University of Manchester

THE ETHICAL CONCERN OF BIASES IN MEDICAL AI: WHO IS RESPONSIBLE FOR WRONG MEDICAL DIAGNOSIS DUE TO BIAS IN BLACK-BOX ALGORITHMS? A RESPONSE TO THE RESPONSIBILITY GAP.

‘A dissertation submitted to The University of Manchester for the degree of Bachelor of Science in the Faculty of Humanities.’

PHIL30002.

2022

Student ID Number: 10519580

Word Count: 5975 words

School of Natural Sciences

Supervised by

Dr. Jon Bebb

Acknowledgements

I would like to thank Dr. Jon Bebb for the insight and knowledge given on the topic as well as the guidance provided through the different stages of the dissertation. I would also like to thank my friends and family, in particular Claudio and Martina, for showing unconditional support and appreciation of my work.

Declaration

I declare this dissertation is my original work. No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Abstract

The aim of this discussion is to respond to the issue of the opening of a responsibility gap when assigning responsibility to biased medical diagnosis output by machine learning algorithms. The results will eventually show that there is no responsibility gap when AI is implemented in healthcare, since multiple agents can be ascribed responsibility. The first stage of the discussion will be focused on addressing the ethical issue of bias in AI machines; it will be shown how algorithms might perpetuate racial and gender discrimination. Secondly, the essay will focus on understanding who should be held responsible of unjust outputs produced by black-box algorithms, considering that no agent can comprehend the decision making of these systems. The essay will conclude that a key to ascribing responsibility when opaque AI is used in healthcare is to divide the notion of responsibility from blame, a strategy that enables the individuation of different senses of responsibility. Accordingly, both the physicians and developers should be held responsible but not blameworthy.

Contents

Chapter 1

Introduction

Machine learning algorithms elaborate an endless amount of data to provide efficient solutions to problems that for the human mind might be overwhelming and computationally long. These algorithms are particularly used in healthcare with the aim of enhancing clinical diagnosis, treatment recommendations and the healthcare system in general by improving its efficiency through automation. Notwithstanding the positive effects that implementing AI can bring to the healthcare system, there are worrying downsides and threats that medical AI has uncovered. These AI systems have been discovered to perpetuate injustices and discrimination, due to bias being found in their training data. In fact, machine learning systems require data to be trained on, but the information available in our world contain several sources of bias. Once these biased data are fed into the system, the machine cannot identify them and they will eventually alter the output, resulting in discrimination and harm. More strikingly, it is not straightforward to ascribe responsibilities for the harm created by increasingly complex and opaque AI systems, given that they cannot be understood nor controlled by the human mind. Thus, a responsibility gap is found. If AI is to be conclusively implemented in the healthcare systems, it must neither bring further any inequalities, nor leave a responsibility gap if harm is caused.

This essay aims at solving the responsibility gap, thus showing that responsible agents can be found. To reach such aim, Section 2 introduces the concept of bias and how it is found in healthcare. In this section, I will conclude that bias cannot be completely avoided since (a) in some cases it is useful to factor differences between groups, and (b) the sources of bias are multiple and ubiquitous, thus it is impossible to eliminate them from the data. Then, Section 3 will introduce the definition of black-box algorithms and show how the implementation of these algorithms will open a responsibility gap. Section 4 will then outline two examples of racial and gender discrimination provoked by black-box systems. These examples will unveil the need of finding a responsible agent for cases where physical and social harm is created. Who is to be held responsible if an algorithm makes errors based on biased data, which eventually create social injustices and physical harm? Answering this question is not straightforward given that many agents take part into the process of developing and implementing an AI machine. Thus, Section 5 focuses on individuating responsible agents when black-box algorithms are implemented in medical diagnosis, and their performance is discriminatory and harmful. Gunkel (2020) argues that the implementation of black-box AI systems opens a responsibility gap because it does not seem intuitive to hold any agent responsible of outputs produced by a machine that is not understandable by the human mind. Gunkel (2020) has argued that this gap cannot be filled, therefore AI poses a serious ethical problem when implemented in healthcare. I will show that this is not the case, and the aim of Section 5 is to provide a sense of responsibility according to which multiple responsible agents can be found. The section will first introduce the idea of understanding ”responsibility” as a polysemantic term, a strategy that will reveal multiple senses of responsibility. Then, I will propose a thought experiment describing a specific scenario of collaboration in medical diagnosis between AI and a physician, where the final diagnosis is biased, wrong, and harmful. Eventually, I will argue that both the physician and the developer of the machine have failed to fulfil their duties and as such are responsible, however they are not blameworthy. In fact, if they have not violated the obligations imposed by their duties, then they should not be blamed for such diagnosis. Overall, when responsibility is detached from blame, then the responsibility gap can be filled.

Chapter 2

How is bias found in healthcare?

According to the dictionary definition, bias is an intentional or unintentional preference towards a group or population over other groups. Usually, the term bias presents a negative association given that the outcomes of a biased prediction will result in discrimination towards a specific group, resulting in the perpetuation of gender and racial disparities, for instance. In machine learning, bias is often found in the data on which algorithms are trained. When it comes to learning, algorithms are trained on a “training data set”, which is fed to the machine to teach the algorithm how to make novel predictions or recommendations (Tan, 2019). If the training data set is biased, then the system is likely to output biased results. The sources of bias are various; in fact, biases could be mirroring social and historical discrimination, or be the outcome of algorithms introducing biases themselves by being unreasonably selective. Hence, it is important to understand what types of bias are found in medical data and how these will affect physicians’ practice. In medical algorithms, Cirillo et al. (2020) individuate the following sources of bias:

1. Historical bias arising from data reflecting previous discrimination, e.g., dark-skinned individuals are marginalised from accessing lung cancer treatment, given that the guidelines for accessing the screening were based on a past study where 94% of the participants were white people; as a result, they

do not meet the guidelines due to being underrepresented in historical data (Lewin et al., 2015).

2. Representation bias occurring when parts of the data set are underrepresented, e.g., a lower percentage of men is diagnosed with depression, since men are less likely to see a doctor, a psychologist or a psychiatrist; as a result, the female population is the primary focus of research on depression and it accounts for the majority of the data available for AI training (Norman, 2004).

3. Measurement bias occurring when data are approximated to some ideal quantity, here the labels given to data are not mirroring useful differences and are approximated to the most common - ideal - quantity, e.g., different symptoms of cardiac arrest are experienced by the two sexes, but this is not appropriately taken into account because only those symptoms suffered by the larger portion of the population are considered (Beery, 1995) (See Section 4.2).

4. Aggregation bias taking place when distinct groups with different conditional distributions are aggregated into one model, e.g., different levels of diabetes are found in different ethnicities, but this is not reflected into the data set (Spanakis and Golden, 2013).

5. Evaluation bias arising when the evaluation of an algorithm does not reflect the population, e.g., skin cancer diagnosis algorithms failing to detect dark-skinned female faces due to benchmark data coming from white men (Esteva et al., 2017); here dark-skinned people are not underrepresented in the data set, rather the benchmark is set on lighter skinned individuals so the algorithm passes the evaluation test merely by hitting the biased benchmark rather than being evaluated against an inclusive benchmark.

6. Algorithmic bias occurring when a wrong label is used consciously or un-

consciously, e.g., the US has implemented an AI system able to recommend which patient will receive more benefit from specific treatments to enhance the efficiency of the hospital's admission system. The algorithm individuates patients with a high hospital care rate as the patients who would benefit more from being admitted (Obermeyer et al., 2019). Here, the label is biased because in the US there is unequal access to hospitalisation given the costs of such access, so rather than predicting hospitality access history, the algorithm is predicting healthcare costs. Hence, less wealthy people will not be guaranteed access to healthcare given their low rate of previous admissions.

7. Instruments bias arising when only a part of the population has access to health record equipment resulting in a misrepresentation of the population, e.g., populations in unwealthy countries might not have access to wearable health tracking devices, excluding them from being represented in the data set (Cirillo et al., 2020).

For AI algorithms to be safely implemented in the healthcare system there is a need to acknowledge and address the ethical issues arising from the presence of bias in data. Eradicating such presence is not an easy task given the multiplicity of the sources of biases. Scholars have argued that to ensure fairness in data processing and modelling, developers should aim at removing sex and gender differences, as well as disparities in ethnicities (Mehrabi et al., 2019). However, in precise medical diagnosis it would be counterproductive not to consider gender or race differences, given that these groups do experience different symptoms or illnesses in the real world, and not reflecting such instances would result in being unfairly dangerous for the health of minority groups.

To reach an accurate diagnosis, several factors need to be included for a better and more precise outcome. Indeed, algorithmic fairness is difficult to define and is highly context-sensitive, in some cases it might be useful to consider differences between populations whereas this might lead to injustices when applied in other contexts (McCradden et al., 2020). It is fundamental that we find ways to mitigate biases that might result in unfair predictions. As Rajkomar et al. (2018) argues, ensuring equal outcomes is indispensable to fairness, which implies that minority and protected groups will experience equal benefit compared to the majority group, when being diagnosed with AI systems. Ensuring such algorithmic fairness is duty of the developers of the system, however it is not always possible to know whether the algorithm will introduce further disparities even if the training data set is claimed to be unbiased. As a result, physicians must be aware that there is the possibility of AI biased outputs, and if such discriminatory outputs are reached, then they might be involved when responsibilities of such error are assigned.

Overall, bias is an intrinsic characteristic of data, it is hard if not impossible and counterproductive to present a data set without any bias. However, the level of bias that is now present in algorithms used in healthcare is not under control and in need of both regulations and more awareness. According to the non-maleficence principle (Varkey, 2021) - the obligation of the physician not to harm patients - any novel medical tool must be assessed in favour of patient safety. Bias in algorithms threaten this ethical obligation, and as such there is the need to (a) understand how to not violate this principle, and (b) ascribe responsibility to an agent, or more, if that obligation is not fulfilled.

Chapter 3

Black-box algorithms and the responsibility gap

Several types of algorithms are used in healthcare. The most problematic category is black-box algorithms; thus, this dissertation will focus on cases where these algorithms are used by physicians and the result of such implementation causes harm and discrimination. What is concerning about black-box algorithms, or self-learning algorithm, is that they can enhance pre-existing disparities between groups without requiring any large amount of bias in their data. These algorithms are self-learning; this implies that they keep learning from the data available and from those they find online, to improve their performance (Vayenad et al., 2018). Indeed, if a self-learning algorithm is trained on a slightly biased data set, it will look for more biased data to enhance its performance.

An even more worrying consequence is that by continuously learning on their own, the reasoning behind these algorithms is extremely obscure. As Durán and Jongsma (2021) argue, a black box algorithm is a system whose workings are so complex that the human mind cannot interpret its reasoning. Indeed, since the functioning of these algorithms is opaque, physicians cannot explain the reasoning behind the diagnosis or recommendation made by black-box machines. This is a dramatic issue for healthcare given that ascribing responsibility in the medical environment revolves around the explanations of the diagnosis given by physicians. Thus, black-box algorithms posit a serious ethical problem: they are actively used in healthcare, but physicians do not and cannot have any understanding of how they work. So, can physicians be held responsible for the outputs of a system they cannot understand? If not, who is responsible for biased outputs that lead to discrimination and serious health complications? As Gunkel (2020) argues, black-box algorithms open a responsibility gap, given that it seems unintuitive to assign responsibility to any agent if the system that produces the error cannot be understood by them. I will argue that a responsibility gap is found if one takes responsibility to be connected to blame; however, this gap can be filled with the introduction of multiple senses of responsibility.

To reach this conclusion, it is first useful to outline a few cases in which these algorithms have caused a real harm and therefore have exposed the healthcare system to a crucial ethical issue.

Chapter 4

Examples of biases in medical AI algorithms

This section introduces two examples of bias in the healthcare environment that led to an injustice towards minority groups based on their race (Section 4.1) and gender (Section 4.2). The introduction of these real cases is key to understanding how pressing the issue of responsibility is when talking about opaque, self-learning algorithms, given the great harm that their outputs create.

4.1 Racial discrimination

An AI system for diagnosis of skin cancer has recently been put into trail; this system applies a deep convolutional neural network algorithm, which is considered one of the most complex models of a black-box algorithm (Esteva et al., 2017). This machine is able to diagnose malignant skin lesions from image processing, by being trained on a data set containing skin images. In the history of skin cancer, people with pale skin have suffered from this illness in greater numbers than people with darker skin. As a result, given the vast availability of data on White people, the algorithm has trained itself on a biased data set and it does not produce accurate diagnoses to people with other types of skin colour. Indeed, if the system is not trained with an inclusive data set, then it will produce biased responses when used in a healthcare setting presenting a diverse population. This is an example of evaluation bias resulting in racial discrimination, where all the available data belonged only to a specific group.

4.2 Gender discrimination

On top of instances of racial discrimination, cases where applying AI systems has resulted in gender discrimination can also be found. The following example shows the results of the implementation of an AI algorithm used in detecting cardiac arrest symptoms (Beery, 1995). This device enables the immediate diagnosis of heart attacks by being trained to recognise symptoms such as chest pain. The bias in this case arises from the fact that male and female individuals present different type of cardiac arrest symptoms. In fact, women are more likely to suffer from nausea and abdominal pain, but given the small number of women suffering from heart attack, these symptoms are not part of the data set on which the AI system has been trained. Historically, men have suffered from heart attack more than women, thus, the data available is dominated by men, with women occupying only 25% of the training data set of the algorithm (Petursdottir, 2021). Overall, by focusing the diagnosis on symptoms suffered by male individuals, female patients are more likely to be misdiagnosed, putting them at risk of a fatal heart attack.

Chapter 5

Who is responsible for biased, discriminatory and harmful AI outputs? A response to the responsibility gap.

The previous discussion has reached the conclusion that it is almost impossible to have an unbiased data set, given the multiple origins of bias. Moreover, bias might sometimes be desirable to provide an accurate diagnosis. How can one make sure that AI is ethically implemented in the healthcare system then? To respond to this pressing question, it is fundamental to understand who should be held responsible if the AI system fails and produces a biased judgment, where the judgment is then reflected in the diagnosis given by the physician who is not aware of the wrongness and bias contained in the output. Therefore, it is fundamental to assess whether any of the agents participating in the creation and implementation of AI are responsible, or if a responsibility gap is created. If such gap is found, then AI systems cannot be safely implemented in healthcare given that no one can be held responsible for the harm caused by its actions.

First, we need to establish the reliability of AI algorithms. In fact, for physicians to be justified in trusting black-box algorithms, it is necessary that those algorithms are reliable. Assigning reliability to such systems is a hugely debated topic, considering that full reliability cannot be attributed as long as examples where these systems fail can be found. For the scope of the present discussion, I will assume that these systems are reliable in the sense that the physicians are justified in employing and trusting them even though they might fail in an infinitesimal number of cases. This assumption is justified by the fact that these algorithms are based on a reliable process that is highly probable to output trustworthy results; in fact, algorithms that pass the training stage are the ones that output satisfactory results ≈99% of the time (Durán and Formanek, 2018). Moreover, if checked against tools already in use in healthcare, black-box algorithms are as accurate, if not more. For instance, MRI have been shown to reach an accuracy rate of only ≈90% (Taylor et al., 2019). Indeed, I will assume the reliability of black-box algorithms for the present discussion.

As we have seen, multiple agents are involved in the creation and implementation of AI systems: from the ones collecting and organising the training data set, the one developing the system's algorithm, the people performing analysis and checks and finally the one implementing the system, i.e., the physicians. It is not easy and might be controversial to assign responsibility only to one of these agents, given that the harmful output of the system could have been caused by an error made in any of the stages. More importantly, I claim that due to the different aspects of automation and the several sources of error that could result in a wrong diagnosis, there is the need to introduce a distinction between different senses of responsibility. Capturing responsibility as a polysemantic term is not a novel strategy in philosophy and reflecting this multiplicity in meaning will unveil different responsible agents, and thus solve the responsibility gap.

The discussion on responsibility will be structured as follows. First, three different senses of responsibility will be outlined. Then, a thought experiment will be introduced to set a precise scenario on which the discussion will focus. This case will show an instance of wrong medical diagnosis which sees the cooperation between a black-box algorithm and the physician. Note that the discussion will not include self-learning systems whose decisions are blindly followed by the doctor. Finally, I will conclude that both the physician and the developer must be held responsible for the wrong diagnosis reached due to bias in the AI system, even if they have performed no wrong action. The fact that the diagnosis is not a result of wrongdoings makes them not blameworthy, but still responsible under a specific sense of responsibility.

5.1 Different senses of responsibility

The term “responsibility” can be given different senses. This idea was first introduced by Hart (1968) who provides a list of four meanings of responsibility: role-responsibility, causal responsibility, capacity-responsibility and liability responsibility. When assessing the kinds of responsibility resulting from artificial intelligence, this list is slightly altered. Thus, I will outline three different senses of responsibility that apply to cases of AI implementation in healthcare. With the introduction of this distinction, I will be able fill the gap that Gunkel (2020) has argued to occur when black-box algorithms are used.

Commonly, we use the term “responsibility” when we associate blameworthiness to an action; however, it is not only through blame that we understand whether an agent is responsible. We can individuate the following types of responsibility: retrospective responsibility or culpability, prospective responsibility and moral accountability (Sio and Mecacci, 2021).

Retrospective responsibility is defined as blameworthiness for wrongdoing resulted from intention, knowledge or control (Tannenbaum, 2018). This sense of responsibility is centered around the moral judgment of an agent's actions, where the judgment leads to consequences such as feeling remorse, being blamed etc. (Garrath, 2006). If an agent is responsible in the retrospective sense, then they are blameworthy.

Prospective responsibility is strictly related to the notion of "duty" (Garrath, 2006). Someone is responsible in this sense if they failed to perform their duty or to meet the obligations of their duty. Note that in the case of physicians, such duties and obligations are contained in the Hippocratic Oath, which states that doctors are responsible for treating patients to the best of their ability without causing any harm (Knott, 2021). It is important to note that failing to meet the obligation of one's duty can be the result of both a wrong and a good action. A fireman, whose obligations include rescuing trapped people and animals (Department for Communities and Local Government, 2012), might fail to save a cat in a burning house even if they have performed to the best of their ability. Indeed, prospective responsibility can still be connected to blame but this is not a requirement. One can still be responsible in this sense without being blameworthy.

Finally, moral accountability is defined as the responsibility of a person to provide an explanation of the reasoning behind their actions to others (Sio and Mecacci, 2021). People might account someone moral responsible with the intention of judging and blaming their actions' explanations, but it might as well be used in a less threatening way, as being expected to respond to one's actions without being open to blame. Indeed, someone who is morally accountable for their actions might be blameworthy if their explanations are unjustified or might not hold the blame if their explanations were logical and justified.

These three senses of responsibility will now be applied to the following thought experiment to understand what agent and in which sense of responsibility they should be held responsible for.

5.2 A thought experiment

As I have explained previously, the implementation of AI systems in the healthcare environment sees the participation of a long chain of agents. To make the analysis of responsibility clearer, the following scenario will involve the participation of only a few agents. Imagine a case where a group of developers have built a reliable AI system able to detect skin cancer (See Example 4.1). In building this machine, the developers of the system have followed their duty's obligations to build a fair AI machine meeting the ethical requirements of the healthcare system outlined in the Hippocratic Oath. Specifically, for the machine to meet such regulations, it has shown a 99% success rate. The AI machine is then assigned to a doctor, who has been informed and trained by the developers on its opacity in reasoning since the system is a black-box. The doctor is now giving a diagnosis to a dark-skinned patient who has been called by the hospital for a general screening (note that the patient has no symptoms of any illness). The physician is aiming at providing a beneficial diagnosis to the patient to the best of their abilities and tools, so they will implement the AI system in order to have a second opinion to confirm or revaluate their hypothesis. Indeed, this is a case where the aim of the system is to help the doctor in performing a more efficient and comprehensive diagnosis. After having examined all the evidence, the doctor has good reason to think that the patient does not have skin cancer but there are still a few factors leaning towards the opposite diagnosis. In this case, the AI system is in favour of the doctor's most probable option and the physician takes this additional component as a further and final factor towards the "no skin cancer" diagnosis. The physician presents their diagnosis to the patient, but after a period of time, the diagnosis is found to be wrong, and it is discovered that the AI output was inaccurate due to its data set being biased. As a result of the diagnosis, the patient encounters serious health complications due to having an undiagnosed and unmedicated skin cancer.

In this case, all the evidence was leading to a wrong output and the machine has given the extra factor leaning the doctor's decision towards a diagnosis that was harmful and discriminatory.

5.3 Who is responsible and in what sense of responsibility?

The scenario poses a serious problem since it seems intuitive not to hold any agent responsible given that they performed good actions and the error in the diagnosis is partially the result of an unexplainable but reliable system. However, given that the diagnosis has resulted in harm and injustice, some responsibility must be ascribed.

According to the previous definitions of responsibility, both the developers and the physicians cannot be ascribed retrospective responsibility. This is because they both lack knowledge, intentions, and control. The doctor lacks knowledge, given that they do not fully understand how black-box algorithms function. The developers are lacking a different form of knowledge, since they do understand the theory behind the functioning of the system, but their knowledge does not cover the underlying workings of the system as well as how the output produced can be applied in healthcare; they do not understand how the machine has reached that output even though they have developed the system. Besides knowledge, both agents present the right sort of intentions: the developers have built the system in such a way that would be beneficial to the user, and they have worked to the best of their ability to reach such aim; in the same way the physician has implemented a logical and evidence-based analysis, balancing all the factors - including the output of the AI system - with the aim of presenting a beneficial diagnosis to the patient. Finally, the physician does not have control over the machine in the sense that they cannot alter the system's output. The same line of thought can be applied to the developers, who in addition have no control over how the machine will be used by the physician.

Overall, the various agents in this scenario have not caused harm due to maliciousness, recklessness, or negligence. The physician has correctly adopted the ends available to them: implementing the AI systems as a consultant, doing their own research and assessing the evidence neutrally; in addition, the physician's beliefs were reasonable and logical given that they were the conclusion of evidence-based analysis. The physician drew the appropriate conclusions according to the results collected and they acted on such conclusion in the appropriate way, i.e., by providing the final diagnosis. The same applies to the developers: they have implemented their knowledge correctly and with the correct aim of producing a machine that could be beneficial for the physician and the patient, their beliefs for approving the machine were reasonable given the reliability of its results and they correctly acted on such conclusion by putting the machine on the market. Indeed, no agent is to be held retrospectively responsible, and as such, physicians and developers should not be blamed for the outcome of the diagnosis. For Gunkel (2020), a responsibility gap is found because (a) no agent can be blamed for harmful and discriminatory diagnosis and (b) no agent can be held morally accountable for such outcomes. I have shown the plausibility of claim (a) and I will now analyse why (b) is also true. However, I will claim that even though (a) and (b) are plausible, Gunkel's conclusion is incorrect, and the responsibility gap can be filled.

Following the previous discussion, the process behind black-box systems is unexplainable to the user and the people who have developed it. Indeed, as Grote and Berens (2020) argues, by not being able to give explanations of these systems, physicians and developers cannot be held morally accountable of the actions resulting from a biased and mysterious machine. In the thought experiment I have introduced, the only justification the physician can give of their diagnosis is that "all factors considered, the no-skin-cancer diagnosis was the most probable one". However, when justifying the reasoning behind all factors, the physician cannot account for the reasoning behind the AI system, which was used as the final factor leading to the diagnosis - note that all the evidence were already leading to such diagnosis, so the AI is the final but not the most important factor. As a result, they are not able to provide a complete explanation of the decision-making process due to one of the factors being unexplainable; indeed, the physician cannot be held accountable for the outcome. Following the same line of thought, even the developers cannot explain why and how the machine has reached such output. The only justification they can give is explaining what training data set was used and what type of algorithm is the machine employing to reach the desired output, which we have assumed to have been chosen to the best of the developers' abilities. Therefore, by lacking the ability to explain the output of the machine due to its opaqueness, the developers are not morally accountable for the wrong diagnosis.

As Gunkel (2020) has claimed under the term "responsibility gap", the fact that a black-box AI system is helping the doctor in making a diagnosis makes the agents in the scenario neither blameworthy nor morally accountable for the outcome of such diagnosis. If no machine was in the picture, then the doctor would be morally accountable, given that they would be in the position to fully explain their actions. In some cases, they would even be considered blameworthy if they are recognised to have performed wrong actions. If I were to conclude discussion at this point, as Gunkel does, then not only would we have a responsibility gap, but we would have given motivations to those scholars who argue that it is the machine who is responsible. However, as Siponen (2004) argues, one would end up living in a society that blames technology for humans' doings. Indeed, the gap must be filled.

What gives the intuitions that an agent is responsible for the harmful consequences observed in the scenario presented above is the fact that the obligations at stake in the healthcare environment have not been fulfilled. More precisely, the physician has not satisfied their duty's obligations not to cause any harm and act in the patient's best interest. Accordingly, I argue that the physician must be held responsible in the prospective sense since they have not met the obligations elucidated in the Hippocratic Oath. Does this mean that they should be blamed? I have already concluded that they are not to blame in the retrospective sense; however, it might still be the case that they are blameworthy in the prospective one. It is important to highlight that the responsibility gap is created because the physician does not seem to be blameworthy of their action and many scholars argue that if the physician is responsible, then they must be blameworthy (Gunkel, 2020; Matthias, 2004). Thus, for these authors, responsibility is strictly connected to blame. I believe that this does not happen in this case, because responsibility and blameworthiness are not always related. In fact, the physician has not satisfied their obligations, which makes them responsible, but these obligations were not violated: the physician has failed in their duty by still performing good actions. The failure was not a result of wrongdoings, and as such, the physician should not be considered blameworthy. Yet by failing to satisfy their obligations of not harming the patient, the physician's actions are morally insufficient, therefore they are responsible for the harm caused. The sense in which these actions are insufficient is that the agent has failed to perform what they were morally required to do. However, no blameworthy conditions of such performance can be identified.

According to this conclusion, the responsibility gap individuated by Gunkel is filled. However, I claim that the physician is not the only agent who is responsible for the wrong medical diagnosis; the developers share such responsibility. The developers have failed to prevent the machine from producing a biased output, which is included in their duties of building a fair and ethical AI system that can be safely implemented in healthcare and can benefit the physician's practice as well as the patient's health. Can the developer be considered blameworthy? They have acted to the best of their possibility, so the result is not a consequence of wrongdoings; one could argue that they could have chosen a different training data set, but this would have not changed the outcome, considering that an unbiased training data set is impossible to be obtained (See Section 2). Hence, they had the right intentions, their actions were adequate to their obligations since those obligations were not violated. Overall, the developers are responsible in the prospective sense, together with the physician, but without being blameworthy.

Overall, I have proven that the responsibility gap claimed by Gunkel is filled when considering different senses of responsibility and separating these senses from the notion of blameworthiness. Agents can be ascribed responsibility even if their actions are good, but given such goodness, they are not blameworthy. In the scenario presented, the physician and the developer are to be held responsible because they have not fulfilled their duties, but they cannot be morally blamed because such outcome was not reached by wrongdoings performed by the agents. Gunkel might be correct in individuating a gap; however, the gap is in blameworthiness rather than responsibility, which does not posit a crucial problem for the ethical implementation of AI in healthcare.

Chapter 6

Conclusion

This thesis has investigated one of the most pressing ethical concerns raising from the implementation of AI in healthcare; in fact, the presence of biased AI systems perpetuates injustices based on gender and race. More importantly, the harm caused by such discrimination is left without an agent to be held responsible. If AI is to be implemented in the healthcare system, no responsibility gap can be left unsolved. Indeed, I have proved that responsible agents can be found when reflecting on different senses of responsibility.

First, I have analysed the concept of bias and its sources to understand whether its presence could be completely avoided or not. Given that unbiased data sets are impossible to obtain and might not even be desirable, one must understand how to work with such data and reflect on the ethical responsibility of misjudgements due to biased AI diagnosis. More importantly, it is important to ascribe responsibility when complex and opaque algorithms are used, such as black-box algorithms. Scepticism can be found among scholars on assigning responsibility to physicians implementing these mysterious systems, given that they cannot understand the functioning of such machines. Accordingly, Gunkel claims that a responsibility gap is created when opaque AI is implemented in healthcare: when black-box algorithms are used, physicians should not be held responsible if the system makes an error given that they cannot understand the reasoning behind such output. I have argued that the responsibility gap can be filled by eviscerating the notion of responsibility and divide it from the idea of blameworthiness. Therefore, I have introduced the concept of retrospective responsibility, moral accountability and prospective responsibility. Additionally, not only is the gap filled but more than one agent should be held responsible. Indeed, according to the definition of prospective responsibility, both physicians and developers are responsible, but they are not blameworthy if they performed according to their obligations, i.e., with the interest of being beneficial to others to the best of their ability. This discussion has been elucidated by the introduction of a thought experiment based on a specific case of medical diagnosis involving the cooperation between the physician and a black-box machine. Both the developers and the physician present in the case examined were responsible by failing to satisfy their duties even though the failure was not the result of wrongdoings. Overall, they are responsible but not blameworthy and the responsibility gap is closed.

Bibliography

Beery, T. A. (1995), 'Diagnosis and treatment of cardiac disease: Gender bias in the diagnosis and treatment of coronary artery disease', *The Journal of Critical Care* **24**, 427 – 436.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S. and Mavridis, N. (2020), 'Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare', *Digital Medicine* . Accessed: 2022-04-01.
URL: *https://doi.org/10.1038/s41746-020-0288-5*

Department for Communities and Local Government (2012), 'Roles and responsibilities of fire rescue authorities'. Accessed: 2022-04-01.
URL: *https://www.local.gov.uk/topics/fire-and-rescue/fire-role-models/roles-and-responsibilities-fire-rescue-authorities*

Durán, J. M. and Formanek, N. (2018), 'Grounds for trust: Essential epistemic opacity and computational reliabilism', *Minds and Machines* **28**, 645–666.

Durán, J. M. and Jongsma, K. R. (2021), 'Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai', *Journal of Medical Ethics* **47**, 329–335.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. and Thrun, S. (2017), 'Dermatologist-level classification of skin cancer with deep

neural networks'. Accessed: 2022-03-03.
URL: *https://licensing.eri.ed.ac.uk/i/*

Garrath, W. (2006), 'Responsibility', *Internet Encyclopedia of Philosophy* . Accessed: 2022-04-10.
URL: *https://iep.utm.edu/responsi/*

Grote, T. and Berens, P. (2020), 'On the ethics of algorithmic decision-making in healthcare', *Ethics and Philosophy Lab* **46**, 205–211. Accessed: 2022-02-01.
URL: *http://jme.bmj.com/*

Gunkel, D. J. (2020), 'Mind the gap: responsible robotics and the problem of responsibility', *Ethics and Information Technology* **22**, 307–320.

Hart, H. L. A. (1968), 'Punishment and responsibility', *Oxford University Press* .

Knott, L. (2021), 'The hippocratic oath and good medical practice'.
URL: *https://patient.info/doctor/ideals-and-the-hippocratic-oath*

Lewin, G., Morissette, K., Dickinson, J., Bell, N., Bacchus, M., Singh, H., Tonelli, M. and Jaramillo, G. A. (2015), 'Recommendations on screening for lung cancer', pp. 425–432. Accessed: 2022-04-01.
URL: *www.cmaj.ca/lookup/suppl/*

Matthias, A. (2004), 'The responsibility gap: Ascribing responsibility for the actions of learning automata', *Ethics and Information Technology* .

McCradden, M. D., Joshi, S., Mazwi, M. and Anderson, J. A. (2020), 'Ethical limitations of algorithmic fairness solutions in health care machine learning', *The Lancet Digital Health* **2**, e221–e223.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2019), 'A survey on bias and fairness in machine learning'. Accessed: 2022-04-01.
URL: *http://arxiv.org/abs/1908.09635*

Norman, J. (2004), ‘Gender bias in the diagnosis and treatment of depression’, *International Journal of Mental Health* **33**, 32–43.

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019), ‘Dissecting racial bias in an algorithm used to manage the health of populations’. Accessed: 2022-02-10.
URL: *https://www.science.org*

Petursdottir, T. (2021), ‘Is ai sexist?’. Accessed: 2022-03-11.
URL: *https://www.cambridgenetwork.co.uk/blog/ai-sexist*

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. and Chin, M. H. (2018), ‘Ensuring fairness in machine learning to advance health equity’, *Annals of Internal Medicine* **169**, 866–872.

Sio, F. S. D. and Mecacci, G. (2021), ‘Four responsibility gaps with artificial intelligence: Why they matter and how to address them’. Accessed: 2022-04-01.
URL: *https://doi.org/10.1007/s13347-021-00450-x*

Siponen, M. (2004), ‘A pragmatic evaluation of the theory of information ethics’, *Ethics and Information Technology* **6**, 279–290.

Spanakis, E. K. and Golden, S. H. (2013), ‘Race/ethnic difference in diabetes and diabetic complications’, *Diabetes Epidemiology* . Accessed: 2022-04-01.
URL: *http://www.cdc.gov/*

Tan, S. (2019), ‘How to train your ai’. Accessed: 2022-04-15.
URL: *https://medium.com/revain/how-to-train-your-ai-98113bdac101*

Tannenbaum, J. (2018), ‘Moral responsibility without wrongdoing or blame’, *Oxford Studies in Normative Ethics* pp. 124–148.

Taylor, S. A., Mallett, S. and Ball, S. (2019), ‘Diagnostic accuracy of whole-body mri versus standard imaging pathways for metastatic disease in newly diagnosed non-small-cell lung cancer: the prospective streamline l trial', *The Lancet Respiratory Medicine* **7**, 523– 532.

Varkey, B. (2021), 'Principles of clinical ethics and their application to practice', *Medical Principles and Practice* **30**, 17–28.

Vayenad, E., Blasimmeid, A. and Cohen, I. G. (2018), 'Machine learning in medicine: Addressing ethical challenges'.