementation of RRI practices in SMEs and concluded that although SMEs are unaware of the concept of RRI, these companies have RRI practices in place under different names.

This research extends from the work of Auer and Jarnai (2017) to study if the same results apply for the case of large corporations. We found that a similar situation applied to large companies by doing a case study at Philips. People working in the company were also unaware of the concept of RRI. However, several RRI practices were developed under the names of CSR and business principles. Our advice is that further RRI implementation in large companies should build upon the CSR practices already in place. According to the interviewees, relating responsible practices to familiar terms could facilitate the acceptability of RRI actions.

On the other hand, there is a contribution to the further refinement of the PRISMA roadmap when applied in large companies. First, the interviewees emphasized that the roadmap could change from a "risk prevention" approach to a "risk/benefit balance" approach. The risks could be weighed against the benefits to define what risks are worth taking and which ones should be avoided at all costs. However, we highlight that this step should never compromise the key ethical principles of the technology studied. Second, in the case of large companies, the assessment of RRI implementation could be better performed if it consists of discussions with the ethical board of the company instead of the revision of KPIs. Large corporations have already thousands of KPIs for product development and adding more KPIs increases complexity and slows the process of innovation.

6.3 Practical contribution

Most of the research in RRI has been oriented to offer governance recommendations to policy-makers in the public sector (Pacifico Silva, et al., 2018; Sun & Medaglia, 2019). It is an important topic in the academic world and in policy circles, however, only very few companies are aware of the concept and have implemented RRI practices in their innovation processes (van de Poel et al., 2017). This work had a very practical approach in order to start

generating awareness of the concept of RRI in industry. We analyzed the case of AI technology development in Philips and provided specific recommendations for the company. For the specific case of this work, the recommendations and RRI actions suggested will be taken into account in the refinement of the Philips' code of ethics in data management and AI. Besides, the roadmap designed will be included in the framework of "best practices" for the development of AI-enabled products in Philips Research. Finally, the results of the study will be presented in a "DoVo" (abbreviation for Donderdagochtend-Voordracht), which is a one-hour internal seminar series where the latest innovations and pieces of research are presented to the Philips Research community worldwide.

6.4 Limitations

The main limitation of this research was the fact that it only covered the case of one company working on AI for healthcare. This can introduce biased because the answers of the interviewees could be biased towards the culture and way of thinking of Philips. Besides, the interviews were limited to a number of ten people as they were the only experts available to take part in the research. This selection bias could affect the reliability of the results. For that reason, a process of triangulation with the literature was developed to reinforce or refute the results from the interviews. However, this only mitigated the bias but did not avoid it. We could not interview doctors and patients using Philips' AI products because the internal regulations of the company do not allow to disclose their identity. Organizing a workshop would have been a good alternative to get to know the point of view of other stakeholders. For these reasons, we cannot generalize the results of this thesis to be applied in the entire domain of AI development in healthcare. The results of the interviews only apply to Philips and only represent Philips perspective.

6.5 Future research

Further research can be carried out in more medical technology companies working on AI to assess how are they implementing RRI practices. In addition, it would be interesting if a study could focus on assessing the opinion of clinicians and patients regarding the use of AI for medical delivery. The answers of these stakeholders can be compared to the answers of people working in the companies to find a balanced approach that contributes to the creation of a comprehensive framework of responsible innovation for AI in healthcare. Research can also be done in more large companies to identify other situations in which RRI practices need to be reframed or in which a different strategy might be more suitable to tackle ethical issues. Finally, it would be interesting to perform the stages of experimentation and validation of the PRISMA roadmap to assess if the methodology is successful or if there are more limitations that were not taken into account. Finally, it would be interesting to check which of the methodologies for evaluating RRI performance ends up being better, KPIs or meetings with the ethical board.

6.6 Fit of the research in the Management of Technology program

The MOT program aims at improving the quality of technology and innovation management by educating responsible decision-makers (TU Delft, 2019). One of the main goals of the program is to emphasize the importance of the connection between society and technology. This works makes a good fit with the objectives of the MOT program because it analyzes how a novel technology (artificial intelligence) can be implemented in the field of healthcare in a responsible way. By analyzing the risks that AI might pose to society, we proceeded to formulate RRI practices to avoid the ethical risks and enhance the social acceptability of the technology. The MOT master encourages students to think beyond the financial indicators of performance in a company. Social and ethical considerations should also be taken into account in the business strategy of a company. This research fits in that vision, because the RRI roadmap designed is going to be included in the "good business practices" that steer the process of technology development within Philips.

6.7 Personal reflection

The elaboration of this work has been my first approximation to qualitative research. As an engineer, I focused most of my academic life on dealing with numbers and making quantitative studies. However, I discovered that qualitative research is a valuable methodology to obtain primary data. Especially, I found out that interviews provide a lot of information that would be almost impossible to get from literature or desk research. I really enjoyed performing interviews because in a short amount time of time I could learn as much as in long hours of reading articles. Besides, by working on this thesis, I learnt quite a lot about a topic that has always drawn my attention: Artificial intelligence. It was really interesting to learn all the potential applications that this technology has in the medical field as well as the ethical concerns that its introduction entails. I also found really motivating that Philips had decided to implement the RRI roadmap designed in this thesis in its "best practices". It is comforting to know that the effort spent in the last 6 months will not end up standing on a shelf.

However, there were also important difficulties during the process. It was particularly challenging to balance the practical outcome that the company wanted with the research outcome expected by TU Delft. There were many discussions with my company supervisor to shape the research in such a way that it also represented a scientific work. In addition, ideating the thesis proposal from zero to defining a path of research was also very difficult. Initially, I had many ideas in mind and it was hard to establish a structured research process.

If I had the opportunity to do this research again, I would change many things. First, I would focus more on the scientific side than on the practical side of the research. Second, I would interview stakeholders non-related to Philips to mitigate the selection bias and to obtain different points of view. It would have been quite interesting to interview patients and

clinicians that do not work for Philips. Besides, I would have liked to interview policymakers in the field of AI to understand better the issues taken into account to develop regulations. Additionally, I would have organized a workshop with several stakeholders to discuss about the ethical concerns of AI. It would have been really insightful to see how the different stakeholders interact. Finally, if time was not a constraint, I would have liked to work on the phases of experimentation and validation of the PRISMA methodology. Maybe, a lot more issues would appear that I did not consider when designing the first version of the roadmap.

Bibliography

Academy of Medical Royal Colleges. (2019). *Artificial Intelligence in Healthcare*. Retrieved from https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf

Accenture. (2017). *Artificial Intelligence: Healthcare's New Nervous System*. Retrieved from https://www.accenture.com/us-en/insight-artificial-intelligence-healthcare

AMA, American Medical Association. (2019). AMA Principles of Medical Ethics. Retrieved from https://www.ama-assn.org/about/publications-newsletters/ama-principles-medical-ethics

Anderson, M., & Leigh, S. (2019). How should AI be developed,validated and implemented in patient care? *AMA Journal of Ethics*, *21*(2), 125–130. https://doi.org/10.1001/amajethics.2019.125

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016, May 23). Machine Bias. *ProPublica.* Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Auer, A., & Jarmai, K. (2017). Implementing responsible research and innovation practices in SMEs: Insights into drivers and barriers from the Austrian medical device sector. *Sustainability (Switzerland)*. https://doi.org/10.3390/su10010017

Barben, D., Fisher, E., Selin, C., & Guston, D. H. (2008). Anticipatory Governance of nanotechnology: Foresight engagement and integration. In *The Handbook of Science and Technology Studies*.

Bartoletti, I. (2019). AI in healthcare: Ethical and privacy challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-21642-9_2

Batayeh, B. G., Artzberger, G. H., & Williams, L. D. A. (2018). Socially responsible innovation in health care: Cycles of actualization. *Technology in Society*. https://doi.org/10.1016/j.techsoc.2017.11.002

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology*.

Baumann, M. F., Brändle, C., Coenen, C., & Zimmer-Merkle, S. (2019). Taking responsibility: A responsible research and innovation (RRI) perspective on insurance issues of semi-autonomous driving. *Transportation Research Part A: Policy and Practice, Volume 124, 557-572.* https://doi.org/10.1016/j.tra.2018.05.004

Becker, A. (2019). Artificial intelligence in medicine: What is it doing for us today? *Health Policy and Technology*, *8*(2). https://doi.org/10.1016/j.hlpt.2019.03.004

Bhatnagar, S., Cotton, T., Brundage, M., Avin, S., Clark, J., Toner, H., … Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv Preprint ArXiv:1802.07228*.

Bird, S. Barocas, S. Crawford, K. Diaz, F. & Wallach, H. (2016) Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. *Workshop on Fairness,*

Accountability, and Transparency in Machine Learning, 2016. Available at SSRN: https://ssrn.com/abstract=2846909

Bostrom, N., & Yudkowsky, E. (2011). The Ethics of Artificial Intelligence. *Medicine*. https://doi.org/10.1016/j.mpmed.2014.11.012

Bresnick, J. (2018, September 17). Arguing the Pros and Cons of Artificial Intelligence in Healthcare. *Health IT Analytics.* Retrieved from https://healthitanalytics.com/news/ arguing-the-pros-and-cons-of-artificial-intelligence-in-healthcare

Brouwer, H., Hiemstra, W., van der Vugt, S., & Walters, H. (2013). Analysing stakeholder power dynamics in multi-stakeholder processes : insights of practice from Africa and Asia. *Knowledge Management for Development Journal*.

Busch. (2011). Standards. Recipes for Reality. MIT Press, Cambridge, MA.

Callon, M, Lascoumes, P. & Barthe, Y. (2019). Acting in an Uncertain World: An Essay n Technical Democracy. MIT Press. Cambridge. MA.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2783258.2788613

Chockley, K., & Emanuel, E. (2016). The End of Radiology? Three Threats to the Future Practice of Radiology. *Journal of the American College of Radiology*. https://doi.org/10.1016/j.jacr.2016.07.010

Collingridge, D. (1980). *The social control of technology.* Pinter London.

Crawford, K. (2016, June 25). Artificial Intelligence's White Guy Problem. *The New York Times.* Retrieved from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Denzin, Norman K. (1973). The research act: A theoretical introduction to sociological methods. New Jersey: Transaction Publishers.

Dickson, B. (2018, November 1). IBM, Harvard develop tool to tackle black box problem in AI translation. *VentureBeat.* Retrieved from https://venturebeat.com/2018/11/01/ibm-harvard-develop-tool-to-tackle-black-box-problem-in-ai-translation/

Dignum, V. & Bieger, J. (2019). *Artificial Intelligence and Ethics at the Police*. Retrieved from http://designforvalues.tudelft.nl/2019/ai-and-ethics-at-the-dutch-police/

Doorn, N. (2014). Assessing the future impact of medical devices: Between technology and application. In: van den Hoven, J., Doorn, N., Swierstra, T., Koops, BJ. & Romijn, H. (Eds) *Responsible Innovation 1: Innovative Solutions for Global Issues*. https://doi.org/10.1007/978-94-017-8956-1_17

Doorn, N., & Nihlén Fahlquist, J. (2010). Responsibility in Engineering: Toward a New Role for

Engineering Ethicists. *Bulletin of Science, Technology & Society*. https://doi.org/10.1177/0270467610372112

Dreyer, M., Chefneux, L., Goldberg, A., von Heimburg, J., Patrignani, N., Schofield, M., & Shilling, C. (2017). Responsible innovation: A complementary view from industry with proposals for bridging different perspectives. *Sustainability (Switzerland)*. https://doi.org/10.3390/su9101719

e-Health Initiative. (2018). *Artificial Intelligence in Healthcare*. Retrieved from https://www.proclinical.com/blogs/2019-5/the-top-10-medical-device-companies-2019

Ellis, M. (2019). *The top 10 medical device companies*. Retrieved from https://www.ehidc.org/resources/artificial-intelligence-healthcare

EPSRC, Engineering and Physical Sciences Research Council. (2019). *Artificial Intelligence Technologies*. Retrieved from https://epsrc.ukri.org/research/ourportfolio/researchareas/ait/

European Comission. (2011). A Renewed EU Strategy 2011-14 for Corporate Social Responsibility. *European Comission*: Brussels, Belgium.

European Comission. (2018). Ethics Guidelines for Trustworthy AI. *European Comission*: Brussels, Belgium.

European Group on Ethics and New Technologies. (2018). *Artificial Intelligence, Robotics and 'Autonomous' Systems*. Retrieved from https://ec.europa.eu/research/ege/pdf/ege_ ai_statement _2018.pdf

Fay, J. (2019, June 18). IBM's answer to AI bias? Don't leave spotting it to humans alone… *DevClass.* Retrieved from https://devclass.com/2019/06/18/ibms-answer-to-ai-bias-dont-leave-spotting-it-to-humans-alone/

FDA, (2019). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - *Discussion Paper and Request for Feedback. Food and Drug Administration:* White Oak, Maryland, USA.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*. https://doi.org/10.1007/s11023-018-9482-5

Forrest, A. (2019, July 12). Facebook data scandal: Social network fined $5bn over 'inappropriate' sharing of users' personal information. *The Independent.* Retrieved from https://www.independent.co.uk/news/world/americas/facebook-data-privacy-scandal-settlement-cambridge-analytica-court-a9003106.html

Frakt, A. (2016, July 11). Using the Web or an App Instead of Seeing a Doctor? Caution is Advised. *The New York Times.* Retrieved from https://www.nytimes.com/2016/07/12/upshot/using-the-web-or-an-app-before-seeing-a-doctor-caution-is-advised.html

Gartner. (2019). Gartner Hype Cycle. Retrieved from https://www.gartner.com/en/research/methodologies/gartner-hype-cycle

General Data Protection Regulation (GDPR). (2018). *General Data Protection Regulation (GDPR) – Final text neatly arranged*. Retrieved from https://gdpr-info.eu/

Gillespie, S. (2018, October 15). The Oxford spinout company using AI to diagnose heart disease. *University of Oxford*. Retrieved from https://www.research.ox.ac.uk/Article/ 2018-10-15-the-oxford-spinout-company-using-ai-to-diagnose-heart-disease

Global Market Insights. (2019). *Healthcare Artificial Intelligence Market Size Report*. Retrieved from https://www.gminsights.com/toc/detail/healthcare-artificial-intelligence-market

Greton, C. (2017, June 24). The dangers of AI in heath care: risk homeostasis and automation bias. *Towards Data Science*. Retrieved from https://towardsdatascience.com/the-dangers-of-ai-in-health-care-risk-homeostasis-and-automation-bias-148477a9080f

Grove-White, R., Macnaghten, P. & Wynne, B. (2000). Wising Up: The Public and New Technologies. *Centre for the Study of Environmental Change*. Lancaster, UK.

Groves, C., Frater, L., Lee, R., & Stokes, E. (2011). Is There Room at the Bottom for CSR? Corporate Social Responsibility and Nanotechnology in the UK. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-010-0731-7

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism: Clinical and Experimental*, *69*, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

Hamid, S. (2016). The Opportunities and Risks of Artificial Intelligence in Medicine and Healthcare. *The Babraham Institute, University of Cambridge*, (Summer 2016), 1–4.

Hart, R. (2017a, May 23). When artificial intelligence botches your medical diagnosis, who's to blame? *Quartz*. Retrieved from https://qz.com/989137/when-a-robot-ai-doctor-misdiagnoses-you-whos-to-blame/

Hart, R. (2017b, July 10). If you're not a white male, artificial intelligence's use in healthcare could be dangerous. *Quartz*. Retrieved from https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/

Hoffmann-Riem, H., & Wynne, B. (2002). In risk assessment, one has to admit ignorance. *Nature*. https://doi.org/10.1038/416123a

Iatridis, K., & Schroeder, D. (2016). *Responsible Research and Innovation in Industry: The Case for Corporate Responsibility Tools*. https://doi.org/10.1007/978-3-319-21693-5

Johnson, J. A. (2006). Technology and Pragmatism: From Value Neutrality to Value Criticality. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2154654

Kerr, A., Hill, R. L., & Till, C. (2018). The limits of responsible innovation: Exploring care, vulnerability and precision medicine. *Technology in Society*. https://doi.org/10.1016/j.techsoc.2017.03.004

Kirsner, S. (2018). The Biggest Obstacles to Innovation in Large Companies. *Harvard Business Review*.

Löfwander, S. (2017). About Artificial Intelligence, Neural Networks & Deep Learning. *Ayima.* Retrieved from https://www.ayima.com/blog/artificial-intelligence-neural-networks-deep-learning.html

Macnaghten, P., & Chilvers, J. (2014). The future of science governance: Publics, policies, practices. *Environment and Planning C: Government and Policy*. https://doi.org/10.1068/c1245j

Macnish, K. (2012). Unblinking eyes: The ethics of automating surveillance. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-012-9291-0

Marr, B. (2019). Deep Learning vs Neural Networks – What's the Difference. *Bernard Marr & Co.* Retrieved from https://bernardmarr.com/default.asp?contentID=1789

Marshall, A. & Davies, A. (2018, May 24). Uber's self-driving car saw the woman it killed, report says. *Wired.* Retrieved from https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/

Martinuzzi, A., Blok, V., Brem, A., Stahl, B., & Schönherr, N. (2018). Responsible Research and Innovation in industry-challenges, insights and perspectives. *Sustainability (Switzerland)*. https://doi.org/10.3390/su10030702

Merriam, Sharan B. (2009). *Qualitative research: A guide to design and implementation*. 2nd ed. San Francisco, CA: Jossey-Bass.

McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*. https://doi.org/10.1145/1045339.1045340

Microsoft. (2018). *Healthcare, Artificial Intelligence, Data and Analytics.* Retrieved from https://www.digitaleurope.org/wp/wp-content/uploads/2019/02/Healthcare-AI-Data-Ethics-2030-vision.pdf

Minsky, C. (2018, December 14). One former Google exec says there's no hope for Europe's artificial intelligence sector. *Sifted.* Retrieved from https://sifted.eu/articles/interview-google-kaifu-lee-ai-artificial-intelligence/

Miskinis, C. (2018, November 15). *Improving healthcare using meducal digital twin technology.* Retrieved from https://www.challenge.org/insights/digital-twin-in-healthcare/

Mittelstadt, B. (2017). 'The doctor will not see you now': The algorithmic displacement of virtuous medicine. In: Otto, P. & Graf, E. (Eds) *3TH1CS - The Reinvention of Ethics in the Digital Age*. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3298923

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.

Nielsen, J. (1993). Usability Engineering. *Morgan Kaufmann Publishers Inc*.

Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., & Abedi, V. (2019). Artificial Intelligence Transforms the Future of Healthcare. *American Journal of Medicine*. https://doi.org/10.1016/j.amjmed.2019.01.017

Nowatzke, B. (2019, August 12). The Novartis Data Scandal explained. *Pearl Pathways.* Retrieved from https://www.pearlpathways.com/the-novartis-data-scandal-explained/

Nuffield Council of Bioethics. (2018). *AI in healthcare and research*. Retrieved from http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future-big data, machine learning, and clinical medicine. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMp1606181

OECD. (2017). Enhancing the contributions of SMEs in a global and digitalised economy. *Meeting of*

the OECD Council at Ministerial Level. Paris 7-8 June, 2017. Retrieved from https://www.oecd.org/mcm/documents/C-MIN-2017-8-EN.pdf

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy*. https://doi.org/10.1093/scipol/scs093

Pacifico Silva, H., Lehoux, P., Miller, F. A., & Denis, J. L. (2018). Introducing responsible innovation in health: a policy-oriented framework. *Health Research Policy and Systems*. https://doi.org/10.1186/s12961-018-0362-5

Parks, J. A. (2010). Lifting the Burden of Women's Care Work: Should Robots Replace the "Human Touch"? *Hypatia*. https://doi.org/10.1111/j.1527-2001.2009.01086.x

Perrin, N. & Mikhailov, D. (2017, November 3). Why we can't leave AI in the hands of Big Tech. *The Guardian.* Retrieved from https://www.theguardian.com/science/2017/nov/03/why-we-cant-leave-ai-in-the-hands-of-big-tech

Philips. (2018). *Using AI to meet Operational, Clinical Goals.* Retrieved from https://www.philips.co.in/healthcare/nobounds/four-applications-of-ai-in-healthcare

Philips. (2019a). *More than a century of innovation and entrepreneurship.* Retrieved from https://www.philips.com/a-w/about/company/our-heritage.html

Philips. (2019b). *How we create value for our stakeholders.* Retrieved from https://www.philips.com/a-w/about/company/our-strategy/how-we-create-value.html

Philips. (2019c). *Philips Annual Report 2018.* Retrieved from https://www.philips.com/c-dam/corporate/about-philips/sustainability/downloads/other/philips-full-annual-report-2018.pdf

Philips. (2019d). *Philips Talks: Winning with Artificial Intelligence – in all we do* [video file]. Retrieved from https://www.streaming.philips.com/talks/ (access limited to employees)

Philips. (2019e). *Adaptive Intelligence. The case of focusing AI in healthcare on people not technology*. Retrieved from https://www.philips.com/c-dam/corporate/newscenter/global/standard/resources/healthcare/2018/ai-in-healthcare/artificial_intelligence_position_paper.pdf

Philips. (2019f). *Philips at IFA 2019.* Retrieved from https://www.philips.com/c-e/ifa.html

Philips. (2019g). *Philips and Microsoft HoloLens 2: could augmented reality change the face of image guided therapy?*. Retrieved from https://www.philips.com/a-w/about/news/archive/standard/news/articles/2019/20190313-philips-and-microsoft-hololens-2-could-augmented-reality-change-the-face-of-image-guided-therapy.html

Platania, R., Zhang, J., Shams, S., Lee, K., Yang, S., & Park, S. J. (2017). Automated breast cancer diagnosis using deep learning and region of interest detection (BC-DROID). *ACM-BCB 2017 - Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. https://doi.org/10.1145/3107411.3107484

PRISMA Project. (2019). PRISMA RRI Exemplar Roadmap. Retrieved from https://prod-edxapp.edx-cdn.org/assets/courseware/v1/12af5c3d47db3ec4fd48a57fac6f3e1e/asset-v1:DelftX+RI102x+2T2019+type@asset+block/PRISMA_RRI_Exemplar_Roadmap_June__2019.pdf

Ranschaert, E. R., Duerinckx, A. J., Algra, P., Kotter, E., Kortman, H., & Morozov, S. (2019).

Advantages, Challenges, and Risks of Artificial Intelligence for Radiologists. In *Artificial Intelligence in Medical Imaging*. https://doi.org/10.1007/978-3-319-94878-2_20

Ravetz, J. R. (1997). The science of "what-if?" *Futures*.

Ross, C. (2018, July 25). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *STAT*. Retrieved from https://www.statnews.com/2018/07/25/ibm-watson- recommended-unsafe-incorrect-treatments/

Santoni de Sio, F. & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI*. https://doi.org/10.3389/frobt.2018.00015

Schutt, R. K., & Chambliss, D. F. (2013). Chapter 10: Qualitative Data Analysis. In *Making Sense of the Social World: Methods of Investigation*. https://doi.org/10.1136/ebnurs.2011.100352

Sekaran, U. & Bougie, R. (2013). *Research methods for business. A skill building approac*h. 5th ed. John Willey: UK

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-010-9234-6

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human Computer Studies*. https://doi.org/10.1006/ijhc.1999.0349

Smith, J. (2015). Qualitative Psychology: A Practical Guide to Research Methods. 2nd ed. SAGE Publications.

Sonck, M., Asveld, L., Landerweerd, L. & Osseweijer, P. (2017). Creative tensions: mutual responsiveness adapted to private sector research and development. *Life Sciences, Society and Policy*. 13:14.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*. https://doi.org/10.1016/j.robot.2016.08.018

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

te Kulve, H., & Rip, A. (2011). Constructing Productive Engagement: Pre-engagement Tools for Emerging Technologies. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-011-9304-0

Timmermans, J. (2017). Mapping the RRI landscape: An overview of organisations, projects, persons, areas and topics. In *Responsible Innovation 3: A European Agenda?* https://doi.org/10.1007/978-3-319-64834-7_3

TU Delft. (2019). *MSc Management of Technology*. Retrieved from https://www.tudelft.nl/onderwijs/opleidingen/masters/mot/msc-management-of-technology/

TU Delft MOOC. (2019). Module 2: Foundations for RRI. *In DelftX RI102x: Responsible Innovation – Building tomorrow's responsible firms*. Retrieved from https://courses.edx.org/courses

/course-v1:DelftX+RI102x+2T2019/course/

van de Poel, I., Asveld, L., Flipse, S., Klaassen, P., Scholten, V., & Yaghmaei, E. (2017). Company strategies for responsible research and innovation (RRI): A conceptual model. *Sustainability (Switzerland)*. https://doi.org/10.3390/su9112045

van de Poel. I. & I., Royakkers, L. (2011). Ethics, technology and engineering. Oxford: *Wiley-Blackwell.*

Van Wynsberghe, A. (2015). Healthcare Robots: Ethics, design and implementation. 1st ed. Routledge.

Verschuren, P., & Doorewaard, H. (2010). Designing a Research Project. 2nd ed. The Hague: Eleven International Publishing

von Schomberg, R. (2012). Prospects for technology assessment in a framework of responsible research and innovation. In D. Marc. & B. Richard (Eds.), *Technikfolgen abschätzen lehren* (pp. 1-19). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93468-6_2

Wynne, B. (2011). Lab Work Goes Social, and Vice Versa: Strategising Public Engagement Processes. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-011-9316-9

Yaghmaei, E. (2016). Addressing responsible research and innovation to industry - Introduction of a Conceptual Framework. *ACM SIGCAS Computers and Society*. https://doi.org/10.1145/2874239.2874282

Appendix 1 – Interview Protocol

Type of interview: Semi-structured interview
Duration: 60 minutes

Objectives:

- ✓ To understand what are the main drivers and challenges for the development and introduction of AI technologies in healthcare.
- ✓ To understand how are the practices of Responsible Research and Innovation (RRI) currently implemented in Philips for the development of AI products.
- ✓ To identify potential improvement opportunities as well as challenges for the deployment of RRI strategies.

Outline of the interview:

1. Introduction (5 mins)
2. Drivers and challenges for AI implementation (15 mins)
3. Ethical concerns of AI in healthcare (20 mins)
4. RRI and its current adoption by Philips (15 mins)
5. Closing (5 mins)

The interviews follow a semi-structured approach. The questionnaire is the guideline during the interview. However, some questions or topics can change depending on the progress of the interview and the job position of the interviewee.

Introduction:

I will start with an introduction of myself and a brief explanation of the project. Then, I will proceed with the following questions:

- What is your current role within Philips?
- What is your background (previous jobs, industries)?
- How is your position related to the field of artificial intelligence?

Drivers and challenges for AI implementation:

- What are the main drivers to develop AI technologies for healthcare?
- How can Philips take advantage of its current capabilities to become a leader in AI healthcare solutions?
- How will people benefit from AI healthcare solutions?
- What are the main technical challenges for AI development?
- What are the most important regulations that have to be complied to introduce AI in healthcare (GDPR, HIPAA, etc)? What are the main regulatory challenges for AI introduction in healthcare?
- What are the main ethical challenges for AI introduction in healthcare?

Ethical concerns of AI in healthcare:

- Is the top management of the company committed to address ethical issues regarding AI?
- Are there any corporate social responsibility practices already implemented by Philips for AI development?
- Do the AI products developed by Philips enhance or replace human capabilities?
- How do Philips ensure the safety of patient data? Are there any risk assessments for data transfer?
- How do Philips validates the quality of the data used to trains its AI algorithms?
- How do Philips ensure that the personal biases of software designers are not embedded in the algorithms?
- Automation bias is the propensity for humans to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct. Has Philips already implemented any characteristics in its AI decision-making systems to avoid this bias (e.g. percentage of accuracy)?
- Are there any measures already taken to ensure the transparency of the algorithms? Is there any procedure in place to explain until some extent how the algorithm reached a specific outcome?
- Why should doctors and patients believe in the AI systems developed by Philips if there is no way to explain how do they arrive at an outcome?
- How do Philips ensure that telehealth applications do not contribute to the social isolation of elderly patients?
- How does Philips ensure that its AI products will not only benefit the wealthiest populations, stretching the inequality in healthcare access? (Knowing that Philips is a premium brand)
- Do you think that internal ethical policies for AI product design stifle innovation?
- Do you think that taking into account ethical issues before product development increases the social acceptability of AI products?

RRI and its current adoption by Philips:

- Are you familiar with the concept of Responsible Research and Innovation?

1. Anticipation
    - Are ethical, legal and social concerns discussed since the initial phases of technology development?
    - Is there a privacy officer involved since the early stages of technology development?
    - Are the potential impacts, positive or negative, of the new product evaluated early in the design phase?
    - Are the potential risks and different scenarios assessed at the beginning of AI product design?

2. Reflexivity
    - Are ethical audits a common process inside Philips? Are people encouraged to think about the social impacts of their work?
    - Is there constant training and education in CSR practices in the AI R&D teams?
    - Do employees follow a code of ethics for AI systems design?
    - Do employees follow a code of ethics for personal data management?

3. Inclusiveness
    - Who are the most important stakeholders that Philips takes into account to develop and introduce AI solutions (academy, government, regulatory institutions, patients, doctors, public, suppliers, etc.)?
    - How is the decision-making power distributed among stakeholders?
    - How are end-users (patients) involved in the innovation process? How are doctors involved? How is the general public involved?
    - How does Philips embed the values of the different stakeholders in AI product design?
    - In the case of AI, does Philips adopt an open-innovation strategy or a closed-innovation strategy?
    - Is there any collaboration with policymakers and authorities during the design process of AI applications?

4. Responsiveness
    - Currently, physicians are held responsible for the outcomes of treatment decisions taken with the help of AI. However, this may change in the future and technology companies behind AI algorithms might be held responsible as well. Does Philips already has a potential strategy to deal with this issue?
    - How do Philips respond to the feedback from the medical community on its AI products?
    - Have Philips AI products experienced opposition from the medical community? How has Philips dealt with that?
    - How flexible are the AI research and innovation teams to reshape the design of a product?

Closing:

- Do you have additional suggestions or comments that can contribute to this research (colleagues to contact, company documents to consult or articles to consider)?
- Are you available for further questions in the future?
- Do you want to be informed of the outcomes of this research?

Appendix 2 – Coding Tables

A2.1 Coding of drivers

Interviewee	Empirical data	First-order coding	Second-order coding
G	"Healthcare is quite expensive; it is a substantial part of GDP in different countries. Up to 18% in Western societies. AI could help us to develop more affordable care"	Offer more affordable care	Economic drivers
G	"The objective is the reduction of healthcare cost"		
H	"In healthcare, the main driver is basically cost. If the cost is downsized for hospitals, it will be translated to lesser cost for patients"		
I	"We expect to see a big increase in our revenue once the AI technologies that we are developing are fully embedded in the market"	Increase revenue	
B	"Another important driver is that Philips wants to be the leader in the digital transformation of healthcare"	Be the leader in the digital transformation of healthcare	Organizational drivers
I	"In Philips we aim at disrupting the health industry, by proving digital solutions that are better and more reliable"		
D	"Philips is really good at processing data and at combining it with the contextual knowledge that we have from the clinical world. That is what makes Philips unique"	Take advantage of the contextual knowledge of Philips in the medical world	
E	"Philips wants to keep its position as one of the industry leaders on the healthcare technology market. So, we have to work on innovative technologies, such as AI, that help us gaining advantage over our competitors"	Keep the position as one of the industry leaders in medical technology	
F	"The main driver of AI in healthcare is to be able to analyze the immense amount of health data that we collect every day, to reach more people and offer better care"	Reach 3 billion people a year by 2030	Social drivers
I	"The vision of Philips is to improve the lives of 3 billion people a year by 2030. That is where AI plays a role, because with AI technologies analyzing millions of patient data in seconds, you can reach more people than ever"		
B	"We try to bring the value of different technologies to improve health. It comes from the vision of the company to reach 3 billion people before 2030"		
A	"Overall, I would say that the main driver to develop AI in healthcare is to make the world a better place by adding value to global health"	Make the world a better place	

C	"AI will be a fantastic tool for the physicians to take back the old habit of taking care of patients"	Improve the experience of patients and HCPs	Technical drivers
G	"Some drivers are improving the patient's experience, the healthcare employees' experience and the healthcare outcomes"		
J	"AI can reduce the high load of routine work in the hospital"		
J	"AI can be used to diminish the amount of false alarms in clinical practice"		
G	"There is always the ongoing need to continue technological innovation and people always want to explore territories where nobody has gone before"	Human need for technological innovation	
F	"There is challenge in balancing the individuals' right to privacy versus society's need to move ahead, even if that means that you are aggregating data from individuals that may have a risk of being identified"		
A	"Within Philips there is a driver to understand what patient data is telling us in order to make more efficient diagnosis"	Improve technological efficiency and accuracy	
B	"The main driver is to achieve more accurate detection, and faster and better solutions"		
C	"The main driver is to make sure that all the information that is necessary for the patient to get a proper treatment at a proper cost is available to the one that needs to act on it"		
D	"Introducing AI in cases that extend beyond the human comprehension is the way to go. I think it would help to improve healthcare in a way that it becomes more efficient"		
E	"In Philips, we are trying to improve the MRI machines' scan time. Usually, it takes around 20 minutes to generate an image. We can bring it down to 5 minutes"		
H	"Another driver is repeatability. You can reproduce things that are done by machines instead of humans"	Improve repeatability of healthcare outcomes	
J	"AI helps to combine vital signs such as heart rate, respiration rate, oxygen saturation, etc. to detect patient deterioration earlier"	Improve the quality of healthcare outcomes	
J	"The symptoms of a particular patient can be compared to the ones of large populations to aid in the diagnosis and treatment of diseases"		

Table 13. Coding of drivers for AI development

A2.2 Coding of non-ethical challenges

Interviewee	Empirical data	First-order coding	Second-order coding
E	"Philips is quite closed when it comes to research. Publishing internally is much more appreciated than publishing externally in conferences and so on"	Closed-innovation approach in some teams within Philips research	Organizational challenges
I	"We recently developed data principles. However, making the change from principles to practices is still missing"	Going further from data principles to data practices	
A	"In a big company, it is always hard to change direction. There are many people working on staff and to coordinate them is not an easy task"	Lack of flexibility	
B	"In general, in R&D things are not very flexible. Sometimes you spend a lot effort for some strategic changes. When the situation changes and you can continue the people that was involved is already gone"		
C	"I don't think that Philips has a good record in being quick, fast and flexible under different situations. We usually take time to develop our processes and once a process is decided, it is almost written in stone."		
C	"Philips is not the flexible one that quickly walk with the changing environment"		
D	"Being a big company gives you the opportunity to create a meaningful impact at a large scale. However, you need to give up a lot of flexibility"		
D	"There is still no regulation about the level of accuracy in which your algorithm can change from saying "Look at this, please define what it is" to "That's a tumor". I hope that accuracy to be 99.99%"	Lack of clear regulation (accuracy)	Regulatory challenges
A	"One challenge is probably legislation. Legislation differs quite a lot and specifically in access to data"	Lack of clear regulation (data access)	
A	"Who owns the data? Is it the patient? Is it the hospital? Is it the physician? Where does that go? That discussion is not fully clear"	Lack of clear regulation (data ownership)	
C	"Regulation should be developed to establish clearly who owns the data. Also regarding how to compensate patients for the data. We don't want to see patients selling themselves for care"		
A	"There is still not regulation for dynamic algorithms in which the system keeps learning in the field when new data is fed"	Lack of clear regulation (dynamic algorithms)	
F	"We also struggle within Europe, because we have 28 different interpretations of the GDPR, specifically because healthcare is member state. So, we can deal with GDPR but dealing with different flavors in different countries is making it more difficult than necessary"	Lack of clear regulation (GDPR interpretation)	

G	"GDPR has just been implemented since last year. GDPR as such is not really the issue. The interpretation of the GDPR is the issue. The different nations within the EU are dealing with that in a different way"		
J	"The interpretation of GDPR is not clear yet. The legislation still needs to be developed further"		
A	"Health data is not that easily accessible and European Legislation around privacy is very strict on how data can be accessed or how data can enter the EU"	Regulation is very strict and perceived as too stringent (External)	
A	"Regulation for the sake of regulation will stifle innovation"		
B	"To be honest, for patient data is really difficult to get a massive dataset due to the privacy regulations"		
B	"If you want to study or collect some data, you need a lot of effort to get the approval and consent of patients and hospitals. It brings a lot of barriers."		
C	"We tried to acquire a database from the USA to use it in Europe. For that, we needed to involve a third party to do a privacy assessment according to the GDPR. It took so much time that we just decided to do the tests in the USA"		
D	"We didn't make it easier for ourselves with the GDPR, but we did it right. I think we need it, but it doesn't make it easier to develop something"		
D	"Regulators take an extra level of safety in what they produce, because they simply do not understand some of the aspects of the technology"		
J	"The Netherlands is really stringent when it comes to interpreting the GDPR"		
B	"Encouraging the GDPR in the company has two sides. On the one side it is very good because we are aware of the ethical and privacy problems but, on the other hand, it slows down the process of innovation"	Regulation is very strict and perceived as too stringent (Internal)	
C	"In Philips, we take the lowest possible risk and that's where the restrictions come into place. Not because of GDPR but because of our interpretation of GDPR"		
E	"I think that Philips is too much on the safe side when interpreting the GDPR. We could be more aggressive because GDPR is a guideline, then it is up to you how to implement it"		
E	"Internal data policies in Philips have a big impact. Sometimes, I just want to test an idea with regular data, not patient data or anything like that. Then, I need to make a project proposal to make an experiment that will take one afternoon. I am not going to work on a project proposal for a week for something that I can test easily in one afternoon. Then, I prefer not to do the experiment"		
H	"Despite being a global company, Philips adopts a western perspective in AI and is very protective regarding data. It makes the process a lot longer"		

J	"The ICBE is very strict with GDPR and it is hampering scientist more than it needs to. It has a big impact on innovation"		
H	"In healthcare, AI solutions have to be much better than the current solutions (even if they are cheaper) to overcome the privacy concerns"	AI solutions have to be much better to be accepted	Technical challenges
A	"Privacy breaches cause anger and disappointment, we should avoid them at all costs by implementing strong cybersecurity practices"	Privacy and security threats (Strong cybersecurity solutions)	
B	"There are many risks regarding privacy and security. The consequences may be huge"		
B	"In healthcare, especially when you develop decision-support systems for clinical decisions, you really need to understand what's going on in your AI model"	Difficulty in explaining the 'black-box'	
B	"Making AI more explainable and transparent is a huge challenge"		
B	"For me, it is very difficult to understand how DL systems come to a decision, how the model exactly works. Unless you go to every neuron of the neural networks, which is massive"		
I	"Opening the 'black box' is really difficult because in deep learning the machine learns itself and re-adapts"		
E	"There are techniques being developed to make the "black-box" a "grey-box". However, we won't be able to get the full "white-box"		
H	"Integration is a big problem. Medical devices are connected to an entire ecosystem. If you have a digital add-on, you are constrained to what already exists. It has to fit in. You cannot have a doctor with two computer systems"	Difficulty to integrate new AI systems in the existing infrastructure	
J	"IT integration is a huge challenge, because there are hundreds of different systems in which AI has to fit in"		

Table 14. Coding of non-ethical challenges for AI development

A2.3 Coding of ethical challenges

Interviewee	Empirical data	First-order coding	Second-order coding
E	"The company has to be careful in the kind of applications that are pursued, because you get more responsibility. If the company wants to have more responsibility because there is a business opportunity, then they have to be held accountable. It is a choice but if you get the benefits you should also get responsibility"	Higher responsibility for healthcare technology companies	Accountability
J	"There is a lot of discussion about who should be responsible for a decision. If the AI is wrong, whom should we blame? The scientist, the technology company, the insurance company, the hospital, the doctor?"	Lack of clarity on who should be held responsible for a decision	
A	"When people do not see the added-value of ethically responsible practices, there is no much we can do"	Unawareness of ethical impacts (customers)	
D	"I take a plane and I don't really understand how the plane works, but I trust it to stay in the air. So, why do I need to know the exact content into the ""black-box". If a regulation appears to prevent a bias, it is founded on the fear of bias"	Unawareness of ethical impacts (doctors)	
D	"As a doctor I don't feel like knowing how AI is working, I am only interested in the outcomes. It has nothing to do with the interaction doctor-patient. So, we don't bother, a different institution can look at it, not the end-user"		
A	"Normally, researchers are focused on their projects and not much on the impacts of them. Reflection is often a luxury and not so much an internal part of the process"	Unawareness of ethical impacts (scientists)	
B	"Sometimes, effects on society are not clear. In that case, you have to continue with the project and see what happens in the market"		
D	"Many scientists do not like the confrontation with the boards that ensure compliance of ethical, legal and privacy aspects"		
I	"If I speak from a data science perspective, the moral part is a hindrance. In principle, if you have free reign and a lot of data you will do a lot more and faster"		
A	"Developers must not consciously introduce biases. However, it is very hard to control this"	Bias embedded by the software developers	Data bias, fairness and equity
F	"We try to mix the software development teams to avoid against potential biases but there is not fully proved way against having an employee that is deliberatively or inadvertently biased"		
H	"It is extremely hard to avoid personal biases from software developers"		
F	"It is hard to find as many computer scientists, female, form emerging countries in comparison to computer scientists, male, from the USA or Europe"		

A	"Perhaps, the most prominent ethical challenge will be bias, where datasets are collected only in certain contexts and then applied in a wider sense. So, the data is not representative of how the applications might be used"	Presence of biased data	
A	"Stereotyping and unfair discrimination also play a role. In Africa, for example, a lot of health data is not digitized. It makes it inaccessible for us. That means that data cannot be drawn into AI applications. By ignoring that data, we can actually discriminate against populations or regions"		
C	"Access to high-quality unbiased data is a big challenge"		
E	"If you don't have enough data, even if it is high quality, you have a mismatch with the reality"		
F	"If there is garbage data in or if the data labelling is incorrect you will learn mistakes instead of learning the right thing"		
H	"Even if the algorithm works 99% of the cases but the 1% is 100% of a group it is still a huge number"		
G	"The second big challenge is access to high-quality data. We need data that can represent the case 100%"		
D	"Ensuring data quality in healthcare is very difficult. I have my doubts whether you can get a dataset with the level of quality to train for the recognition of a disease like tuberculosis. This disease comes with all different kinds of shapes, sizes, colors and smells. It manifests in ways that you'd never expect"		
E	"It is really difficult to validate the quality of data if you don't have the quantity. You can only detect biases if you have enough data, you cannot do it if you don't have the data"		
D	"There was a high blood pressure medication developed specifically for Afro-American groups in the US. It worked really well in clinical trials, but people didn't want to be discriminated in this way. So, even if you have a bias that may be good for the patient, people might not accept it"	Reluctance to 'good' bias	
A	"Ensuring that AI products not only reach wealthy populations is a big challenge. The company is committed to reach 3 billion people by 2030, so that requires innovative thinking from everyone in the company"	Unequal distribution of care	
C	"Today, healthcare costs are rising so much and access to care is not fairly distributed"		
H	"The west has this idea that if you have a product to sell, you will try to sell it in the wealthiest markets"		
H	"The state of the art of AI is pretty much elitist. If you want to have a fancy facial image recognition software, you need to have a fancy phone that can handle it. If you have a medical device that is so small that you can wear it as a wearable (the smaller the device the more expensive it is), it would be something that poor people won't be able to pay for"		

F	"If the doctors are not willing to adopt changes in the way that they provide care, in the way that they structure their workday, in the way that they rely on cues from the system combined with the medical knowledge; if the education of the doctors doesn't change and they continue to be educated only based on paper, AI will take a very long time to be adopted"	Changes in the role and education of the healthcare professionals	Effects on healthcare professionals (HCPs)
G	"AI will be a similar type of revolution in which the doctor will need to learn to deal with AI, where AI will assist him and the doctor will get a different role"		
D	"There is a shortage of doctors. With AI you can have somebody with a bachelor performing certain type of surgery. You don't need a full training as a GP"		
H	"Automation bias is a tough challenge. You need to be careful on how you frame your product. If you suggest that you are outcome is almost 100% accurate, the doctor will act less critically"	Risk to fall in the "automation bias"	
J	"Automation bias is difficult to control. Sometimes people now that the value dos not make sense, but they just trust the value. With AI it will be even worst"		
J	"If the AI is right 98% of the time, it would be difficult to ensure a critical behavior from doctors"		
A	"AI applications can be seen as the super expert that replaces the expertise of the health professionals. I can imagine some health professionals seeing it as a risk that they are going to be replaced, but similar things have been seen in other technologies"	Threat of replacement of healthcare professionals	
D	"Some doctors are scared that they will be replaced by AI, but only because they don't know what AI is"		
D	"The role of the doctor is changing. Radiology will radically change over the next decades because looking at the picture and pointing out the tumor is not fundamental. Let us let AI do that"		
D	"If AI is presented as a technology with the potential of taking over the doctor, it will create fear and lack of understanding amongst practitioners"		
F	"In many areas, the medical community is extremely concerned that some tasks will be eliminated and some jobs will be lost"		
H	"I do not think that is the intention but I cannot say that doctors are not going to be replaced. If that happens, it wouldn't be intentional"		
I	"In the case of oral healthcare. If I find that a procedure that a dentist makes is very inefficient. They will feel threatened"		
B	"It shouldn't be compulsory for people to use certain technologies. Everyone should be able to choose. In telehealth, people should also have the option of going to the hospital and talk to a nurse or doctor"	Patient autonomy might be compromised	Effects on patients
C	"Sometimes we do not involve patients in innovation, because we are not marketing to patients"		

H	"Personally, I would imagine people feel less lonely just by seeing people even if you don't talk to them. You get that in a hospital, not with telehealth"	Social isolation (telehealth)	
G	"The first challenge is willingness of society to go to a path of predictive healthcare enabled by AI"	Willingness to change from reactive to predictive care	
C	"We will never have an agreement with the patient to sell us the data, and that is possibly not allowed because patients will be in a disadvantaged position. Maybe they will start selling themselves for care and that should never happen"	Unfair commercialization of patient data	Privacy and security
G	"You need to have access to data of individual patients and they may question whether there is responsible use of the data. If it is not being misused for other purposes"		
F	"If you learn from the average doctors you will be pulling the best care down to average level and that is not where we want to go. You cannot generalize from it"	AI has to be better than the best doctors	Reliability and Safety
D	"As a doctor, I'm allowed to make mistakes. If I give a proper explanation of my reasoning, the mistake might be admissible. But people will never accept it when it comes from a machine"		
F	"Humans tend to forgive human errors but humans do not tend to forgive machine errors. In healthcare, machines have to be much better than humans to be accepted and to replace even the simplest human tasks"		
D	"The biggest challenge is to understand what the bias is of the AI we develop. It is really difficult to implement contextual information in the algorithm"	Difficulty to introduce contextual information from the clinical world	
F	"If you make data analysis without making clinical sense out of it, you will not be able to create hypothesis. Therefore, you will not be able to create cause and effect analysis, and then you cannot create a solution"		
F	"The main problem is not access to data; the main problem is making sense out of the data, creating solutions out of it and making sure they are adopted"		
C	"If the data is not representative of the disease, then you will train the machine to do something wrong"	Wrong diagnoses	
C	"If you give the wrong advice because of errors in our system, we are liable and there is a patient on the table"		
E	"In decision-making applications, it would be difficult to have flexibility. In that case, you are pushing doctors to take decisions and you can induce biases"		
F	"AI has potentially more risk of malfunctioning than hardcore old-style software. Old solutions are more controlled because they are derived from a model and requirement specs"		
B	"Sometimes AI looks fancy but it doesn't really work. To keep the high expectations scientist resort to 'window dressing'"	"Window-dressing" by technology developers	Transparency and Trust
F	"One of the main problems of AI in healthcare is that the dynamics and expectations are going faster than what you can deliver in a reliable manner"		

A	"Patient autonomy has to be respected and the AI application must not become the final decision-maker. Sometimes, users may not even know that they are interacting with an AI application, which is not a transparent approach"	Lack of trust towards new technologies	
B	"Sometimes doctors are very conservative and do not trust new technologies, they want to be sure"		
F	"Over the history of AI, a huge part of the investment has been wasted by a few things going wrong and society thinking that AI is not developed enough yet"		
F	"I think that we are in an inflection point where if we start to see AI misbehaving and doing harm to people or patients, it will be discarded like the last times"		
B	"For instance, you give some values of risk to the doctor, but they don't understand what's going on, so when they make decisions they will be confused and it is also risky if the model makes mistakes"	Lack of understanding of the AI solution	
E	"The main issue is the control and explicability. The trustworthiness of the solution because you don't want to have a black-box kind of approach. You have to be able to explain why you made a decision"		
F	"If data is too easily accessible, there is a tendency to not understand the problem. People just get the data, throw a deep learning model at it, don't understand the solution and put it out there"		
H	"The main challenge is explicability. In healthcare, some algorithms are very advanced and it is very hard to put them in the market because people want to understand what causes the result they see"		
H	"Sometimes research is harmful, because you spend months building your algorithm and it gives garbage out, then you have to kill your own proposition"		

Table 15. Coding of ethical challenges for AI development

A2.4 Coding of RRI actions

Interviewee	Empirical data	First-order coding	Second-order coding
A	"During the review process at the ICBE, we analyze the long-term consequences of a proposed project. We check if it is beneficial for society"	Analyze the potential impacts of AI systems since the initial phases of development	Anticipation & Reflexivity
F	"A systematic analysis of what could happen if AI goes wrong is done by the privacy experts and the security experts"		
H	"Before we develop our algorithm, we have to go through a long process where we get in touch with the privacy and ethical committee"		
F	"What we are trying to do is to collect high quality and representative data, well curated and combined with medical data. Then, you can create hypothesis and validate them, then you can do good in a controlled manner. That is a steady but slower approach to the problem"	Design rigorous procedures to ensure that data has a high-quality, it is well-curated and representative of the population	
F	"We do not believe in learning from uncurated datasets because it induces sensitivity to human error"		
G	"Our data scientists are busy 90% of the time curating the data, checking if data is representative of the real situation"		
J	"Diversity in data has to be ensured and validated since the beginning of product design"		
A	"The company is working on an AI ethics code. It was implemented by high-level management"	Develop and implement data and AI principles (codes of conduct)	
F	"We are currently developing the data principles and AI principles. Data principles are more about privacy and security of data. AI principles are more about benefits, do not harm, be transparent, be fair in the sense of equitable access and don't discriminate"		
B	"More work can be done from Philips to educate the PHD students in the company to embrace ethical awareness. This is very important in the education of a scientist"	Educate PHDs within the company in the ethical aspects of research	
G	"AI products have to be clinically validated. Normally, we do an auto-validation before applying for EU validation"	Ensure that AI-enabled medical devices are clinically validated	
B	"Philips is really committed to helping the doctors, not replacing them"	Ensure that the vision of the company with AI is aimed at supporting the doctors, not replacing them	
C	"We will never take the decision on behalf of the doctor. They will always have the autonomy to take the final decision"		
D	"The advice power of the doctor shall remain in place. AI should never take over the doctor"		
D	"AI should never be framed as the technology with the potential to replace the doctors"		

F	"For the moment, we support and augment the clinicians, we do not aim to replace them"		
G	"We have taken the position that the machine does not make the decision. It's the doctor who make the decision"		
H	"Right now, the doctors have agency and they should act upon things"		
J	"Any AI solution should support clinicians in doing their work but it should not take over their work"		
D	"Some years ago Philips established the ICBE, which verifies the ethical use of running experiments"	Establish an ethical, social and legal monitoring board	
D	"There is a very early involvement of the ICBE in research. They proceed with an ethical, privacy, Q&R and medical efficacy check since the start of a project and during its different phases"		
G	"When we start innovation projects, we have the ICBE committee evaluating the research purpose and what it is required for that purpose. Whether it is ethically sound and respecting all the appropriate regulations"		
A	"The privacy legal department of Philips is very strict with data management. Patient data should be duly recognized and duly protected"	Establish clear and rigorous practices of data management regarding privacy and consent	
B	"Philips started developing data transfer tools two years ago and within Philips the awareness is very good. We are not allowed to transfer data to the outside world, so no USB or online data transfer platforms"		
B	"Before starting a project, you need to clearly explain to the ICBE what are you going to do with the data"		
C	"If you do a data study, you always need to get a privacy assessment"		
E	"When you start a project and you need patient data, you need to get the approval of the ethical board. The same applies for the hospital"		
E	"We have started this year a center of expertise on data science and AI, that will help people inside the company to do a proper validation of the data"		
G	"We have already a corporate rulebook on how to deal with data"		
J	"There is a secure data transfer procedure to send and receive information from hospitals"		
G	"We comply with the five ethical principles for the responsible application of AI developed by the European Commission: Beneficence and non-maleficence, transparency, no discrimination, explicability, and human oversight"	Every AI application developed has to comply with the ethical principles of AI established by the European Commission	

B	"Data is always anonymized and there is an expert from Philips evaluating the quality of anonymization"	Guarantee that patient data is anonymized	
H	"We have to make sure that data is anonymized"		
I	"Regarding AI, one of the departments has taken the decision to organize workshops were researchers, privacy officers, legal officers and ICBE sit around the table and look at the challenges of AI applications"	Organize workshops with internal stakeholders to reflect on the challenges of AI	
A	"We send constant communications to our employees about ethics and responsibility to make them aware of these issues"	Train employees in ethical and privacy principles	
A	"Researchers have to go to a formal privacy training session before they can send proposals to be reviewed"		
B	"We have regular newsletters from the legal and privacy office. Also, there are Philips University Courses, which are compulsory. These courses are about privacy issues, ethics and the ICBE"		
C	"We constantly have the Philips University Trainings about business principles, ethics, privacy, and that kind of discussions"		
E	"We have the ICBE to establish the CSR practices and checks. However, as a scientist I have my own awareness. The company must help to create this personal awareness in their employees"		
J	"Every employee has a mandatory ethical training every year"		
A	"Our privacy officers go to the legal department to understand the regulations, and how we interpret them"	Understand and implement the legislation early on the technology development process	
A	"We are in the process of defining clear pathways to access the data within the legislation. We want to help researchers to make use of those pathways"		
F	"In the CoE of Data Science and AI, we develop capability plans, helping the businesses to develop AI solutions and looking at legal and ethical compliance frameworks"		
F	"We comply with the GDPR and take legal requirements into account"		
G	"We have, of course, implemented the GDPR in our processes and procedures with all the compliance with the law"		
B	"We do cross-population validation to avoid biases"	Design cross-population studies to validate AI applications	Inclusiveness
E	"Academy is also an important stakeholder. I am actually leading external collaborations with universities"	Enhance the collaboration with academy for AI related purposes	
F	"We publish as much as possible in research, so that the scientific community can reproduce and find mistakes"		

A	"More actions have to be designed to ensure that software designers do not embed biases in the algorithms"	Ensure a high level of diversity in AI software development teams
F	"We try to balance our teams as much as possible. You have got multinational and multicultural teams"	
A	"Philips has established research labs in developed and developing countries. Products are being taken to many different countries. It improves equality"	Establish AI research centers in different income-level countries
F	"You need the agility of people who are excellent with the tools, to explore. But you also need us to say, ok this is not as good as the hardcore solution that we had before, so we are not going to use it"	Find a balance in the teams between innovativeness and experience
A	"Healthcare professionals are included constantly in focus groups to assess the quality of our research"	Include healthcare professionals and patients in the process of technology development
I	"Certain applications are shown to patients and health care professionals to take their opinions into account"	
C	"To ensure trust, you should include physicians in the development to give you advice"	
C	"Normally, we ask doctors if according to their standards we should do things differently"	
E	"Doctors and patients are the main stakeholder, and that is why I joined the company"	
F	"The most important stakeholders are clinicians, by far. They are followed by payers, patients and society"	
J	"Philips is changing from a technology push company to a market oriented company. For that, you have to involve the end users from early on"	
J	"We involve clinicians early on. They are since the brainstorm sessions to develop a product"	
D	"Patients should be involved in focus groups shaping regulations for AI in healthcare. If you can explain them what you do and why, they will be more willing to accept the technology"	Involve patients when developing AI regulations for healthcare
C	"To reduce the bias, scientist should work together with clinicians and experts that provide and annotate the data"	Make sure that health care professionals are included in the process of selecting and curating the data
F	"Introducing nurses and doctors in the process of curating the data is absolutely required because otherwise you will be learning from garbage and your prediction will be as bad as the data that you learnt from"	
D	"Philips has a lot of co-creation. We put together hospitals, patients, former companies and other device companies together in a room and we ask the following question: What kind of product would be nice for you?"	Set co-creation exercises with a wide range of stakeholders

B	"We do a lot of co-creation sessions with physicians and patients. We talk to each other to avoid surprises. They contribute to the ideation of products"		
B	"Nowadays, the company is more digital-oriented. For that reason, innovation must be open and we should collaborate with hospitals, universities and society"		
H	"Doctors, scientists, hospital managers and patients are all important in the focus groups when designing AI solutions"		
I	"Philips is working together with Microsoft to use some of their knowledge or platforms in which they have the expertise"	Work together with companies that have expertise in other fields	
F	"We have a branch called Healthworks, which scouts thousands of AI startups and gradually filters them to few hundred. Then a few dozen, etc. So, we have worked with startups in the space"		
C	"We give the doctors a percentage of accuracy to help them with their decisions"	Always include the percentage of accuracy of an AI solution	Responsiveness
A	"We inform our patients when they are using AI applications, to give them the autonomy to use it or not"	Always inform the patients or customers that they are interacting with AI and not a real person	
B	"Consumer monitoring applications should only be used for pre-screening. If you have a bigger problem, you should go to the doctor to have a further check or diagnosis"	Communicate that consumer-monitoring applications should only be used for pre-screening of diseases	
H	"The diversity of the data used to train the algorithm must be explained in advance. So, if the clinician knows that he is dealing with a population that was not taken into account in the data, he can be more alert"	Communicate to the HCPs the limitations of the AI solution	
B	"We have also the opportunity to do something non-sense or interesting on Friday afternoons, so there is no limitation by any stakeholders. Some very interesting products have come out from this Friday afternoon experiments. This encourages scientists to have some new discoveries, which is interesting"	Develop activities of experimentation and innovation without regulatory constraints (bootcamps, hackathons, etc.)	
J	"Philips is starting to develop bootcamps events to explore the potential of technology propositions"		
J	"The solutions that you design have to fit in the workflow already developed by the clinicians, you cannot create a new way of working for them"	Develop AI solutions that can be easily integrated in the existing systems	

E	Some companies are already working on machine learning techniques to avoid personal biases from the software developers. This is very new, but Philips should start working on it.	Develop and implement AI solutions to detect and avoid biases from software developers
C	Data principles must be translated into practices in all the research teams within the company.	
B	"In telehealth, people should always have the option to go to the hospital and talk to the nurse or the doctor"	Ensure that AI applications do not compromise patient autonomy
C	You should never have data scientists working by themselves when developing the algorithm. It must be done in a team and there must be peer review discussions going on.	Ensure that data scientists work in teams and that there are peer-review discussions going on
G	"The intended use should be well described and the system must function according to its fully specified intended use"	Ensure that your AI application is only employed for its intended used
G	"We have a cybersecurity team dealing with the safety of patient data"	Establish a strong cybersecurity team
A	"We do data security risk assessments lead by data security risk officers to act when data transfer may pose a risk"	Establish periodic data security risk assessments
H	If we skip the doctor and we cause damage to the end-user, yes, we are liable. Also, If you do not train them very well (case of Boeing with Ethiopian Airlines' pilots) the company is also liable.	Educate the HCPs on how to AI-enabled systems and on how to interpret the "black-box"
F	"We try in Philips research to say that you should commit yourself to a previous AI solution or a previous coded solution. If you replace the human, your previous device is the human. You have to outperform"	Only put in the market AI applications that are at least as good as the best doctor or the best solution already existent
F	"You need to learn from the best, so that your system is ok in comparison with the best. Only then you can move society and care ahead"	
H	"You are liable if you cannot prove that you are doing at least what a human doctor would do"	
A	"Some of our telehealth systems still have to implement interaction features to avoid social isolation of elderly patients. Short messages or updates from the doctor"	Implement interaction features in telehealth applications
A	"We have underestimated the value of social networks. A social network where telehealth patients can interact could be a nice way to avoid social isolation of patients"	
B	"Communication in telehealth can be improved with screens, VR or other technologies so that isolated patients can communicate with people at a different location. This can enhance engagement of the patients"	

A	"Privacy officers are involved right from the start, even before the formal review but also during the formal review there is a privacy officer that is a permanent member of the ICBE"	Involve privacy officers during all phases of technology development
B	"In the company, the privacy officers are very active shaping the data studies related to the GDPR"	
B	"We always need to consult a privacy officer if we want to collect data"	
H	"In Philips we get reviews and complains from users, then we iterate. The bias has at least a corrective measure"	Iterate based on the feedback from users
B	"We check the feasibility of the algorithm with small datasets, but we don't draw any conclusion based on those small datasets. We plan a new prospective study and we test the model with a large dataset"	Never draw conclusions based on small datasets
F	"Every document and every phase of AI development has to be faithfully validated"	Validate every document and every phase of AI development
G	"If you develop a learning algorithm that improves over time, you need to maintain a report on the learning of the application. So, that you can always monitor if the system is within the boundaries of intended use"	
A	"In the privacy office of Philips we are collecting a lot of information on the legal challenges that we face. We share these outcomes with the legislators to see how we can deal with these challenges"	Work closely with legislators on addressing legal challenges
C	"We work together with regulators, For example, we are collaborating with FDA and other companies to establish the regulation for dynamic algorithms that keep learning in the field when new data enter"	
D	"Companies, hospitals and the public should be more involved in shaping EU AI regulation"	
E	"For healthcare, I think we should go beyond GDPR. There should be a European entity where you can donate your data because the final goal is worth it. Some people donate their bodies when they die, it could be similar for data"	
F	"We participate in discussions about the freedom and the individuals' right of privacy versus aggregating data and moving science ahead"	
F	"We have the FDA providing discussion papers to all academics and industry and we comment there regularly to see how the things are moving forward. There are also very close discussions with governments and parliaments in multiple regions"	
A	"We have embedded the practices of the European Commission AI ethics code to deal with transparency issues. We should be able to explain how the algorithm works"	Work on new ML technologies to try to make the "black-box" more transparent
A	"The company should move a bit away from the "black-box" to be able to explain how we arrive at solutions at least at a very general level"	

B	"So far, we have tried to make the "black-box" a "grey-box" by trying to understand some things"		
B	"We are trying to identify the main features of the model to see how they influence the outcomes. It makes the "black-box" a bit more grey and easy to interpret"		
E	"If you made a decision, you have to be able to explain why you made a decision"		
F	"For deep learning, which is difficult to explain, we try to prove the transparency by explaining how we validate the outcome. It may be difficult to understand how it is exactly working, but we can at least show the curated dataset, the performance and the comparable human performance"		

Critical actions identified by the interviewees

Table 16. Coding of RRI actions to improve social acceptability

A2.5 Coding of technologies/products

Interviewee	Empirical data	First-order coding	Second-order coding
B	"If we talk about consumer applications, such as baby, pregnancy, or sleep monitors, these are short-term introductions as you can quickly put them in the market"	Consumer applications - Monitoring devices	Short-term
F	"For the short-term, we have already launched an intelligent shaver. We have also launched and AI-enabled solution for sleep-support. We have launched AI solutions in radiology to measure breathing and heart rate"		
J	"In patient monitoring, there are already products in the market"		
J	"In radiology there are already applications that show the radiologist the best locations to focus on"	Decision-support systems in radiology	
E	"In the short-term, I think that improving the image acquisition time of MRI machines is feasible"	Faster image acquisition time of MRI machines	
C	"In the short-term we can expect to see improvements is the operation side of the procedure. It will be faster, more efficient and less wasteful"	Operational applications - Hospital systems	
H	"Most of the applications for hospitals right now have to deal with flow optimization"		
E	"In the medium-term, you can skip the image analysis from the radiologist. You can go directly to diagnosis-support"	Diagnosis-support applications	Medium-term
H	"I see telehealth playing a role in the medium-term. Because we have the problem of aging population and hospitals are constantly full"	Telehealth	
F	"In the long-term, we can expect that most of our solutions will have an AI component visible to the user or working in the background"	All Philips products will be AI-enabled	Long-term
G	"In the long-term we expect to connect society, healthcare systems, hospitals and caregivers in an organized way. This will help us to offer predictive healthcare instead of reactive healthcare"	All-encompassing applications - Predictive care	
D	"In the long-term you can expect an all-encompassing thing. Siri can tell you what are the pros and cons of your decisions every day to help you steer your decisions in the healthiest way"		
C	"In the long-term, we change from procedures to outputs. Outcome-based applications that are personalized for the patient"	Personalization of health	

Table 17. Coding of AI technologies/products

Appendix 3 – RRI Roadmap

A3.1 Case Description

The company

Philips is a leading health technology multinational company established in Eindhoven in 1891 (Philips, 2019a). The mission of the company is 'to improve people's lives through meaningful innovation', with the *vision* of 'improving the lives of 3 billion people a year by 2030' (Philips, 2019b). Philips is a leader in diagnostic imaging, image-guided therapy, patient monitoring, healthcare informatics, and personal health (Philips, 2019c). The strategic focus is based on generating constant innovations to deliver on the Quadruple Aim of value-based healthcare: improved patient experience, better health outcomes, improved staff experience, and lower cost of care (Philips, 2019c).

The interest of Philips in the field of artificial intelligence has increased considerably in the last few years. However, the company prefers to talk about "adaptive intelligence", which refers to the use of AI to help analyzing large quantities of data to generate outcomes that support and empower people. Adaptive intelligence combines the power of AI with the contextual knowledge of Philips in the clinical world (Philips, 2018).

RRI commitment

- Philips is developing a code of data principles and AI ethics. This code will be based on the report "*Ethics Guidelines for Trustworthy AI*" made by the High-Level Expert Group on Artificial Intelligence at the European Commission (European Commission, 2018). Different regulatory boards within the company are reviewing it. The code follows exactly the same structure as the document from the European Commission, which is based on seven principles: "1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination and fairness, 6) environmental and societal well-being and 7) accountability". However, the indications given are very general. For example, "Identifiable and discriminatory bias should be removed in the collection phase where possible" is the main suggestion to avoid bias. The RRI roadmap and the results of this work are more practical in nature. For instance, instead of saying that identifiable and discriminatory bias should be avoided, we propose clear actions such as involving HCPs and patients in data curation and selection, or making culturally diverse software development teams. These recommendations represent a step further. The recommendations formulated in this work go from principles to practices.
- ***Motivation for RRI:*** Better understanding of the ethical, legal and social impacts and uncertainties related to the introduction of AI in healthcare. Exploring ways to improve the social acceptability of the AI products developed by the company.

Context

- ***Type of organization:*** MNC (77,000 employees)
- ***Country:*** The Netherlands (headquarters)
- ***R&I Projects selected:*** All the projects related to the field of medicine in which AI is involved
- **Technology:** Health monitoring devices, healthcare operational systems, medical devices
- ***Regulatory regimes relevant for Philips:*** GDPR, medical devices, AI principles
- ***Type of R&I activities:*** In-house and cooperative research
- ***Type of business:*** Business to business, business to consumer.
- ***Time to market (indicative):*** Depends on the product (see Figure 18)
- ***CSR Policies:*** Compliance with GDPR, Constant checks on product development from the Internal Committee of Biomedical Experiments (ICBE), 'Healthy people, sustainable planet' program, rigorous involvement of privacy officers in research.
- ***RRI maturity level:*** Strategic. According to Yaghmaei (2016), companies at this level consider RRI as an important part of their strategy. In the case of Philips, there are practices that can be considered as RRI actions, because they take into account dimensions of anticipation, reflection, inclusiveness, and responsiveness. For example, the creation of the Internal Committee of Biomedical Experiments (ICBE) to analyze the impacts of the products developed by the company reflects anticipation (Philips, 2019c). The introduction of HCPs in the process of data selection and curation reflects inclusiveness. The objective of the company of "augmenting the abilities of doctors" instead of replacing them shows a process of reflection (Philips, 2019e). The active role of privacy officers in research to evaluate continuously whether privacy is being respected is an action of responsiveness (Philips, 2018). Section 4.5. has a list of 48 RRI actions developed by Philips and related to their corresponding RRI dimension. All these practices are core to the strategy of AI technology development within the company. Philips is at a strategic maturity level in RRI but these practices are labeled as CSR and not RRI. This work introduced the concept of RRI for the first time in Philips.

Materiality & experimentation

- ***Key stakeholders:*** The key stakeholders were identified from the interviews. We asked directly to the experts, which were the most important stakeholders for AI research and development. Most of the interviewees mentioned the following stakeholders: Company (R&D, management and legal), AI scientists, privacy officers, ICBE, clinicians, external regulatory bodies.
- ***Key ethical, legal and social issues:*** The key concerns of AI in healthcare were identified in Chapter 2. These are accountability; data bias, fairness and equity; effects on HCPs; effects on patients; privacy and security; reliability and safety; and transparency and trust. (See section 2.2. for a more detailed explanation).

A3.2 RRI Roadmap

RRI vision

Develop AI technologies in healthcare to improve the lives of 3 billion people a year by 2030 (Philips, 2019b). Deliver on the Quadruple Aim of value-based healthcare: improved patient experience, better health outcomes, improved staff experience, and lower cost of care (Philips, 2019c).

R&I Technologies and products

- ***Consumer-oriented applications:*** Shaver Series 7000, Sonicare DiamondClean Smart, Pregnancy+, SmartSleep Analyzer
- ***Operational applications:*** PerformanceBridge, HealthSuite Insights
- ***Decision-support systems in radiology:*** IntelliSpace PACS with Illumeo
- ***MRI machines:*** Ingenia MRI machines
- ***Diagnosis-support applications:*** IntelliSpace Precision Medicine
- ***Telehealth:*** CareSage
- ***Digital twin:*** Philips digital twin concept
- ***All-encompassing applications – Predictive care:*** Wellcentive

Drivers and challenges for RRI

The drivers and challenges mentioned below are based on the results of the interviews (see sections 4.1 and 4.2).

Drivers

- Offer more affordable care
- Improve technological efficiency and accuracy in healthcare outcomes
- Keep the position as one of the industry leaders in medical technology
- Improve the experience of patients and clinicians

Challenges

- Lack of clarity on who should be responsible for an AI decision
- Lack of diversity in software developers
- Changes in the role and education of clinicians
- Lack of willingness to change from reactive to predictive care
- Difficulty to introduce contextual information from the clinical world

Risks and barriers to be addressed by RRI actions

The risks and barriers mentioned below are based on the results of the interviews (see sections 4.3 and 4.4).

- Unintended discrimination
- Privacy breaches

- Social isolation in telehealth applications
- Opposition from the medical community
- Reinforcement of inequality in healthcare
- GDPR is very strict
- Impossibility to explain the "black-box' in ML systems
- Complexity and lack of flexibility in Philips

RRI actions

The RRI actions mentioned below are based on the results of the interviews (see sections 4.5 and 5.6).

Anticipation and reflection

- Design procedures to ensure that data has a high-quality, it is well-curated and representative of the population
- Establish clear and rigorous practices of data management regarding privacy and consent

Inclusiveness

- Design cross-population studies to validate AI applications
- Ensure a high level of diversity in AI software development teams
- Involve clinicians and patients in the process of technology development
- Make sure that health care professionals are included in the process of selecting and curating the data

Responsiveness

- Implement internal policies that allow AI scientists to make quick changes once a project has started
- Develop activities of innovation and experimentation without regulatory constraints
- Develop and implement AI solutions to detect and avoid biases from software developers
- Ensure that data scientists work in teams
- Educate the HCPs on how to use AI-enabled systems and on how to interpret the "black-box"
- Implement interaction features in telehealth applications
- Work on new ML technologies to make the "black-box more transparent"
- Work closely with legislators on addressing legal challenges

Roadmap design

The aspects relevant for the RRI uptake by the company have been synthesized by the author in an overall diagram, following the visual approach described in the PRISMA exemplary roadmap (Figure 15).

The RRI roadmap developed in this work is a useful starting point for RRI uptake. If the abovementioned issues are answered in a collaborative practice with a wide range of stakeholders, Philips could move up to a civil level of RRI maturity, in which, the company converts in a role model that promotes RRI principles within the business environment and society.

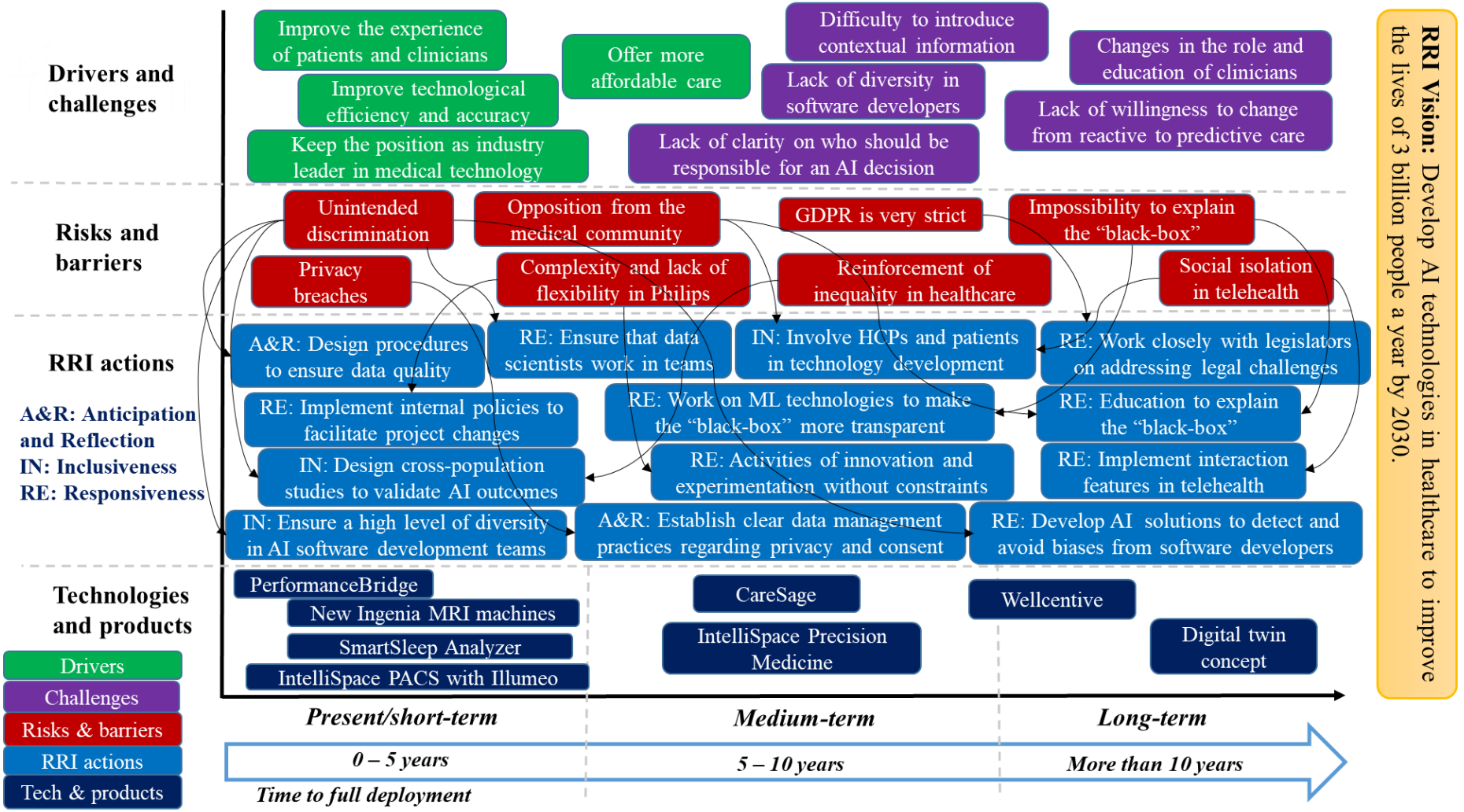

Figure 15. RRI roadmap for AI in healthcare (Philips case)