Epistemic (in)justice, social identity and the Black Box problem in patient care

Student name: Muneerah Khan

Student number: 550470

Submitted to the Faculty of Health Sciences, University of Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree Master of Science in Medicine (Bioethics and Health Law).

Supervisor: Dr Cornelius Ewuoso

Qualifications: PhD

Position: Senior Lecturer

Date: 27 Feb 2024 (Revised Submission)

Academic Integrity Declaration

I Muneerah Khan

(Student number: 550470)

am a student registered for MSc (Med) Bioethics and Health Law in the year 2024.

I hereby declare the following:

1. **I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.**
1. **I confirm that the work submitted for assessment for the above course is my own unaided work except where I have explicitly indicated otherwise[1].**
2. **I have followed the required conventions in referencing the thoughts and ideas of others.**
3. **I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.**

Disclosures:

1	**I have made use of AI based large language model applications (such as** ChatGPT, BERT, LaMDA, PaLM and LLaMa).[2]	**Yes**	**No**	**Specify:**
2	**I have made use of paraphrasing software tools (such as** QuillBot, Parafrasist, Ivypanda, or the paraphrasing functionality in grammar assistance tools).[2]	**Yes**	**No**	**Specify:**
3	**I have made use of spelling and grammar assistance applications (such as** Grammarly, Word spelling and grammar functionality, Grammarcheck, Writer, Zoho.)	**Yes**	**No**	**Specify: Grammarly**
4	**I have made use of referencing software tools, (such as** Zotero, Mendeley and EndNote)	**Yes**	**No**	**Specify: Mendeley**
5	**I have received assistance with my spelling and language from another person.**	**Yes**	**No**	**Professional or non-professional?**

[1] Students may talk to one another about their work and they may seek assistance from the academic staff of the Steve Biko Centre. Seeking the assistance of a Centre lecturer other than your unit/block head is discouraged and is only permissible with the consent of your unit/block head or the Centre Director. Assistance with spelling and grammar may be obtained from professional or non-professional language editors. Seeking assistance from other academics, lawyers or experts is prohibited.

[2] Making use of these tools is prohibited. But, if you have done so, it is better to disclose this, than to try to conceal it.

6	**I have received any other kind of assistance from another person, other than the unit/block head.**	**Yes**	**No**	**Specify:**

Signature: _____M Khan_____

Date: _____27 Feb 2024_____

TABLE OF CONTENTS

DEDICATION

This Research Report is dedicated to my family, who have shown unwavering support in my pursuit of greatness.

ACKNOWLEDGEMENTS

I want to express my deepest gratitude to Dr Cornelius Ewuoso for his guidance throughout the project. I would also like to acknowledge The Steve Biko Centre for Bioethics for the knowledge and resources that facilitated this project's success. Equally, I would like to acknowledge my Wits RHI Shandukani team for accommodating and supporting me through this process. Finally, it is important to mention that this report is being considered for publication in the journal Medicine, Healthcare and Philosophy and may appear in a print version before the examiner receives it.

LIST OF ABBREVIATIONS

- **AI –** Artificial Intelligence
- **IBM –** International Business Machines Corporation
- **PwC –** PricewaterhouseCoopers
- **IDI –** In-depth Interview
- **GDP –** Gross Domestic Product

ABSTRACT

This research report draws on (for the first time) the moral norms arising from the nuanced accounts of epistemic (in)justice in the work of decolonial scholars, and social identity in the work of relational autonomists to defend the thesis that using AI *in patient care* in light of the Black Box problem is deeply problematic and is ethically impermissible. This does not necessarily doom AI since it may be used for other purposes within the healthcare system. The report highlights what needs to happen to align AI with the moral norms it draws on. Deeper thinking – from backgrounds other than decolonial scholarship and relational autonomy – about the impact of AI on the human experience needs to be done to appreciate any other barriers that may exist. Future studies can take up this task.

CHAPTER ONE

GENERAL INTRODUCTION

Background

Artificial Intelligence (AI) – described as the "simulation of human processes by machines" (Burns et al., 2023) – is increasingly being integrated into different industries. One of the earliest conceptualizations of AI was by the mathematician Alan Turing in the 1950s when he developed the "imitation game" to test if machines were capable of achieving human intelligence (Turing, 1950). Initially, AI extended to *simple* systems that were limited to "if-then rules" (Kaul et al., 2020). Since the 1950s, AI has evolved to include complex algorithms such as Machine Learning and Deep Learning, creating a platform to optimize operations in a variety of industries (Kaul et al., 2020). These industries include financial services, aviation, insurance, telecommunications, and healthcare (Hastings, 2022). From recognising abnormal transactional patterns for fraud detection in the finance industry (Regunath, 2021) to providing solutions to air traffic control during unfavourable conditions in the aviation industry (Saini, 2022), the uptake of AI has taken the world by storm. In 2022 alone, the *IBM Global AI Adoption Index 2022* reported a 35% increase in the adoption of AI by various companies and reported that a further 42% expressed an interest in exploring AI (International Business Machines Corporations [IBM], 2022). A report released by PricewaterhouseCoopers (PwC, 2017) on the economic potential of AI estimated that AI could contribute to a 14% increase in Gross Domestic Product (GDP) by the year 2030. This potential for economic growth is likely to see adoption figures steadily increase over the years to come. The impact of AI has been so prominent that it is said to be at the core of the Fourth Industrial Revolution.

The healthcare industry remains one of the front-running industries likely to benefit from AI. The applications of AI in healthcare are numerous. Within the healthcare context, AI promises to revolutionise healthcare delivery in ways that will allow extraordinary machine functionality to move beyond science fiction. From brain-computer interfaces that restore communication channels in patients with neurological dysfunction to the development of radiology tools capable of performing "Virtual Biopsies", and smart devices designed to monitor critically ill patients (Bresnick, 2018), the sophisticated application of AI in healthcare holds much promise.

Globally, AI is being more readily adopted in patient care to deliver personalised care, with a projected timeline of limited use by 2025 and extensive use by 2035 (Bohr & Memarzadeh, 2020). However, the full adoption of AI to optimise and/or deliver tailored care to patients' specific needs will likely be inhibited by the AI Black Box problem. Broadly, "[a black box problem] occurs whenever the reasons why an AI decision-maker has arrived at its decision are not currently understandable to the patient or those involved in the patient's care because *the system itself* is not understandable to either of these agents" (Wadden, 2021). In other words, a Black Box is problematic – within the context of patient care – because of a *current* gap in *understanding* between the AI and the users of the AI, including the potential beneficiaries – like the patient – of the technology. This gap exists because *the system itself* is not understandable. Some studies explain that this is not the same as one's inability to understand a diagnosis (Wadden, 2021). For example, symptomatic anaemia is understood universally much in the same manner, and an AI might suggest a blood transfusion as a suitable mode of treatment. However, what values of the patient did the AI consider

in its output? Why did the AI conclude that a blood transfusion is warranted? This question would be relevant to certain cadres of patients, such as a Jehovah's Witness whose religious beliefs do not allow them to receive blood transfusions.

Yet, the non-understandability of AI is not the only problem a Black Box creates. A Black Box problem also occurs because the system itself is *currently unexaminable.* In other words, the system is opaque. That is, the user cannot work their way back from the outputs to interrogate its justification for that output. This may lead to concerns such as, what *degree of priorities* are placed on what values or variables? What *reasons* are provided for what prognosis?

A Black Box may become examinable or understandable in the future. However, the current inability to examine these AIs creates several gaps relevant to clinician-patient relationships. Notably, it creates a knowledge/information exchange gap. For example, what cannot be understood cannot be adequately shared between the machine, the user (clinician) and the benefactor (patient). Yet, sharing relevant material information about a patient's clinical care is essential to developing trust between a clinician and a patient. It allows for transparency of clinical processes and promotes patient autonomy through the informed consent process. In the absence of a lucid explanation on how AI technology works (or its internal operations) in patient care, trust, transparency and autonomy are undermined.

The Black Box problem also creates a responsibility gap. The responsibility gap implies that a health professional cannot be held accountable, liable, and culpable for

an (intentionally wrongful) action. To understand how, the reader should take note that control condition (that is, one is the agent of – or has sufficient control over – the action) and epistemic condition (that is, one is fully aware of the action) are core conditions for impugning responsibility (Neri et al., 2020). The Black Box nature of AIs implies that clinicians can neither control nor know what variables are considered in an AI output. They cannot tell whether and what consideration an AI gives problematic variables like gender, race, orientation, etc., to prevent harmful consequences to patients (Afnan et al., 2021). Yet, clinicians ought to be accountable or responsible for their actions. Clinicians ought to be able to explain, justify and defend why implementing one intervention is the preferable, rational and reasonable option in place of another. And in the event that an action results in harm, they ought to be liable and culpable, that is, morally blameworthy for harmful wrongs.

Suppose a course of treatment is AI-generated. In that case, one may ask who should be responsible when something goes wrong if clinicians cannot be blameworthy for such actions? AIs lack the freedom and agency to be held (civilly) liable for their actions (Neri et al., 2020) and, thus, compensate patients for any harm. In the absence of an agent who can rightfully be liable for AI actions, wrongful and intentional wrongs can hardly be determined and redressed. Precisely, negligence frameworks require clinicians to compensate for actions they are responsible for (Sullivan & Schweikart, 2019). Suppose clinicians and AI cannot be held accountable for AI-generated outputs. In that case, patients cannot be compensated for wrongs arising from such outputs, specifically in settings where AI adoption is still in its infancy, and the facility and/or health insurance compensation regulations and guidelines have not been amended to include AI use and its shortfalls.

The knowledge exchange and responsibility gaps undermine the requirement to foster a patient's informed decision-making capacity and create an informed decision-making capacity gap in the process. A clinician cannot share information that they do not have/know, and a patient cannot understand what is not shared. Why is one course of treatment preferable? When understanding is absent, a patient cannot make informed health decisions. The reader should notice that a decision-making capacity gap is created not because a clinician cannot provide *any* information. How much information a clinician ought to provide to honour a patient's right to informed decision-making is a fascinating question, but not the key issue here. Instead, the decision-making capacity gap that is described in this paragraph implies AI outputs are, in fact, *decisions* without justifications and, thus, override a patient's autonomy (Quinn et al., 2022).

The preceding knowledge exchange, responsibility and informed decision-making gaps raise tough questions regarding integrating AI into patient care. Should AI be used in patient care despite these gaps? One must address this question since it directly impacts a clinician's obligation of good clinical practice and a patient's right to self-determination.

Two emerging normative concepts – epistemic (in)justice and social identity – may offer new and valuable insight into whether AI should be used in patient care. Epistemic (in)justice can be described as "a wrong done to someone specifically in their capacity as a knower" (Fricker, 2007). These wrongs may also be expressed through the discriminate acquisition and transmission of knowledge. At the centre of

AI technology is the processing and output of knowledge, and critical thinking around these sources of knowledge needs to be done to fully assess the extent to which the patient and their value systems are considered relevant in the use of AI in their care. Greater insight into this can be achieved by considering epistemic (in)justice in the scholarship of decolonial scholars. Decolonial scholars are individuals who challenge any hegemony or silencing and adopt practices, theories and pedagogies that decentralise oppressive colonial power. The work of decolonial scholars is critical to ensuring academic diversity and giving agency to previously underrepresented groups. That is, when considering the use of AI in patient care, not only should the sources of knowledge used be considered, but they should also be inclusive and fair.

Social identity, another emerging concept, reverberates that people are socially embedded within their communities and environment. Drawing on the scholarship of relational autonomists, it highlights the importance of social systems in the development of personal autonomy. These social systems are driven by values, shared beliefs, and responsibility to more than just one's self, and influence the health decisions of patients and the care that they are open to receiving. Social identity is an imperative part of patient care and holistic management. As such, deep consideration needs to be given to whether opaque systems place any value on it (social identity) in its knowledge production.

In the subsequent section, this report describes various normative positions and their underlying principles with regard to the use of AI in patient care in light of the Black Box problem.

Literature Overview and Critique

Should the Black Box problem prevent the application of AI in patient care? In this section, existing normative positions on the use of Black Box AI in patient care are described.

One justification *against* AI use for patient care in light of the Black Box problem is what Suzanne Kawamleh (2022) calls the *explainability imperative,* which states that AI's ethical application in *safety-critical* sectors such as medicine requires explainability. That is, the 'how' and 'why' a machine came to a certain outcome needs to be apparent. Evidently, the explainability argument is imperative to foster an individual's right to self-determination grounded in the principle of autonomy. In order to fulfil the explainability imperative, some scholars call for explainable models of AI as alternatives to Black Box AI, such as like White Box or Grey Box AI. A White Box AI model is one in which the internal processes are transparent, therefore allowing the user to understand how the machine came to a certain decision, and a Grey Box AI model is one that combines the technology of both the Black Box and White Box models (Pintelas et al., 2020). However, while White Box AI may be interpretable, it lacks the accuracy expected from the opaque Black Box models (Pintelas et al., 2020). This would raise the question of whether accuracy should be sacrificed for transparency, particularly in a high-stake industry such as healthcare.

By contrast, propositions in favour of the use of AI in patient care are partly supported by empirical validation, whereby AI accuracy is measured and validated through scientific analysis. These propositions, which draw on empirical validation, appear to be informed by consequentialism. Precisely, emphasis is often placed on what AI can

do and how that may ultimately benefit many rather than on the AI's internal processing. For example, if the logic of the AI output is sound and the results are reproducible, then this would be considered sufficient justification to warrant its use.

Against this background, Alex London (2019) argued that opaque decisions are common in medicine. According to London (2019), explainability and examinability of how results came about are less important than the machine's ability to produce useful/accurate results. Juan Durán and colleagues (2018) echoed similar thoughts in their paper on epistemic opacity, proposing a framework that allows for trust in computer simulations. They called it c*omputational reliabilism,* which states that the results of a simulation can be justified if it is produced through a reliable process that generally produces valid results (Durán & Formanek, 2018). Durán revisits this with Jongsma in a different paper where they argue that *computational reliabilism* "offers the right epistemic conditions" for trust in medical AI (Durán & Jongsma, 2021). Durán and Jongsma (2021) also claim that Black Box algorithms may be deemed reliable if the clinician operating that technology can justify the nominal beneficial results.

Other studies have explored the implications of AI's opaqueness for informed consent. Respect for autonomy appears to be the basis for these claims. This has been a grey discussion area since these contributions do not expressly preclude the use of AI in patient care. Instead, these contributions justify the need for alternative forms of consent to bridge the knowledge exchange, responsibility, and informed decision-making gaps. In these works, two divergent ideas of informed consent have been explored concerning AI in patient care. On the one hand, some scholars explore

whether and how opacity limits a patient's ability to give comprehensive informed consent (Astromskė et al., 2021). On the other hand, other studies interrogate whether disclosure of the *use of medical AI* for medical decision-making during the informed consent process is sufficient to honour the patient's right to self-determination. For example, Glenn Cohen (2020) considered the normative construe of assessing the actions of a "reasonable" person in relation to the United States (US) Case Law on Informed Consent with regard to the obligation to disclose when using medical AI. Cohen (2020) further measured opaque AI systems against the "Transparency Standard" described by Howard Brody and concluded that even if the use of opaque medical AI was disclosed to a patient, it was unlikely to meet the standard proposed by Brody, which requires that it be discussed "*out loud in language understandable to a patient."* Moreover, if the requirement for informed consent is to disclose how and why an opaque medical AI came to a specific medical decision, then opaque medical AI would essentially violate this normative standard by default (Cohen, 2020), as it would not be possible with opaque systems.

Amann and colleagues (2020) have also explored how the use of opaque medical AI may jeopardize informed consent and acknowledged that an absence of ethical agreement exists on whether it is necessary to disclose the use of AI in the decision-making process. Nonetheless, they still argue that a failure to do so may result in a hindsight demand for justification, which may not always be possible in the case of opaque systems (Amann et al., 2020).

Various accounts of trust have also been described in relation to the application of Black Box AI in clinical care, both as a barrier and/or as a merely perceived barrier to

informed consent. These views lean towards motivation-based and/or virtue ethicist-based accounts of trust, focusing on *encapsulated interests* (Hardin, 2002) and intent of *goodwill* (Baier, 1986). In his argument, Joshua Hatherley (2020) highlighted that trusting relationships can only be formed between “beings with a form of agency.” Yet, the Black Box problem’s responsibility gap implies that agency is undermined. Thus, trust in an AI only amounts to reliability as it lacks the right motivation to act in the patient’s interest and cannot shoulder normative obligations. Mark Ryan (2020) similarly argued that AI does not have the capacity to be trusted when measured against affective and normative accounts of trust. That is, effectively, AI lacks the ability to be “moved” by the trust placed in them that would motivate them to act freely and out of goodwill (Ryan, 2020). Normatively, AI lacks the ability to take moral responsibility for the task it has been trusted with and, therefore, is incapable of being trusted (Ryan, 2020).

Andrea Ferrario and colleagues (2021) differed in their approach to trust in AI and stated that trust in medical AI does need to be defined by qualities that only humans can have. Instead, through repeated use of medical AI, the clinician will develop certain beliefs on the capabilities and margin of error expected from the AI, to trust its performance at a subsequent use without needing to update their beliefs on its trustworthiness (Ferrario et al., 2021). This view focuses on validation as a basis of trust instead of motivation.

Other normative issues that have been identified include the responsibility gap (Afnan et al., 2021) owing to the lack of accountability of Black Box AI. The addition of explicability as a fifth bio-medical principle in addition to the existing four has also been

explored as an avenue to refine AI ethics (Ursin et al., 2022). However, the existing four principles – autonomy, beneficence, non-maleficence, and justice - were found to cover all bases that render the proposed addition of a fifth principle negligible (Ursin et al., 2022).

The current arguments on whether AI should be used in light of the Black Box problem, although valid, are influenced by individualistic thinking reflective of the West. Some pertinent gaps in knowledge still exist. Little work, *if any,* has been done on two emerging normative concepts that need to be considered concerning the use of Black Box AI in patient care. They include epistemic (in)justice in the work of decolonial scholars and the relational autonomists' thinking about social identity. Both of these concepts have been briefly described in the background. Notably, suppose Black Box AI creates knowledge, responsibility, and decision-making gaps. In that case, what do the moral norms arising from the nuanced accounts of epistemic (in)justice and social identity imply for *whether* and *how* AI ought to be used in light of the Black Box problem? What valuable insights do these moral norms give rise to concerning how Black Box AI is integrated into patient care? What knowledge production regarding how best to foster meaningful healthcare experiences for patients can be inferred from these nuanced accounts? What do the moral norms arising from these accounts imply for fostering a patient's values and preferences, given the Black Box problem?

This project draws on a combination of the underexplored works on epistemic (in)justice and social identity by decolonial scholars and relational autonomists to address these gaps. AI use in patient care has an overarching human component, and deep consideration needs to be given to this human aspect when assessing whether

AI should be used in light of its Black Box problem or it runs the risk of prioritising scientific accuracy over human value systems. Epistemic (in)justice and social identity offer valuable insights into the human aspect of AI use, and drawing on decolonial scholars is critical to transformation and inclusivity. It is also intrinsically valuable to interrogate these questions from these under-used perspectives to foster diverse academic scholarship on the ethical discourse of the AI Black Box problem in patient care.

Research Question

The project's specific question is: What do the moral norms that arise from epistemic (in)justice and social identity grounded in the scholarship of decolonial scholars and relational autonomists imply for whether AI ought to be used in patient care in light of the Black Box problem?

Rationale for the Study

This project is being undertaken with the dual purpose of addressing an identified research gap and contributing to the advancement of knowledge in this area. This is particularly important given the current relevance of AI and its increased uptake in healthcare. Notably, the view that writings of decolonial scholars and relational autonomists can be combined to yield valuable insights into how we use medical AI in patient care in light of the Black Box problem has not been explored previously, or at least to a significant degree. Reasonably, this should be of interest to academic scholars.

There is also instrumental value in undertaking this project. Precisely, the project is socially valuable. Integrating AI into patient care in light of the Black Box problem is complex, with many social implications. Exploring ethical concepts, which have received little attention, will allow this project to provide insight into creating a more equitable, just, and sustainable system for integrating AI into society. Furthermore, this project will raise critical insight into the social components that influence the implementation of AI into healthcare beyond just empirical results. These, in turn, contribute to the project's scientific and health value. Through the contribution of knowledge, this project can help shape policies for AI use in patient care and influence decision-making processes to guide clinicians on whether and how to integrate medical AI into their practice.

Thesis

Drawing on the moral norms arising from a combination of epistemic (in)justice in the scholarship of decolonial scholars and social identity in the scholarship of relational autonomists, this report contends that AI use in patient care in light of the Black Box problem is deeply problematic and is ethically impermissible. However, there may still be value in its use outside of patient care.

Research Aim

The primary aim of the project is to defend the thesis.

Research Objective

(1) To outline the relevant moral norms that can arise from the writings on epistemic (in)justice in the work of decolonial scholars, and social identity in the scholarship of relational autonomists.

(2) To draw on the moral norms outlined in *objective one* to interrogate whether AI ought to be integrated into patient care in light of its Black Box problem.

(3) To address potential objections to this project's thesis and/or the justifications supporting the thesis.

Research Design and Methods

This research report is a Type I project. Precisely, it is a primarily conceptual and normative bioethical project. As a result, the research method will primarily be desktop-based research, and relevant literature will be derived from databases such as Google Scholar, Elsevier, PubMed, and the Wits Library database. This may include, but is not limited to, journal articles, dissertations, books, and web-based articles.

Argumentative Strategy

To realise objective one, I will first give a comprehensive overview of epistemic (in)justice and social identity as the underlying concepts that will be used in my thesis defence. I will draw on the works of decolonial scholars and relational autonomists to provide nuanced accounts of these concepts. Once I have broadly outlined these ethical concepts, I will outline the relevant moral norms (for thinking about the medical AI Black Box problem) that arise from a combination of these concepts, such as inclusivity of sources of knowledge and respect for people and their values.

To realise objective two, the project draws on these moral norms (objective one) to argue that AI use in patient care in light of the Black Box problem is deeply problematic. I will show that three things regarding AI data need to be considered when making an allowance for its use in patient care – the source of knowledge for the training data, how that data is processed and whether the AI can provide sufficient information regarding its output data for informed consent to be obtained. Epistemic (in)justice in the works of decolonial scholars calls for *de-silencing* and *inclusivity.* Sources of knowledge used for AI would require extensive restructuring to achieve this, and to bridge the gap between credibility granted to scholars with scientific knowledge and patients who have the best knowledge of themselves as the recipients of AI-influenced care.

I will proceed to show that a patient's social identity, values, and choices are relational to their contextual existence. That is, relationality enhances autonomy rather than inhibits it, as humans are socially embedded creatures who, by nature, are influenced to think about the world in a way that the world was taught to them. Therefore, social identity becomes a key influencing factor in the decision-making process, and Black Box AI undermines this through its opaque nature. It also fails to guarantee that the values of patients are being taken into consideration in the processing of data by the AI. Furthermore, this opaque nature acts as a barrier to ensuring the inclusivity of various groups and their nuances. Opaque systems also cannot provide adequate information regarding the AI output for a clinician to obtain informed consent, forcing patients to make important health decisions on sub-minimal information. These feed into the gaps created by AI – knowledge exchange, responsibility and informed

decision-making gap. I will also show the added perplexity of introducing advanced technology into disadvantaged communities who have little knowledge of it and how this will affect their social experience, including the risk of AI potentiating bias.

To realise objective 3, the project addresses potential criticisms. For example, one might argue that with the emergence of less opaque systems, such as explainable AI and White Box AI, the Black Box problem becomes irrelevant. In response, I demonstrate how these new AI formats are subject to complex proprietary patents, which may leave the clinician and patient in a metaphorical Black Box space with AI creators.

Furthermore, a critic may point out that my decision to draw on two concepts – epistemic (in)justice and social identity – from different scholarships is ill-advised since it undermines my capacity to think deeply about the concepts to determine their implications for AI use in light of its Black Box problem. In response, I demonstrate how this is not problematic since the primary objective of the project is, in fact, *evaluative* – to defend the usefulness of these concepts for AI use in patient care – rather than *descriptive* – that is, how they are understood. The project provides further explanation.

Research Outcomes and Outputs

This project is a key part of the requirement for attaining my master's degree in Bioethics and Health Law at the University of Witwatersrand. Equally, the project aims for at least one publication in a Bioethics journal. Notably, I have now submitted a

manuscript to the journal *Medicine, Healthcare and Philosophy,* and the manuscript may appear as a published print before the examiner examines this project. Finally, I would also endeavour to share insights from this project at a scientific meeting.

Limitations

This study has limitations. One helpful way to address the gaps created by the medical AI Black Box problem is to ask clinicians and patients if they oppose its use, given these gaps. Yet, this project is a primarily conceptual project that broadly explores the implications of scholarships on decoloniality and relational autonomy for AI use in light of the Black Box problem. This implies that the project's outcome is neither generalisable nor necessarily representative of the views of clinicians and patients. Empirical studies and multiple perspectives are still required to contribute a more robust and comprehensive insight into whether and how AI should be used for patient care in light of the Black Box problem. Furthermore, AI in healthcare is also still relatively new, although rapidly developing. As such, it is difficult to predict what other gaps Black Box AI in patient care may create following further advancement in this technology to address them in this project fully.

Overview of Chapters

1. General Introduction
2. What are the moral imperatives of epistemic (in)justice and social identity?
3. The moral imperatives of epistemic (in)justice and social identity for medical Black Box AI use in patient care
4. Addressing potential objections

5. Conclusion
6. References

Ethical Approval

This mostly conceptual study does not involve research with humans and animals. Against this background, I have not applied for ethical approval from an Ethics Committee.

CHAPTER TWO

WHAT ARE THE MORAL IMPERATIVES OF EPISTEMIC (IN)JUSTICE AND SOCIAL IDENTITY?

Introduction

In the previous section, I articulated three gaps that the Black Box AI creates: responsibility, knowledge exchange and informed decision-making gaps. These gaps have important implications for clinician-patient relationships. Briefly, there are two important stakeholders in a clinician-patient relationship. On the one hand, there is a patient, who seeks medical help for an ailment or concern. The patient is the recipient of the medical care. On the other hand, there is a clinician, who provides medical assistance to a patient when they can. The encounter between the clinician and patient, often within a clinical setting, gives rise to a *contractual* clinician-patient relationship (Solberg & Steinsbekk, 2012). Yet, a clinician-patient relationship is not only defined by this encounter. Magda Slabbert and Melodie Labuschaigne (2022) add that this contract must be *consensual*. This does not need to be formally expressed. Specifically, as Valarie Blake (2012) remarks, “once the [clinician] enters into a relationship with a patient [by either examining, diagnosing, treating, or agreeing to do so]….*a legal contract* is formed in which the [clinician] owes a duty to that patient to continue to treat or properly terminate the relationship.”

Suppose the clinician-patient relationship ought to be consensual. In that case, it ought to satisfy key elements of consent. Notably, the dynamics of *contemporary* clinical contexts often require consent to be informed in order to move the clinician-patient relationship from a paternalistic model that thrived in previous centuries to one that

enhances the patient's decision-making capacity (Hall et al., 2012). Most medical guidelines often specify disclosure of material information, comprehension, voluntariness and competence as core elements of (informed) consent. Disclosure of material information requires that all information that a reasonable person would consider significant in influencing their decision ought to be shared with the patient. Comprehension implies that the patient is able to understand the significance of this material information. Voluntariness means that the patient is happy to proceed, given this material information. Finally, competence implies that the patient making this decision is of a sound mind and a mature age to be allowed to make such a decision.

Suppose an AI suggests that a patient with a terminal illness such as cancer has a poor prognosis and that the best management tool for this patient is palliative care. To fulfil informed consent conditions, a clinician would be required to provide information on how the AI determined that the prognosis is poor and why no further steps are recommended in the patient's management plan. However, the AI Black Box problem has many implications for the requirement to disclose material information to honour the consensual nature of the clinician-patient relationship.

The identified gaps also present a problem with two other factors that influence informed consent. That being the commercialisation and social media hype around AI. With the increasing availability of information online, AI exposure and popularity is constantly growing. These present a challenge for assessing voluntariness. Is the patient's choice being unduly influenced by what they have been led to believe about AI through positive media feedback and online presence, or is the patient's decision

self-determined? Without measures in place to address this, the voluntariness of informed consent would be undermined.

These gaps also pose a challenge to the understanding component of informed consent. The word ‘understand’ itself implies comprehensive knowledge. In light of the knowledge exchange gap, if the clinician, and subsequently the patient, cannot understand, then understanding cannot exist, and without it, informed consent cannot be valid. Could the clinician-patient relationship be consensual in light of this gap?

Currently, AI algorithms are used as supporting tools in clinical contexts, such as interpretation of radiology scans and remote patient monitoring. Yet, the full integration of AI in clinical care is anticipated to occur by 2035, as stated in the previous section. Suppose a clinician allows the medical Black Box AI to absorb their duties in the clinician-patient relationship based on empiric validation of the AI. If there is a shortfall in the validation, and malpractice has ensued, how should responsibility be impugned? Given that the clinician-patient relationship is legally binding, unto whom does the legal action fall?

In this section, I will describe clear norms that seriously consider the implications of the gaps medical Black Box AI raise for the clinician-patient relationships and the gaps it creates. To articulate these norms, I will focus on the thinking about social identity in the work of relational autonomists and epistemic (in)justice in the work of decolonial scholars. Eclectic approaches like the one I adopt in this project are not uncommon and are preferred approaches for thinking deeply about an issue that requires multiple perspectives, as is the case with the medical Black Box AI. Moreover, these moral

approaches offer a perspective beyond simply what the machine is capable of, but also on what the capability of the machines means for the people who are intended to use and benefit from it. It addresses the human problem beyond the science of AI.

Epistemic (in)justice and Decolonised Scholars

In this section, I outline what decolonial scholarship is and the moral norms that arise from epistemic (in)justice in the work of decolonial scholars. Decolonial scholarship is vast, complex and generates many contestations concerning its moral imperatives, goals, and participants. For example, the targets of decolonisation are contested. In this regard, decolonisation is sometimes conceptualised as an anti-western, anti-colonial quest to reclaim Africa's agency and her intellectual, economic, infrastructural, political, and social freedom by shifting the continent to the post-colonial (or post-neocolonial) era. This is the way Ademola Fayemi and Macaulay Adeyelure (2016) describe decolonisation, that is, as the "process of self-critical awareness of foreseeing, discovering and avoiding hegemonic institutionalisation as well as mental colonisation of concepts and disciplines in contemporary *African scholarship."* Conceptualised this way, decolonisation discourse becomes an exclusive reserve of African scholars, walling off the discourse from non-Africans.

Notice that the positionality from, or the context in which this discussion is had, also has implications for the moral imperative of decolonisation. The moral imperative of decolonisation is the norm regarding *what we ought to do to decolonise*. In African health decolonial literature, the moral imperative of decolonisation entails grounding health and health research discourses that have significant implications for Africa in dominant knowledge systems on the continent. Furthermore, at the health and health

research curriculum level, this imperative entails giving greater recognition to African knowledge systems in developing the continent's health and health research curricula. This will also mean that African scholars are afforded primary roles in African health and health research. In the global health discourses, African knowledge systems will have equal status as other knowledge systems in the knowledge production that underlies key international health policies or feeds into global health discourses. At the health journals' level, the contributions and perspectives of African scholars will have equal prominence in journal publications on global issues. This list is not exhaustive. Nonetheless, what I wish to outline here is that foregrounded this way, decolonisation requires recognising, inviting, and granting agency to Africa.

Decolonisation is also sometimes described as a quest to reclaim colonised persons everywhere. In this regard, decolonised scholars, as such, are individuals everywhere who challenge any hegemony or silencing and adopt practices, theories and pedagogies that decentralise oppressive colonial power. For this reason, Rianna Oelofsen (2015) describes decolonisation as "the change that colonised countries go through when they become politically independent from their former colonisers." Nelson Maldonado-Torres (2016) equally echoes a similar view in his description of colonisation and describes decolonisation as "key terms for movements that challenge the predominant racial, sexist, homo- and trans-phobic conservative, liberal, and neoliberal politics of today." Conceptualised this way, decolonisation is placed in the contest of the human quest for liberation from oppression and domination everywhere. On this account, the moral imperative of decolonisation will include de-silencing voices, acknowledging, inviting, and enhancing the agencies of individuals in discourses and issues that have significant implications for them and global issues.

The goal of the preceding paragraphs is not to defend the right way to conceptualise decolonisation conclusively. Instead, I highlight the relevant aspects of decolonisation for my evaluative task in this project. Nonetheless, it is still worth emphasising – as demonstrated by different publications – that decolonisation is commonly accepted as the effort to emancipate previously colonised nations, persons, and communities from the generational impacts of colonial thought, which continue to govern social, political, intellectual and economic structures (Bua & Sahi, 2022; Nordling, 2018; Von Bismarck, 2012). In other words, although scholars conceptualise decolonisation differently, the central point is that it entails undoing marginalisation, de-silencing certain groups, and the domino effect it has on building and progressing inclusive societies. Framed this way, epistemic (in)justice in decolonial scholarship thus becomes a quest to decentralise knowledge production by giving equal recognition to various knowledge systems and knowledge producers on global issues.

Notably, this section draws on the thinking about epistemic (in)justice, primarily – although not exclusively – in the scholarship of decolonised scholars like Miranda Fricker, Frantz Fanon, Nelson Maldonado-Torres, Olufemi Táíwò and Linda Tuhiwai Smith. Miranda Fricker (2007) described epistemic (in)justice as “a wrong done to someone specifically in their capacity as a knower.” Two primary forms of epistemic (in)justice exist, i.e., testimonial and hermeneutical injustice (Fricker, 2007). Testimonial injustice occurs owing to a credibility deficit that is influenced by an identity prejudice (Fricker, 2007). For example, a female academic whose testament holds less value than her male counterpart owing to gender prejudice. On the other hand, hermeneutical injustice occurs because a person cannot express an important aspect

of their social experience due to hermeneutical marginalisation (Fricker, 2007). For example, oppressed people often lack the necessary resources to make sense of their oppression, as was likely the case during the era of slavery.

The reader should observe that both the testimonial and hermeneutic injustices create i) a disadvantage condition since it disadvantages the knower in relation to others, whose thoughts are given serious consideration in the knowledge production, and ii) a prejudice condition, since such exclusion arises from biases against the speaker based on their say, race or gender. Since Miranda Fricker conceptualised the term, other conditions have also been added: i) stakeholder condition (suppose the excluded knower has a stake in the outcome from which they are excluded because of biases against them), ii) epistemic condition (the excluded knower does, in fact, possess the relevant knowledge), and iii) social justice condition (that is, the bias against the knower is indeed connected to larger structural injustices like racism) (Byskov, 2021).

In light of the above, the moral imperative of epistemic justice is *de-silencing* or *inclusion* that respects the agency of others is substantive and transformative. Notably, epistemic (in)justice in the decolonial narrative brings to the fore the necessity of including the voices and knowledge systems that were previously dismissed based on identity prejudices and marginalisation, enabling credibility that previously did not exist, not for lack of knowledge, but for lack of affirmative presence, and understanding how silencing or domination occurs.

Silencing is complex and occurs in various ways not limited to mythologising Western superiority. The preceding reaffirms the view that the West is the only legitimate and civilised source of knowledge (Smith, 1999). Silencing could also occur through cognitive injustice, which is the systematic exclusion of the values and beliefs of those dominated in the knowledge production that feeds into many practices. For example, to mitigate the effect, colonial education was used as a tool to create “indigenous elites”, on the presumption that these elites would align their interests with colonisers (Smith, 1999). This effectively resulted in a dismissal of knowledge and culture that represented where they came from.

This form of cognitive injustice is worth reflecting on, even if briefly. Frantz Fanon, in his work, also reflected on this supposed elite as a native intellectual. He aptly described the role of the native intellectual who, to experience the culture of the oppressors, pledges “his adoption of the forms of thought of the colonialist bourgeoisies”, leading him to mindlessly accept the word of the oppressors as good judgement (Fanon, 1963). Similarly, in his essay, Olufemi Táíwò (2020) described how inviting an individual who somewhat can be associated with a certain group of people is not a true reflection of those people, as the person invited into the room already has a social advantage. Knowledge about the people can only truly come from the people. Táíwò (2020) calls this *elite capture*. Specifically, elite capture is “the control over the political agendas and resources by a group’s most advantaged people” (Táíwò, 2020).

These views echo the recurring sentiments of knowledge of a large part of the population being diminished beneath the towering dominant group, with the invaluable essence of who they are being lost in history. To undo the harm of silencing and foster

a more transformative inclusion, all people, irrespective of who they are or where they come from, ought to be equally included in the development, progression, and distribution of knowledge, especially if they have a stake, or may be significantly impacted by the outcomes that are informed by such knowledge production. This means additional focus needs to be placed on marginalised groups, who have thus far not received the necessary credibility for their contribution to knowledge production. They also often lack the resources to receive knowledge that may be valuable to them in making sense of their social experiences.

Finally, it calls for fair distribution of power and resources. To understand how, the reader will recall that epistemic (in)justice is often connected to larger structural injustices (or social justice condition). Social disadvantage is a generational trauma and may influence identity prejudice. As Táíwò (2020) stated in an essay, the goal (of epistemic justice) is to build (new) rooms where people can sit together rather than limit themselves to navigating around rooms that history has made. Suppose power is what it takes to be in a room where influence occurs. In that case, power should be distributed to all, not merely by representation of marginalised groups in the room but by including the marginalised in developing standards for re/distributing power.

Social Identity and Relational Autonomists

This section outlines the moral norms that can arise from the thinking about social identity grounded in the scholarships of relational autonomists. Autonomy has long been an integral part of ethics and philosophy, with depictions tracing back to ancient Greek Medicine and later to the philosopher Immanuel Kant. In 1979, it reached its pinnacle when (Beauchamp & Childress, 1979) added “Respect for autonomy” as one

of the four principles in their book *Principles of Biomedical Ethics,* which became a foundation upon which modern ethics of medicine would be built. In their book, autonomy is conceptualised as the right to self-determination that often necessitates moral rules like the requirement of informed consent and respect for one's privacy. Jukka Varelius (2006) echoes a similar view in their description of autonomy as self-governance and further describes personal autonomy as "self-rule that is free from both controlling interference by others and from limitations, such as inadequate understanding that prevents meaningful choice." This is in keeping with the traditional concept of autonomy, which is aligned with Western Individualism.

Autonomy has long progressed as a "hyper-individualism" concept informed by more Western or individualistic modes of experiencing the world. The reader should observe that the preceding does not imply that forms of communitarianism cannot be found in the West; in the same way, elements of individualism are also present in the Global South, often acclaimed to be more communitarian.

The hyper-individual model of autonomy was progressed to overcome paternalism and the undue influence that authoritative figures could place on decision-making. However, it problematically ignored "values [like] mutual responsibility, cooperation and care towards others" (Dove et al., 2017). It assumes an impoverished view of humans as isolated beings unbounded by any roots.

In recent years, however, hyper-individualism has been challenged by feminists, communitarians, and identity politics theorists (Christman, 2004). For example, Jennifer Nedelskyt (1989), contended that people's social and political relations impact

how their predispositions, interests, and autonomy develop. These ultimately affect how people process information and use it to make crucial decisions – such as those that can be expected in healthcare and research. Through such discussions, a new concept of autonomy has emerged, namely relational autonomy. Prominent scholars contributing to this relational autonomy concept are Jennifer Nedelskyt, Catriona Mackenzie, Natalie Stoljar, John Christman and Marina Oshana.

Catriona Mackenzie and Natalie Stoljar (2000) have described relational autonomy as an umbrella term that encompasses similar perspectives that are based on the shared premise that people are socially embedded and that a person's identity manifests through their relationships and shared social factors. In other words, people understand who they are based on contextual factors such as religious affiliations, family values/connections, and community normative practices. As Annette Baier (1985) claims, people are actually "second persons", who succeed each other, and their personalities are uncovered through their relations with others.

Whereas autonomy has affirmatively been a process that discourages any external influence in fear of simulating paternalism, which modern medicine has stepped away from, relational autonomy moves away from this stringency to a model that encourages engagement and shared decision-making. It suggests that involving others in important decisions does not conflict with being autonomous. Rather, this affirms the social nature of the individual (Walter & Ross, 2014).

The preceding gives rise to the concept of social identity. Notably, social identity is the sense of self that develops through belonging to a group (Mcleod, 2023). It (social

identity) is a recurring theme among various relational autonomists. Equally, it is a key influence in how people choose to make their decisions.

However, the influence of the social component in relationality is not consistent among relational autonomists, and two broad categories of philosophical thinking have been noted. These are “causally relational” and “constitutively relational” accounts (Baumann, 2008). *Causally relational accounts* put forward that social conditions act as background conditions for realising autonomy without changing the definition of autonomy itself (Baumann, 2008). In contrast, *constitutively relational accounts* believe that the connection between a person’s autonomy and social environment is more fundamentally entwined with “social” forming part of the definition (Baumann, 2008). The overarching theme, however, among relational autonomists is that people do not exist in isolation and are a product of their social environment. A key difference lies in how much weight the social aspect plays in defining autonomy.

Suppose social Identity in the works of relational autonomists implies that an individual’s web of connections and relations plays a crucial role in self-conception and understanding. In that case, one ought to honour individuals’ relations. Individuals are not just *isolated, independent* entities but exist in relations and inter-relations. These relations define and redefine one’s identity as *social.* Notably, social identity is that certain individuals understand their place in the world as related and interrelated (Kajee, 2010). Fragmenting the individual from their social reality, relations or embeddedness gives rise to a crisis of identity and is dislocating (or silencing). Social identity is a key part of meaning-making. Social identity also implies that the weight of decisions should be shared with relevant people who can assist in helping to make

life-altering decisions. Families are a crucial part of the health decision-making process, as even once the patient enters professional care, the family plays a vital role in the care and recovery process (Ho, 2008). Ho (2008) describes healthcare decisions as not “simply individual affairs” but rather as “family events.” Seeking guidance and support in making decisions is not a threat to autonomy. Instead, it is a means to allow autonomy to flourish (Walter & Ross, 2014). It considers the emotional weight of decisions as equally important to that of reason (Walter & Ross, 2014). It also considers the cultural values of joint decision-making that are prominent in minority groups. In Western medicine, minority groups often are faced with challenges such as language and cultural barriers, discrimination, and lack of access to healthcare coverage, creating a sense of isolation for such patients, as they depend on their families for advice and medical care (Ho, 2008).

Furthermore, understanding people as a product of their social self, means that people are not simply accountable to themselves but to those around them who are affected by their decision. For example, a positive result from a genetic test done at an individual level – from the relational autonomy perspective – is relevant at a familial level as a risk factor to other family members. Dove and colleagues (2017) describe a familial approach that draws on relational concepts of autonomy and “emphasizes relational values, such as mutual responsibility (as a manifestation of reciprocity), and interdependence,” to highlight circumstances in which it may be appropriate to share sensitive information. Individuals tend to have “altruistic tendencies” and generally are in favour of the familial approach (Dheensa et al., 2016), highlighting a sense of responsibility towards more than just themselves.

Concluding Remarks

This chapter has interrogated and described key moral imperatives that can be grounded in the decolonial scholarship on epistemic (in)justice and the thinking about social identity in the scholarship of relational autonomists. At first glance, epistemic (in)justice and social identity appear to be distinct moral concepts or, at most, loosely related. However, the moral norms that underlie these concepts are closely linked, and in this concluding remark, I will highlight the overreaching moral norms that arise from the combination of these concepts.

Social identity represents a fixed set of values, rules, and commitments an individual adopts based on their origins. As argued by relational autonomists, it constitutes or influences one's development of autonomy and ability to make choices regarding themselves. The choices made, in itself, are an expression of knowledge. The choices made by individuals represent the history of a person, cultural values, and community norms. These choices, at times, may not align with popular or contemporary views associated with the West but are equally valuable sources of knowledge. Conceptualising, incorporating and distributing this knowledge will be essential in ensuring inclusivity and promoting epistemic justice.

Knowledge is not only an individual prerogative but one that can be progressed through families, communities and their leaders. Inclusivity encompasses the acceptance that individuals may choose to source their knowledge from their close network as a trusted source, and use it to progress their decisions. The credibility of this knowledge should not simply be dismissed based on identity prejudices. Perhaps a community leader who influences a decision is not merely patriarchal and

paternalistic, but one experienced enough to add value to the decision. Perhaps, an individual respects that leader and does not feel pressured but reassured by their input.

Individuals and communities ought to receive a fair distribution of power and resources so that the social narrative changes to one where people are empowered with equal knowledge to further develop their autonomy. Access has long-defined social circumstances, and the lack thereof has led to a poor understanding of certain social experiences. In the face of contemporary medicine, individual autonomy increasingly poses the risk of isolation and trepidation (Ho, 2008), making the need to consider social context more critical than ever.

The moral norms underlying epistemic (in)justice and social identity in the work of relational autonomists call for inclusivity of knowledge and sources of knowledge, empowerment of autonomy through equal resources, and respect for people and their values. I will draw on this moral norm in the next chapter to interrogate i) the permissibility of integrating medical Black Box AI in patient care, ii) how AI, particularly the more advanced ones like Deep Learning algorithms, ought to be integrated into patient care, suppose this is permissible from the perspectives I draw on, and iii) outline concrete actions which ought to be taken; what changes need to occur to align AI broadly with the values and norms I articulate in this section.

CHAPTER THREE

THE MORAL IMPERATIVES OF EPISTEMIC (IN)JUSTICE AND SOCIAL IDENTITY FOR MEDICAL BLACK BOX AI USE IN PATIENT CARE

Introduction

In the previous chapter, I outlined the moral concepts of epistemic (in)justice and social identity and described the moral norms that jointly arise from these moral concepts. These norms include inclusivity of knowledge and sources of knowledge, empowerment of autonomy through equal resources, and respect for people, their values and their relations. This chapter interrogates the implications of these norms for the permissibility or impermissibility of AI use in patient care in light of the Black Box problem. Notably, given that there is an overarching human component in clinical care, the chapter defends why deep consideration ought to be given to this human aspect in assessing whether AI should be used in patient care in light of the Black Box problem. The moral norms that arise from epistemic (in)justice and social identity can enhance the thinking around this. This chapter also addresses what should happen to align Black Box AI with the moral concepts progressed in this report.

Assessing AI Permissibility Given the Black Box Problem

There are different subsets of AI that may be used for patient care. They include Machine Learning and Deep Learning. Machine learning starts with a set of training data, which is fed into the AI technology. Once the data is available, the machine uses the available data to train itself to find patterns and make predictions following what it has learnt. The learning can occur either through supervised, unsupervised, or reinforcement models. Contrarily, Deep Learning is a more advanced brain-logical

complex form of AI that has higher predictive power and the capacity to learn from un/structured large data sets, including continuous data, with *nearly* no human interference. Unlike Machine Learning, layers are hidden in Deep Learning, which makes them Black-Boxed. The analysis below has implications *primarily* for Deep Learning.

Three things regarding data are worth highlighting when considering the permissibility of using AI in clinical care in light of the Black Box problem – i) the origin of the training data put into the machine, ii) how the machine processes the data available, and iii) whether the AI output provides sufficient relevant information for an informed decision to be made. This section defends how the moral norms I articulated in the previous chapter yield the conclusion that medical Black Box AI use for *patient care* is impermissible for the following reasons: i) given that un/recognized prejudices may structure source data or operational environment, ii) medical Black Box AI internal operations are opaque, and iii) this opaqueness undermines the basic requirements for informed consent.

Sources and inclusivity of knowledge are important considerations in the origin of training data. Suppose the AI learns based on the knowledge it is provided with. In that case, it becomes imperative that data generators and creators of AI technology exhibit some level of transparency on what sources of knowledge are being used to collect and group data. What knowledge production is employed to develop standards that are used to group and collect data? As described in the previous chapter, the transcending nature of colonial knowledge continues to dominate as a primary source of knowledge in many scientific and non-scientific fields. As such, *representation* could

be misguidedly progressed through “elite capture”, which is a poor reflection of the people it is meant to be representing, and as such, would be a limiting source to use as training data.

Furthermore, suppose the machine learns from the ongoing data that it is fed, or from the environment (as is the case with Deep Learning), which would likely occur through repeated use in various clinical subsets. In that case, the patients in those clinical settings may become one source of data. Health disparities would directly impact the collection of data in this manner. Health disparities refer to “differences between groups in health insurance coverage, affordability, access to and use of care, and quality of care” (Ndugga & Artiga, 2021). Populations with poor access to and quality of basic care would also inevitably lack access to advanced technology, such as medical AI, to complement their healthcare. As such, there would be an underrepresentation of these groups of people in the data set, and health nuances specific to these people would be omitted. This could affect the accuracy of the output, given that a machine cannot consider unique factors it is not exposed to.

Some may argue that these specific circumstances might be the exception. However, many small exceptions may converge into big omissions. Advanced AI can learn from the environment or field operation in addition to the training data. Even this is problematic since certain prejudices remain in the healthcare context that continue to structure healthcare delivery in ways that disadvantage certain groups or individuals. Specifically, studies continue to show a correlation between ethnicity, health and wealth in healthcare delivery (Hague, 2019). The more advanced AI may learn this pattern and predict lower health outcomes for already disadvantaged individuals like

the poor, thus, causing limited medical resources to be withheld from them. Given the knowledge gap created by medical AI Black Box, it would be (nearly) impossible to detect when a Deep Learning algorithm has behaved unethically this way to correct it.

Secondly, even if the limitations in sources of data had to be overcome with conscientious input of training data that was fair and equitable, a further problem lies in how the Black Box internal processes use that data to produce an output or rank outputs. Suppose there is no discernible way to know what factors were considered in assimilating the outcome. Equally, suppose epistemic (in)justice and social identity as a subsidiary of relational autonomy concur that people and their values should be respected. In that case, (to enable respect) the values and personal nuances of a person need to be considered in decisions that are made about them, or in this instance, in outputs that may influence their healthcare decisions. The Black Box renders knowing this information impossible. At best, it can be assumed that if the output is accurate and in keeping with the patient's values, then the machine might have considered all contributing factors. However, "assumed" and "might" are non-definitive terms. Such output could also be purely coincidental, and the Black Box AI may not have considered factors important to the patient at all. Hidden processes are counterintuitive to the concept of respect – for people, their values, their ideals – and Black Box AI represents just that through its opaque nature.

Another problem that the Black Box presents is the amplification of bias over an extended period of time. This may result from the intentional or unintentional introduction of biased training data. Suppose a machine that has not been calibrated recently is used to measure the vital signs of patients, and faulty data is input into the

machine over a period of time, the machine may incorrectly define a new normal based on the data it was fed. Or suppose stereotypical assumptions about groups of patients (for example, patients > 60 years have decreased renal function) are used to create the dataset that will be used to train the machine. The machine will eventually produce all outputs based on these stereotypical assumptions. This could radically alter the algorithm and skew all future outputs to favour this biased data. The Black Box hinders control over the exponentiation of the bias since the system used to process the information is not controlled by the creators nor understood. Therefore, the output produced may be evidently biased or may potentiate existing real-world biases, which adversely affects the drive for inclusivity. Quality control of data sets may assist in minimizing the potential for bias. However, the threat cannot be entirely eradicated without the ability to quality control the processes as with Black Box AI.

Ultimately, medical AI is only as good as its training data and the context it functions or learns from. Even if full control of data is achieved, what the machine does with data will still be a mystery – one where the opaqueness neither promises inclusivity of knowledge nor respect for people and their values. Nevertheless, even if it did have a higher awareness of such concepts, the Black Box conceptually makes it impossible to know or understand what more needs to be done to ensure that individuals' social relations form the basis of AI predictions.

Thirdly, there is a margin for concern about what should be done with the AI output once it has been processed. In a *patient healthcare* setting, that output could range from organizing patient files to suggesting a potential treatment or prognosis. However, in the *healthcare delivery* context, it could range from triaging emergency

response or care services to planning clinical rotations. Nevertheless, when directly linked to patient care, patients are usually expected to make an informed healthcare decision based on the output produced by the Black Box AI. This is problematic as the word “informed” suggests that sufficient information is required to be shared with the patient regarding how a certain outcome or decision was reached, so any decision taken based on it would be with good rationale. Given the nature of the Black Box, the clinician using the AI would be unable to relate such information. Even if the clinician had to seek clarification from the creator, it would surmount to little as the Black Box can neither be opened nor understood. This would result in the minimum requirement for informed consent not being met, and any decision taken based on this output would merely be consent. Patients’ autonomy ought to be empowered through the distribution of equal resources and/or information. Placing patients in a position where they are forced to make decisions about their health based on sub-minimal information goes against this principle and may lead to a breach of trust or even harm.

Additionally, there is the social self that exists within a network of connections. As such, patient care is not an individual prerogative, and any decision made may affect several closely linked connections. Can AI conceptualize the complexity of social self in processing information to suitably suggest patient care interventions that are beneficial on an individual level and holistically in keeping with the patient’s family dynamic? Given the nature of the Black Box, is it even possible to know? Beyond that, suppose that the patient in question is unconscious or incapacitated, and a proxy is appointed to decide on the patient’s behalf. The proxy would be faced with the dilemma of deciding for someone else based on essentially opaque reasoning. For example, a situation may arise where the AI has determined that the patient should

be taken off end-of-life care. Therefore, the family is requested to agree to it based on the scientific accuracy that is expected from the AI. Accuracy alone may not be sufficient justification for the family, and a further explanation would be warranted. However, the provided explanation would likely be limited to the interpretation of how the clinician “thinks” the AI came to that suggestion.

Furthermore, introducing Black Box AI into community healthcare programs in socially disadvantaged and rural communities may have added perplexities. Suppose a community with little to no contextual background knowledge of AI technology is now informed that a machine will be processing their information and assisting with their healthcare. Having had no previous exposure to such advanced technology, they will lack the ability to fully make sense of this new social and healthcare experience. Two things may result. They could either overestimate its potential and be in blind favour of the AI-suggested output. Or they could implicitly refuse the added dynamic to their care based on mistrust of technology that they are not familiar with. This would further contribute to a lack of representation within the training data set and/or healthcare context, which continues to be structured around prejudices and histories of imperialism. It would also raise the question of whether the AI was built to consider minorities, and given the nature of the Black Box, it would be difficult to tell. That said, the existing output may already be biased given that the training data and the environment would be reflective of the socially advantaged standpoints or prejudices where it existed before its introduction into underprivileged communities.

This feeds into the gaps that AI creates in light of the Black Box problem – knowledge exchange, responsibility, and informed decision-making gaps. The lack of

transparency of creators in what training data is used, the inability of the creators to explain how the Black Box works and the inability of the clinician to explain why the AI output suggested a certain intervention/diagnosis/prognosis creates a three-fold gap in knowledge. The lack of higher awareness in a machine to take responsibility for "wrong" decisions, concurrent with the creators' and clinicians' lack of insight into how the decisions were made, creates a gap in who ultimately takes responsibility. The unavailability of information makes the informed consent process negligible, and what is not known cannot be shared. These gaps create a barrier to fostering patient rights through inclusivity, empowerment of autonomy and respect for persons and their values.

Given the moral norms arising from epistemic (in)justice and social identity, the above depicts that using AI *in patient care* in light of the Black Box problem is deeply problematic, and consequently ethically impermissible.

What Needs to Happen?

A critic may be correct to observe here that this project will not have the impact it intends in the AI community since it has not, at least to a significant degree, recommended concrete steps for aligning AI broadly with cherished values in the scholarship on decolonial and relational autonomy literature (or the concrete suggestions for addressing the gaps medical Black Box AI creates). This section interrogates this question. Specifically, it reflects on what concretely needs to happen to address the Black Box AI problem broadly: at the level of the knowledge production that informs how standards are developed for structuring data collection, gathering or training for AI; at the level of AI developers/creators since they may also have

un/recognised prejudices that can feed into their creation; at the level of AI users like clinicians that implement AI predictions or the hospital administration that recommend AI use; at the level of the potential beneficiaries of AI for whom AI is used to optimize care; and at the level of policymakers and institutions who have the responsibility of safeguarding human welfare.

Ultimately, what needs to happen to address the gaps that medical Black Box AI creates and ensure that AI use promotes the moral norms this project draws on is that AI, including the advanced ones like Deep Learning that create the Black Box problem, must become interpretable and explainable. However, research on more advanced AI interpretability has not yielded material success. In fact, as Tilman Räuker and colleagues (2022) have confirmed, current research on AI interpretability has been mostly unproductive. Under this assumption, what more could be done? What concretely needs to change to foster cherished human values?

First, inclusivity of knowledge production requires that both skilled professionals and the common man be included in the development of standards so that the value of human experience is not overshadowed by scientific superiority, given that both are valuable sources of knowledge. It also allows for the gap to be bridged between those who fall on either extreme of the AI spectrum (from the most knowledgeable to the least knowledgeable), as their experience of the same technology will be significantly different. Furthermore, in order to empower groups of people to appreciate the significance of AI technology and to affiliate it as a social experience that may soon be available to them, it is imperative that they be included in educational programs that share verified information about AI, and in setting the standards upon which

training data should be developed. This can be done through advisory boards open to all people irrespective of their level of formal education and must be reflective of different sub-sets of the population likely to benefit from the technology.

At the level of AI developers/creators, there needs to be a diversity of the team working on developing AI. Movements and proclamations by companies globally are already driving towards employment diversity. However, its impact has been limited thus far. Human resources need to develop policies that seek out marginalised groups. Educational resources need to be made available to these groups to develop them for skilled roles so that they may be competent additions to the workforce. Women, people of colour, minority groups and transgender communities, amongst others, are an important part of the population, and including them on the team of AI developers will be more impactful than simply asking their opinion.

Critical thinking needs to occur at the level of the users, such as clinicians who implement and recommend AI outputs to patients. As powerful and accurate as AI in patient care might be, there is still room for error and bias, as has been shown above. Clinicians need to be hypervigilant in its use and to trust only what they can rationalise through their experience while still considering the human patient that they are managing. That said, an AI may not know that the patient is a Jehovah's Witness, but a clinician would know that and should be able to practically deduce that something like a blood transfusion would not be feasible for said patient. The clinician also needs to be trained to recognise biased and inaccurate output, and a concrete feedback loop needs to exist between the creators and users of AI so that previous shortfalls may enrich new knowledge production. The fear remains that new clinicians entering the

healthcare system in the era of AI, will have increased confidence in AI, and concurrently lack the experience to differentiate otherwise.

As the beneficiaries of AI in patient care, patients must also play an active role in the feedback cycle, with a platform to express their satisfaction and grievances on the user experience. Patients may also be invited to In-Depth Interviews (IDI), to explore the acceptability of incorporating Black Box AI into their care. Furthermore, a patient may be the most suitable candidate in the AI life cycle to recognise prejudice, especially if it is directed at them.

Balancing the human factors to use Black Box AI in a morally permissible way is a task that requires ongoing vigorous work and evaluation. Nonetheless, at the level of potential beneficiaries of AI predictions, that is, patients, beyond emphasizing accuracy, I contend that they should be given an opportunity to decide if and how advanced AI should be integrated into their care. Suppose the accuracy of the machine is the baseline upon which its permissibility is judged, and trust in its accuracy is the cornerstone of healthcare decisions. In that case, medicine has perhaps not moved away from paternalism. The decision maker has simply evolved from man to machine, irrespective of the logic used for such decisions. Involving patients in the decision on whether to use AI, and subsequently the AI output, creates a shared responsibility. That is, a patient choosing to use AI in their care accepts the shortfalls in its functioning and chooses to bear some responsibility should the output result in a negative outcome. Furthermore, the patient choosing AI in their care, knowing that only limited opaque knowledge is available, essentially consent to uninformed

consent, allowing the informed-decision-making gap to perhaps not be bridged but to be overlooked at their own discretion.

Furthermore, patients need to be considered as a credible source of knowledge regarding their own health, and their input ought to be considered when health decisions are being made. Training data sets also need to be built with the input of the subjects (patients) in addition to the evidence-based science. These are critical to overcoming testimonial injustice and promoting fair inclusion of knowledge.

Equally, increasing education on AI and its drawbacks, specifically about the Black Box, needs to occur to empower patients to be more conscientious in their decisions. The concern remains that patients may not be aware of the right questions to ask when presented with the option of introducing AI into their care. This is largely impacted by the media hype surrounding AI that presents an overload of information, some unverified and lacking substance. Hence, the patient may be satisfied with a machine that can provide no answer on how it works. They (patients) may be confident if it meets their needs, even if it goes against their values. Empowering patients with knowledge of AI, will allow them to appreciate the knowledge-exchange gap that comes with AI use and its implications thereof, allowing them to make an informed decision about its use, if not an informed decision on the process that it used.

At the level of training data, creating training data sets that combine the knowledge of the scientists, the clinicians and the patients as all valuable data sources must involve a level of open-mindedness that embraces the knowledge that science needs to be complemented by human judgement and rationale. For example, chemotherapy is a

gold standard for cancer, but is chemotherapy the right treatment for this person with cancer given his psycho, social, religious and moral standing?

Overcoming the gaps created by Black Box AI will require extensive reworking of the existing system frameworks, and this can only be achieved through acknowledging that science ought to be supplemented by human interface and that human interface should not be limited to just scientists and software engineers.

The reader would be correct to observe that AI is not necessarily doomed because of its current Black Box problem. Although much work still needs to be done to make AI align with the values of this report, there are multiple other uses of AI in healthcare that do not directly affect patient care and may create efficiency in healthcare settings that ultimately improve patient care. Concretely, AI, particularly advanced ones like Deep Learning, could be used to plan clinical rotations or triage emergency responses at the hospital and patient administration level. Who ought to have charge of what aspects of a patient who is coding?

Overall, patient administration is an area in which AI may be used to improve current structures. That is file management, organizing and storing digital records, and retrieving historical information, among others. These are labour and time-intensive jobs that historically would have required many additional staff to facilitate completion. However, AI may be a suitable alternative to efficiently manage these tasks, and may have the added advantage of minimising human error and creating easy accessibility routes to retrieving historical data.

Another area that AI may be useful in is patient triage, particularly in high patient influx healthcare settings. Effective triage by AI will assimilate the seriousness and line patients in order of priority. It may also be capable of learning to alert emergency services of hospital capacity and redirect ambulances to other centres as needed. This may simplify the complex hospital diversion system that currently exists. Staff who are generally responsible for these mundane tasks can then be reallocated to the floor, allowing for quicker patient assessment and management turnover over time.

Concluding Remarks

The potential of AI beyond direct patient care is constantly growing, and while its use in patient care may raise some important problems, there may still be room for its use with caution and monitoring.

In its current state, AI use in patient care in light of the Black Box problem remains problematic and ethically impermissible as it presents barriers to inclusivity and quality control of training data. It lacks explainability to account for the data that has been input, and it limits the information available for informed consent to be sufficiently obtained. Also, knowledge used in its development, evaluation and implementation creates an exponential bias that is not reflective of all whom the benefit is intended for.

These ultimately feed into the gaps that AI creates – knowledge exchange, responsibility and informed decision-making gaps. Extensive restructuring of existing systems will need to occur, which requires a level of open-mindedness and patient empowerment for some of these gaps to be somewhat bridged. Even then, it will not

be fully eliminated. In the next chapter, I address some objections that may be raised to my argument and respond to them accordingly.

CHAPTER FOUR

ADDRESSING POTENTIAL OBJECTIONS

In the previous chapter, I demonstrated how the Black Box problem presents barriers to inclusivity and quality control of training data and, in the process, fails to promote key moral norms that arise from epistemic (in)justice and social identity. Its unexplainability also creates challenges for obtaining informed consent. Furthermore, knowledge used in its development, evaluation and implementation creates an exponential bias that is not reflective of all whom the benefit is intended for. I address some potential objections to this project's central argument in this chapter.

Will an interpretable/explainable AI solve the problem *definitively*?

A critic may argue that if Deep Learning opacity could be overcome, then the Black Problem would no longer exist. This will undermine the quality of this project's argument. Precisely, the position this project defends is only relevant because these advanced AI types are opaque. This is the most pressing challenge for integrating this technology into patient care. However, if the processes of the AI were visible and could be understood, then data representation and effective use of all influencing factors, including socioeconomic, religious, cultural and value systems, could be confirmed. Furthermore, sufficient information would be available for users to provide information to patients for comprehensive informed consent. We would be able to study how AI promotes social relations or fosters epistemic justice.

There are many ways to address this criticism. First, interpretable AI research, as previously mentioned, is currently unproductive. Perhaps such studies may yield

success in the future, implying that it could give greater capacity to engineers and users to understand how an advanced AI works to yield or rank different recommendations. In that case, we may, indeed, be able to study how AI fosters epistemic justice or promotes individual social relations. However, the critic would notice that an AI may still fail to promote social relations or foster epistemic justice even if it is interpretable. There may be existing and un/recognized biases that the AI may pick from the training data or domain of operation that could still cause the AI to function in problematic ways. If users and developers are not vigilant, these biases may cause injustice to the patients. What this project's argument points to is that a number of factors must be present to align advanced AI and its use with key values of social relations and epistemic justice beyond interpretability or explainability. For example, our social contexts or environment must also be free of biases that cause harm to others.

Second, this criticism faces another important problem since it does not align with research on AI, particularly opaque AI, broadly. Many authors are convinced that AI opacity will only increase with further advancement in the field. For example, David Beer (2023) has remarked that "the story of neural networks tells us that we are likely to get further away from that objective [of AI explainability/interpretability] in the future, rather than closer to it." That is, an inverse relationship exists between the two. *The greater the impact AI comes to have on people's lives, the less they will understand why and how*. Unlike White Box AI and Grey Box AI, Black Box AI has higher predictive power mostly because of its inherent opacity. Summarily, *advancing* AI technology does not solve the problem of unexplainability (or uninterpretability), but only worsens the same. It seems explainability and interpretability can *effectively* be accomplished

by giving AI less predictive power. However, many AI engineers will be unwilling to follow this path since it would amount to retrogression.

Finally, even if interpretable AI had to surpass all these barriers and the vestiges of unethical biases were eradicated from the training data or domain of operation, AI would still be subject to complex proprietary patents, and the transparency may be limited to the creators of the AI. These AI creators may use the information to address bugs in the software and create updates which they deem necessary without being inclined, or even legally obligated to, share the internal processes. This would result in the clinicians and patients still being left in a metaphorical Black Box, in which they are still blind to the processes of the machine.

Should accuracy be prioritised over the patient's values or knowledge?

Another objection that may be raised against this project is the argument that it unjustifiably dismisses the science of AI backed by concrete scientific research with validated results. AI in patient care has been designed after consultation with experts in the field who have contributed extensive knowledge on the subject matter to ascertain safe and accurate functionality. It (AI) is a product of extensive research. In light of the preceding points, what matters is that an AI can enhance patient care and not whether we can or ought to understand how it does this.

In response, the reader would be correct to observe that the accuracy of the machine was never in question. However, it is far more important that the machine can function ethically. Notably, it would be important to understand how an AI promotes a patient's values to transition patient care in the era of AI to a patient-centred model. In the

Jehovah's Witness example where a blood transfusion is being recommended, it is scientifically accurate that a patient with an extremely low haemoglobin will benefit from a blood transfusion, and scientifically, the blood transfusion if matched correctly, is unlikely to cause any physical harm. However, that does not account for the patient's religious standing on the matter or the psycho-social harm and stigmatization that may result from that blood transfusion. It does not account for whether knowledge from a Jehovah's Witness was used in the creation of training data sets, or if it was included, whether the opaque system considered it when producing the output. It does not provide sufficient information explaining why that is the recommended option when attaining informed consent for that transfusion to occur. It also does not account for what a proxy decision maker should do if they are forced into a position where they are required to make this decision for a Jehovah's Witness, even though it may not reflect their view or that of the healthcare establishment. In the absence of these explanations, AI risks technologising paternalism in patient care. This project raises this concern by drawing on epistemic (in)justice and the thinking about social identity in relational autonomy.

Will the use of AI in patient care truly exclude human oversight?

A critic may object that AI use in patient care will not exclude human oversight. They may point out the numerous propositions that are being put forward to regulate its use in this regard, such as the 2021 proposal for the European Union regulatory framework on artificial intelligence. They may further point out that even if the machine were to be autonomous, the patient as the human counterpart, has the ability to say no if the intervention (for example) does not align with their personal values. The control would

ultimately lie with the clinician and/or the patient and, in the case of shared decision-making, with the family.

In response, the reader should observe that this report does not contest the lack of human oversight in the use of AI in patient care. However, it does contest the equity with which the AI's processes data to give outputs that will influence human decisions, and the oversight needed to overcome the shortcomings of the machine in that regard. Even if global AI regulations were in place ensuring human oversight in its use, and emphasising the rights of patients to say "no", neither of those hold the machine accountable for producing equitable outputs. It also reinforces the idea of imperialism. Why should it be acceptable to allow the use of advanced technology if it only produces outcomes suitable to certain groups of 'elites' that are encompassed in its training data? What power does the option to say "no" hold for people not included in this elite group if the machine cannot produce inclusive outcomes to which they would want to say "yes"? Why should we disregard what the machine is lacking in respect for persons and their values in favour of personal choice? Is there even a choice if all the options are based on one-dimensional imperial knowledge reinforced generationally? The AI's ability to function ethically should not be dependent on its human counterparts' ability to choose correctly.

Does drawing on more than one framework undermine this work?

Another objection that may be raised against my argument is that I have used more than one moral concept to argue my point. One might argue that using two moral concepts – epistemic (in)justice and social identity –undermines my capacity to think critically about the concepts, the contestations behind these concepts and the

implications that they may have for the use of AI in patient care in light of the Black Box problem.

What this criticism points to is that this project has failed to ethically engage the concepts and values it draws on. Ethical engagement with theories and frameworks, as Anye-Nkwethi Nyamnjoh and Cornelius Ewuoso (2023) have argued, ought to be deep engagement that consists of i) critically developing the justificatory work these concepts are meant to serve and ii) engaging with the scholarly debates around these concepts that might, in fact, have implications for the project's thesis. For example, contestations exist regarding whether promoting social relations (a consequentialist reading of social identity in relational autonomy) is more important than how these social relations are promoted (a deontological reading of social relations). Medical AI Black Box is only problematic for social relations if how these relations are promoted matters for social identity. However, some scholars appear to think that what matters is *that* social relations are promoted and not *how* (Jegede, 2009). Suppose a medical Black Box AI can promote social relations. In that case, how it does it is no longer relevant, at least from the consequentialist reading of the social identity in relational autonomy.

In response, although these two moral concepts may seem only loosely related, the moral norms that arise from them do have overarching similarities. Furthermore, this is not problematic as the primary objective of this research report is evaluative. Precisely, it aims to defend the usefulness of these moral concepts in relation to the research question. As opposed to if the project were descriptive, in which case a deeper look into the moral concepts would have been relevant.

Furthermore, Black Box Medical AI that promotes social relations is not any more acceptable if that promotion is not fair and inclusive and respects the person and their values. Suppose an AI promotes social relations, but is only reflective of historical privilege. Then, yes, it promotes social relations, but it fails to guarantee any form of equity. As this project has highlighted, there are multi-dimensions that need to be addressed for the value of promoting social relations to be fully realized.

Notwithstanding, it is worth stating that the requirement to engage concepts deeply is sound. Does the evaluative goal of the project undermine the project's thesis in any way? No, it does not. Suppose that the ethical concepts had to be engaged more deeply and a full descriptive component of this paper had been included, it would not change the argument or the outcome. Deep thinking of these ethical concepts had to have occurred in order to achieve the evaluative goal of this thesis. Reflectively, the arguments made would remain, irrespective of the depth.

Concluding Remarks

Reflecting on the objections above that critics may raise, it is evident that AI's unexplainability raises critical problems for the ethical use of Black Box AI or even more opaque AI in patient care. Progression in the field is likely to create more opaque systems, which only intensifies this problem. Furthermore, even though AI is founded on concrete science, the human experience requires that human factors should be considered for the output to be a holistic fit for the person rather than just a task well done. Ongoing efforts are being made to regulate the use of AI. However, a clinician's duty to make the correct final decision does not account for what AI lacks in

transparency and inclusivity. This project has defended in this chapter that the use of two moral concepts for the purpose of this research paper is appropriate given that the objective of the paper is evaluative, implying that the main objective is to draw on concepts that can enhance our thinking about whether and how AI should be used in light of the Black Box problem.

CHAPTER FIVE

CONCLUSION

The integration of AI in healthcare is an ongoing process, and within the next 5-10 years, it is likely to reach its full potential. As promising as the prospect remains, many challenges with the use of AI in healthcare exist, particularly in light of the Black Box problem. The Black Box problem remains a limiting factor in the use of AI given the accompanying unexplainability of the internal processes that produce outcomes that may directly impact lives, particularly when used for patient care, creating the gaps identified in this paper – knowledge exchange gap, responsibility gap and informed decision-making gap.

Till now, the basis upon which AI has been progressed ethically has been centred on the scientific accuracy on which the machine can be relied. However, it fails to encompass the full human experience that occurs beyond science, involving a socially integrated person with their value system. The moral norms that arise from epistemic (in)justice and social identity in the works of relational autonomy address this knowledge gap.

Inclusivity of knowledge and sources of knowledge, empowerment of autonomy through equal resources, and respect for people and their values play an integral role in assessing the permissibility of using Black Box AI in patient care. It challenges three stages in which Black Box AI processes information - the gathering of training data, the processing of that data and how the output is used in a clinical setting - to demonstrate that Black Box AI use is problematic in patient care. It presents barriers

to the inclusivity of training data, uses opaque systems to process the data whereby factors considered in decision-making cannot be confirmed, and does not provide sufficient information for informed consent.

With the emergence of explainable AI, some might argue that the Black Box problem may become redundant. However, ongoing research continues to show that explainable AI is less accurate than Black Box AI. In a patient-centred industry, whilst accuracy is critical to ensuring optimal care, explainability is far more important to shift care away from (technological) paternalism. However, research shows that AI will only become less understandable to the human mind as it becomes more sophisticated. Others may argue that this thesis dismisses the science behind AI, which has shown promising results in optimising patient care. However, the gold standard of care may be scientifically accurate but not in line with a patient's psycho, social and religious value systems.

Drawing on the moral norms that arise from epistemic (in)justice and social identity in the works of relational autonomists, the use of AI in patient care is deeply problematic and ethically impermissible in light of its Black Box problem. However, its use cannot be dismissed entirely, as it can be a tool used in non-patient-centred tasks to improve efficiency and functionality in the healthcare setting.

Although the project has outlined many recommendations in the previous sections, it is worth stating that to overcome the barriers presented by Black Box AI, a level of open-mindedness needs to be adopted in the AI development process to include various underrepresented groups as valuable sources of knowledge, to train

underrepresented groups with the necessary skills to be part of the developing teams of AI, and to empower patients with the knowledge they would need to make decisions about themselves with the rising use of AI.

Deeper thinking about the impact of AI on the human experience needs to be done to appreciate any other barriers that may exist from backgrounds other than decolonial scholarship and relational autonomy. Future projects can take on this task.

REFERENCES

Afnan, M. A. M., Liu, Y., Conitzer, V., Rudin, C., Mishra, A., Savulescu, J., & Afnan, M. (2021). Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Human Reproduction Open*, *2021*(4). https://doi.org/10.1093/hropen/hoab040

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-01332-6

Astromskė, K., Peičius, E., & Astromskis, P. (2021). Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. *AI and Society*, *36*(2). https://doi.org/10.1007/s00146-020-01008-9

Baier, A. (1985). *Postures of the Mind: Essays on Mind and Morals*. University of Minnesota Press.

Baier, A. (1986). Trust and Antitrust. *Ethics*, *96*(2), 231–260. https://about.jstor.org/terms

Baumann, H. (2008). Reconsidering Relational Autonomy. Personal Autonomy for Socially Embedded and Temporally Extended Selves. *Analyse & Kritik*, *30*(2). https://doi.org/10.1515/auk-2008-0206

Beauchamp, T. L., & Childress, J. F. (1979). Principles of Biomedical Ethics (Principles of Biomedical Ethics. *Oxford University Press*.

Beer, D. (2023, March 31). *AI will soon become impossible for humans to comprehend – the story of neural networks tells us why*. https://theconversation.com/ai-will-soon-become-impossible-for-humans-to-comprehend-the-story-of-neural-networks-tells-us-why-199456

Blake, V. (2012). When is a patient-physician relationship established? *Virtual Mentor*, *14*(5). https://doi.org/10.1001/virtualmentor.2012.14.5.hlaw1-1205

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

Bresnick, J. (2018). *Top 12 Ways Artificial Intelligence Will Impact Healthcare*. https://healthitanalytics.com/news/top-12-ways-artificial-intelligence-will-impact-healthcare

Bua, E., & Sahi, S. L. (2022). Decolonizing the decolonization movement in global health: A perspective from global surgery. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.1033797

Burns, E., Laskowski, N., & Tucci, L. (2023, March). *What is artificial intelligence (AI)?* https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence

Byskov, M. F. (2021). What Makes Epistemic Injustice an "Injustice"? *Journal of Social Philosophy*, *52*(1). https://doi.org/10.1111/josp.12348

Christman, J. (2004). Relational autonomy, liberal individualism, and the social constitution of selves. In *Philosophical Studies* (Vol. 117, Issues 1–2). https://doi.org/10.1023/b:phil.0000014532.56866.5c

Cohen, I. G. (2020). Informed Consent and Medical Artificial Intelligence: What to Tell the Patient? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3529576

Dheensa, S., Fenwick, A., & Lucassen, A. (2016). "Is this knowledge mine and nobody else's? I don't feel that." Patient views about consent, confidentiality and information-sharing in genetic medicine. *Journal of Medical Ethics*, *42*(3). https://doi.org/10.1136/medethics-2015-102781

Dove, E. S., Kelly, S. E., Lucivero, F., Machirori, M., Dheensa, S., & Prainsack, B. (2017). Beyond individualism: Is there a place for relational autonomy in clinical practice and research? *Clinical Ethics*, *12*(3). https://doi.org/10.1177/1477750917704156

Durán, J. M., & Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, *28*(4). https://doi.org/10.1007/s11023-018-9481-6

Durán, J. M., & 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*(5). https://doi.org/10.1136/medethics-2020-106820

Fanon, F. (1963). *The Wretched of the Earth*. Grove Press.

Fayemi, K., & Adeyelure, M. (2016). Decolonizing Bioethics in African. *BEOnline Journal of the Center for Bioethics and Research*, *3*(4). https://doi.org/10.20541/beonline.2016.0009

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: It is possible to trust medical AI. In *Journal of Medical Ethics* (Vol. 47, Issue 6). https://doi.org/10.1136/medethics-2020-106922

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Hague, D. C. (2019). Benefits, Pitfalls, and Potential Bias in Health Care AI. *North Carolina Medical Journal*, *80*(4). https://doi.org/10.18043/ncm.80.4.219

Hall, D. E., Prochazka, A. V., & Fink, A. S. (2012). Informed consent for clinical treatment. In *CMAJ. Canadian Medical Association Journal* (Vol. 184, Issue 5). https://doi.org/10.1503/cmaj.112120

Hardin, R. (2002). *Trust and Trustworthiness*. Russell Sage Foundation.

Hastings, R. (2022, March 4). *Examples of Artificial Intelligence (AI) in 7 Industries*. https://emeritus.org/blog/examples-of-artificial-intelligence-ai/

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics*, *46*(7). https://doi.org/10.1136/medethics-2019-105935

Ho, A. (2008). Relational autonomy or undue pressure? Family's role in medical decision-making. *Scandinavian Journal of Caring Sciences*, *22*(1). https://doi.org/10.1111/j.1471-6712.2007.00561.x

IBM. (2022). *IBM Global AI Adoption Index 2022*. https://www.ibm.com/downloads/cas/GVAGA3JP

Jegede, S. (2009). African ethics, health care research and community and individual participation. *Journal of Asian and African Studies*, *44*(2). https://doi.org/10.1177/0021909608101412

Kajee, L. (2010). Disability, social inclusion and technological positioning in a South African higher education institution: Carmen's story. *Language Learning Journal*, *38*(3). https://doi.org/10.1080/09571736.2010.511783

Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. In *Gastrointestinal Endoscopy* (Vol. 92, Issue 4). https://doi.org/10.1016/j.gie.2020.06.040

Kawamleh, S. (2022). Against explainability requirements for ethical artificial intelligence in health care. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00212-1

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1). https://doi.org/10.1002/hast.973

Mackenzie, C., & Stoljar, N. (2000). Introduction: Autonomy Refigured. In *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self* (pp. 3–31). Oxford University Press.

Maldonado-Torres, N. (2016). *Outline of Ten Theses on Coloniality and Decoloniality **. Frantz Fanon Foundation.

Mcleod, S. (2023). *Social Identity Theory: Definition, History, Examples, & Facts*. Simply Psychology. https://www.simplypsychology.org/social-identity-theory.html

Ndugga, N., & Artiga, S. (2021). Disparities in Health and Health Care: 5 Key Questions and Answers | KFF. *Kaiser Family Foundation*, *Cdc*.

Nedelskyt, J. (1989). Reconceiving Autonomy: Sources, Thoughts and Possibilities. *Yale Journal of Law and Feminism*, *1*, 7–36.

Neri, E., Coppola, F., Miele, V., Bibbolino, C., & Grassi, R. (2020). Artificial intelligence: Who is responsible for the diagnosis? In *Radiologia Medica* (Vol. 125, Issue 6). https://doi.org/10.1007/s11547-020-01135-9

Nordling, L. (2018). How decolonization could reshape South African science. *Nature*, *554*(7691). https://doi.org/10.1038/d41586-018-01696-w

Nyamnjoh, A.-N., & Ewuoso, C. (2023). What Constitutes Ethical Engagement with Africa and the Global South? *The American Journal of Bioethics*, *23*(7). https://doi.org/10.1080/15265161.2023.2207537

Oelofsen, R. (2015). *DECOLONISATION OF THE AFRICAN MIND AND INTELLECTUAL LANDSCAPE*. *16*(2), 130–146.

Pintelas, E., Livieris, I. E., & Pintelas, P. (2020). A Grey-Box ensemble model exploiting Black-Box accuracy and White-Box intrinsic interpretability. *Algorithms*, *13*(1). https://doi.org/10.3390/a13010017

PwC. (2017). *Sizing the prize: What's the real value of AI for your business and how can you capitalise?* https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., & Coghlan, S. (2022). The three ghosts of medical AI: Can the black-box present deliver? *Artificial Intelligence in Medicine*, *124*. https://doi.org/10.1016/j.artmed.2021.102158

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2022). *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks*. http://arxiv.org/abs/2207.13243

Regunath, G. (2021, August 26). *6 Ways AI is Transforming the Finance Industry*. https://www.advancinganalytics.co.uk/blog/2021/8/26/mo482xc608yon5wplj0iobz7cjmtgk

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, *26*(5). https://doi.org/10.1007/s11948-020-00228-y

Saini, H. (2022, March 21). *8 Applications of AI in the Aerospace Industry*. https://analyticssteps.com/blogs/8-applications-ai-aerospace-industry

Slabbert, M., & Labuschaigne, M. (2022). Legal reflections on the doctor-patient relationship in preparation for South Africa's National Health Insurance. *South African Journal of Bioethics and Law*, *15*(1). https://doi.org/10.7196/SAJBL.2022.v15i1.786

Smith, L. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. Zed Books Ltd.

Solberg, B., & Steinsbekk, K. S. (2012). Managing incidental findings in population based biobank research. *Norsk Epidemiologi*, *21*(2). https://doi.org/10.5324/nje.v21i2.1494

Sullivan, H. R., & Schweikart, S. J. (2019). Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA Journal of Ethics*, *21*(2). https://doi.org/10.1001/amajethics.2019.160

Táíwò, O. (2020). Being-in-the-Room Privilege: Elite Capture and Epistemic Deference. *The Philosopher*, *108*(4).

Turing, A. M. (1950). Computing machinery and intelligence-AM Turing. *Mind*, *59*(236).

Ursin, F., Timmermann, C., & Steger, F. (2022). Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics*, *36*(2). https://doi.org/10.1111/bioe.12918

Varelius, J. (2006). The value of autonomy in medical ethics. *Medicine, Health Care and Philosophy*, *9*(3). https://doi.org/10.1007/s11019-006-9000-z

Von Bismarck, H. (2012). *Defining Decolonization*. The British Scholar Society. http://britishscholar.org/publications/2012/12/27/defining-decolonization/

Wadden, J. J. (2021). Defining the undefinable: the black box problem in healthcare artificial intelligence. *Journal of Medical Ethics*, *48*(10). https://doi.org/10.1136/medethics-2021-107529

Walter, J. K., & Ross, L. F. (2014). Relational autonomy: Moving beyond the limits of isolated individualism. *Pediatrics*, *133*(SUPPL. 1). https://doi.org/10.1542/peds.2013-3608D