THE ETHICAL IMPLICATIONS OF AI IN HEALTHCARE

RAND HIRMIZ

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN
PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PHILOSOPHY

YORK UNIVERSITY

TORONTO, CANADA

July 2024

Abstract

This dissertation focuses on the ethical implications of implementing artificial intelligence in healthcare. Three approaches are considered regarding the role that artificial intelligence ought to play in healthcare: 1. the neo-luddite approach, which urges against the implementation of artificial intelligence in healthcare altogether; 2. The substitutive approach, which favours the goal of ultimately substituting artificial intelligence systems for human clinicians; and 3. The assistive approach, which favours the implementation of artificial intelligence systems in healthcare, but only as a tool to assist, rather than replace, human clinicians. The dissertation begins by looking at what excellence in the practice of medicine entails and deriving some formal duties of clinician, before proceeding to assess the risks or threats that different approaches to the use of AI may pose to the practice. The first position evaluated is the neo-luddite approach. In evaluating this approach, some arguments in its favour are considered, including worries regarding biased algorithms, decreased communication with patients, confidentiality, the resurgence of paternalism, the potential for unethical design, overdependence on AI, and inscrutability. After responding to these concerns and rejecting the neo-luddite approach, the dissertation shifts its focus on to the assistive and substitutive approaches. Before evaluating these two approaches, the concepts of "assistance" and "substitution" are analyzed and definitions of assistive and substitutive technology are provided. With these definitions at hand, the remaining two approaches are assessed, beginning with the substitutive approach, which is then rejected for reasons related to both AI's limited capacity to foster a strong patient-clinician relationship (due to its limited capacity for compassion, recognition, communication, creating trust, and respecting patient autonomy) and concerns on a more systemic level (including its likelihood to cause responsibility gaps and produce erroneous and biased decisions). The final part of the dissertation focuses on the assistive approach. Issues related to clinician-AI disagreement and the use of opaque decision support systems are addressed and recommendations for handling these issues (derived after a discussion of the epistemic role of assistive AI) are provided. The dissertation concludes with the recommendation that the assistive approach offers a superior outlook for the future of medicine.

Acknowledgements

I would like to thank Regina Rini for her guidance, patience, and support, for sharing her brilliant intellect with me, and for just being an altogether wonderful, kind, and caring human being. I don't believe there's a single person who hasn't heard me talk about how lucky I am to have had the privilege of working with her. She embodies all the qualities I aspire someday to have as researcher, supervisor, teacher, and person.

I would also like to thank Duff Waring and Parisa Moosavi for their support and encouragement and for their incredibly helpful feedback, without which this dissertation would not have been anything resembling what it is now.

I want to thank Michael Giudice, Robert Myers, and Alice MacLachlan for all that they've done for me throughout my PhD.

I also want to thank everyone at York's philosophy department who made these past few years incredible. Thank you especially to John Buchanan, Andrew Buzzell, Fatima Sadek, Daniel Rodrigues, Nathan Dyck, and Jef Delvaux.

Finally, thank you to my family and to my friends outside of academia who have celebrated with me in the good times and been my shoulder through the bad times. My vocabulary is too limited to express the love I have for you.

Table of Contents

Introduction

As the capabilities of artificial intelligence systems expand, their use is expected to expand accordingly. Given the rapid technological advancement we have achieved in a short period of time and the advancement we continue to aim for, it is no longer difficult to imagine the design and implementation of machines that completely take over the tasks and entire roles once thought to only be able to be carried out by humans. One field in which this is a real possibility is healthcare. We can envision being diagnosed completely by a machine, being operated on by a surgical robot, being given a treatment option by an algorithm, being washed, fed, given medicine, and moved out of bed by a carebot, and all without the aid (or even the presence) of a single human health care provider. In fact, given some of the current existing technology, we have reason to believe that a future in which *all* medical tasks are assigned entirely to machines is not that far out of reach. IBM's Watson for Oncology and Microsoft's Project Hanover, for instance, can analyze medical research from journals and make predictions about the most effective cancer treatments for patients. Tencent's AI Medical Innovation System (AIMIS) combines AI with medical imaging to, within seconds, accurately diagnose disease. The Fraunhofer Care-O-Bot (though primarily used to assist the elderly in nursing homes) can be used to bring patients medicine, food, and water, while RIKEN-TRI's RIBA can move patients from or onto a wheelchair or bed.

Amid all this excitement about technological advancement, however, it is important to pause and consider not only what will be gained, but also what will be sacrificed as AI takes on a greater and greater role in medicine. Importantly, we need to consider what is it that we are willing or not willing to sacrifice. This is precisely the question I aim to answer in this dissertation. I do this by evaluating three different approaches for how to implement AI in healthcare to see which one offers the best outlook on the future of medicine. The first is the neo-luddite approach, which urges against the implementation of AI in healthcare altogether. The second is the substitutive approach, which supports the goal of ultimately replacing human clinicians with AI. The third is the assistive approach, which favours the implementation of AI in healthcare, but only as a tool to assist, rather than replace, human clinicians.

I begin this project with a discussion of the nature of medicine in Chapter 1. Appealing to Eric Cassell's account of healing as the wellbeing of the *whole* patient as well as Mathison and Davis's value-promotion account of medicine, I defend an account of medicine and healing that goes beyond just tending to the physical aspects of the patient and includes among its goals the promotion of patient values. From this account, I proceed to present five duties that I suggest clinicians ought to fulfill in order to exhibit excellence in the practice. These five duties include: compassion, recognition, communication, creating trust, and respecting patient autonomy.

In Chapter 2, I consider the neo-luddite approach and evaluate some arguments in its favour, presented in the form of concerns about the implementation of AI in healthcare. These include concerns about biased algorithms, decreased communication with patients, confidentiality, the resurgence of paternalism, the potential of unethical design, overdependence on AI, and inscrutability. After taking up each of these concerns, I conclude that they are all insufficient to justify refraining from utilizing AI in healthcare altogether.

In Chapters 4, 5, 6, I focus on the remaining two approaches. But before I move on to evaluating either the substitutive or the assistive approach, I use Chapter 3 to provide a conceptual analysis of technological assistance and substitution so that it may become clear what type of technology it is that is being advocated for or against under each approach moving forward. As I demonstrate in this chapter, the line between "assistive" and "substitutive" is not always easy to draw, especially in the face of some hard cases.

Chapters 4 and 5 are then dedicated to evaluating the substitutive approach. In Chapter 4, I evaluate AI's capacity to foster a strong patient-clinician relationship. After demonstrating its limited capacity to fulfill each of the duties discussed in Chapter 1 (namely, compassion, recognition, communication, creating trust, and respecting patient autonomy), I conclude that it is not an appropriate substitute for a human clinician in that respect.

In Chapter 5, I consider two further concerns with substitutive AI on a more systemic level and outside of the patient-clinician relationship. The first issue concerns its limited capacity to detect and correct its own erroneous and biased decisions and the second issue concerns its likelihood to create a responsibility gap in which there is no appropriate agent that may be held responsible for the harms that the AI's actions or decisions may bring about. These concerns serve as additional reasons for why AI would not be a suitable substitute for a human clinician, even outside of the clinical encounter.

Though only one approach remains after both the neo-luddite and the substitutive approaches have been rejected, this remaining approach is far from perfect. In Chapter 6, I consider issues that arise only on the assistive approach. In particular, I focus on the dilemma that arises when considering how to resolve disagreements between human clinicians and AI decision support system, especially when the latter's recommendations are opaque. After providing an account of the epistemic role of assistive AI, where I classify assistive AI as an "epistemic nudge" rather than an epistemic superior, epistemic inferior, or epistemic peer, I conclude by offering a recommendation for handling clinician-AI disagreements on the basis of this account and discuss what this recommendation entails with regard to the allocation of responsibility and accountability in such cases and with regard to the permissibility of utilizing opaque AI in medicine. Given the potential for its issues to be resolved , and considering that the assistive approach does not require us to sacrifice remotely as much as the neo-luddite and the substitutive approaches do, I end the dissertation by recommending the assistive approach as the superior one and the one that we ought to aim for in the future of medicine.

Chapter 1: Healing, Value Promotion, and the Duties of Clinicians

1.1 Introduction

A discussion of the core values of medicine or the duties of clinicians must account for future alterations to the practice that may be brought on by AI as well as general sociocultural changes. With that in mind, the purpose of this chapter will be to discuss what I take the nature of medicine to be and what this means in terms of clinicians' duties, both currently and in the face of any future technological and sociocultural changes. I appeal to Eric Cassell's account of healing as well as Mathison and Davis's value-promotion account to show that value promotion is, and ought to continue to be, an important goal of medicine. I then proceed to present a set of duties (namely, compassion, recognition, communication, creating trust, and respecting patient autonomy) that I suggest all clinicians must be able to fulfill if they are to be able to promote this goal. I end the chapter by stressing the main point I intended to make by beginning this project in this way: that while society's beliefs may change in the future and while technology may alter the techniques of medical practice, this is no reason to conclude that these duties must change with them. On the contrary, if there ever comes a time where we no longer deem these duties necessary of physicians, then this will be a sign that we are failing to promote one of the most important goals of medicine, and thus falling short in achieving excellence in the practice.

1.2 What Healing Is

The traditional view of medicine views bodily health as an objective condition – something that is independent of patients' opinions and desires, and something that can be observed and recognized by professionals (Kass, 1975, p. 22). Moreover, the promotion of bodily health is deemed on this view of medicine to be the sole end of the profession, and all other goals that do not aim at the bodily health of the patient (including, for Leon Kass, "happiness" or "contentment," which he believes to be the task of agents other than doctors) are considered to be "false goals for medicine" (Kass, 1975, pp. 13-14).

In contrast to this traditional view are more patient-centred approaches to medicine that view health as at least involving to some extent a subjective component. Eric Cassell, for instance, promotes a view of sickness that goes beyond the physical, and a view of healing that goes beyond curing disease or treating an injury, and involves, instead, a focus on maintaining or restoring functioning in *any* part of the patient. He explains that well-being is not to be equated with freedom from disease – that a patient can have a disease and maintain a sense of well-being, and vice versa, where a patient can appear to be disease-free according to objective metrics and assessment tools, yet the patient may nonetheless lack a sense of well-being (Cassell, 2012, p. 83). For Cassell, then, the goal of medicine is the wellbeing of the patient, where wellbeing is "not just physical and not just emotional. Well-being is a patient's goal—only the patient knows whether he or she has a sense of well-being" (ibid., p. 83). When patients have a sense of wellbeing, he explains, they "believe they can accomplish their purposes and goals...they can do the things they need and want to do to live their lives the way they want to." (ibid., p. 84). Of course, treating the disease remains an essential aspect of the clinician's role, but on Cassell's account, the purpose of treating the disease is not merely to achieve better results according to quantitative assessments, but to restore the patient's wellbeing and return them to being a "functioning whole person" (ibid., p. 87).

Cassell's account of wellbeing is not accepted by everyone, however. Among his rivals would be those who endorse an "objective list account" of wellbeing, according to which, as put by Parfit, "certain things are good or bad for people, whether or not these people would want to have the good things, or to avoid the bad things." (Parfit, 1984, p. 499). These lists vary among objective list theorists,[1] but things like knowledge, relationships, autonomy, and achievement are some examples of the types of things that some objective list theorists have in mind. This conception of wellbeing is defended by Christopher Rice (2013) against a subjective account as follows:

> Imagine that there are times when some people are conscious but experience no reactive attitudes at all. These people do not desire or take pleasure in things around them and are not motivated to choose or approve of anything...If states of well-being must involve occurrent reactive attitudes, then all people in such circumstances would be equally well-off, since they would all entirely lack well-being. But this is not the only

[1] See, for example, Ross, 2002, p. 134-141; Nussbaum, 2011, pp. 33-34; Rice, 2013, p. 200

> possibility. Although all people in these circumstances would be very poorly off, it seems that some people would be at least a little better off than others on account of the objective goods that are present in their lives. These would be people who are supported by loved ones or who know the circumstances of their lives or other meaningful truths about the world. This suggests that these things can contribute to well-being, whether or not people hold positive reactive attitudes toward them (p. 208).

Though arguing against arguments like the one put forth by Rice would fall outside the scope of this chapter, I would like to, at the very least, put to rest possible worries with Cassell's conception of wellbeing. That is, while I remain unconvinced that people with no reactive attitudes whatsoever would in fact be better or worse off than others with no reactive attitudes due to having or lacking things like support by loved ones or meaningful truths about the world, I will not attempt to take up this subject here. Instead, I would like to point out that Cassell's account of healing and wellbeing can still be useful even to an objective list theorist, for two reasons: first, though some objective list theories (e.g., knowledgism) may not allow room for subjective affective states like preferences, tastes, desires, and interests, this is not true of all objective list theories. In fact, subjective affective states are compatible with several objective list theories (particularly those that include, for instance, pleasure, happiness, and/or achievement) (Fletcher, 2015, pp. 156-157). The second reason is that one may accept that there is a list of objective goods that it makes one's life worse off to not have (regardless of whether or not they acknowledge it), and yet still accept that beyond this threshold and beyond having these basic goods, wellbeing (at least if we proceed to *then* think of it in the sense that does involve reactive attitudes) will at that point depend on the individual. So we may say that beyond the threshold where these objective goods are necessary, only the individual themself can know whether or not they have wellbeing in the sense that Cassell has in mind (that is, wellbeing in the sense of being able to live their lives the way they want to). So we need not be extreme subjectivists to accept Cassell's account – we need only agree that reactive attitudes do also play a role when it comes to functioning and wellbeing. If the objective list theorist is still not convinced, then I would like to appeal to the same method to defend this claim as Rice did to defend his: Imagine that a group of individuals were all given the same lot in life – they all had equal opportunities, equal knowledge about the world, equal loving relationships, etc. and thus, all were equally well off according to any particular objective list of wellbeing. Would we not

say that those who had positive reactive attitudes to their situations (perhaps because of their awareness that these things allowed them to pursue the kind of life they wanted) were better off than those who had no reactive attitudes or those who had negative reactive attitudes to their situations? Intuition about wellbeing and reactive attitudes would tell us that they would, in fact, be better off.

Another concern one may raise at this point with regard to Cassell's conception of well-being as something that only patients themselves would know, is that patients can at times be wrong about what constitutes their own well-being. A patient may insist that well-being for them amounts to being injected with drugs that induce euphoria at all times. Does this mean that the physician is obligated to provide medical assistance with this?

Cassell's conception of well-being does not have to imply these things, however. Patients can certainly be wrong about what they may believe would constitute their own well-being. This may be due to a mental illness, a lack of information about the consequences of what they believe they want, or any other reason (or combination of reasons). The job of the physician is not to simply provide the services requested with no questions asked. Rather, the job of the physician is to try to understand what the patient's goals are and what it would truly mean for them to live the kind of life they want to live. Once the patient's values and life goals are made more clear to the physician, then the physician can perhaps offer suggestions for alternative methods that may allow them to live the kind of life they want, inform of them of the potential long-term consequences of what they are asking for, help them understand why it may be that they are asking for these services from the physician, and, overall, seek to understand and promote their values while helping the patient to better understand them themselves.

Of course, it is possible that what it actually takes to promote a particular patient's well-being requires an intervention that the physician does not approve of or is not comfortable with, such as interventions deemed to be medically futile, proscribed treatments (those prohibited by law or prohibited by widely accepted public policies) or discretionary treatments (those that law and policy permit clinicians to refuse to administer).[2] Important to stress is that despite the importance placed on promoting patients' well-being and despite the emphasis on the subjective experiences of patients with regard to their own health and overall well-being, we need not commit ourselves to an account in which clinicians are under an obligation to provide whatever

[2] See White and Pope, 2015, pp. 79-80.

intervention the patient takes to be important in promoting their purposes or goals. For instance, in 1997 and 1999, surgeon Robert Smith agreed to amputate the limbs of two individuals who had nothing physically wrong with them, but who wanted the operation because "they felt incomplete with four limbs but would feel complete with three" (Dyer, 2000). Even though Dr. Smith performed the amputations, many other doctors turned the two individuals away. We generally do not expect most doctors to abide by patient requests or promote their goals at all costs, and even those who argue that there are cases in which it is permissible to amputate a healthy limb[3], generally do not go as far as to suggest that the doctors who turn such individuals away will have failed to abide by their roles as clinicians or that they ought to be sued for malpractice under "failure to treat" or criticized by their colleagues. This is because, while promoting patients' well-being in the sense of promoting their goals and purpose is important, it is not a goal that cannot be overridden by other considerations. Still, it is far from being an unimportant goal. It is precisely for this reason that I would like to proceed by proposing the adoption of an account of medicine that takes the promotion of patients' values to be an important goal of medicine, but not the *sole* goal. This is the kind of account that Mathison and Davis (2021) propose.

1.3 Value Promotion as a Goal of Medicine

Mathison and Davis promote what they call a "value promotion" view of medicine, which incorporates some aspects of Cassell's account of medicine, but where Cassell's account focuses on restoring functioning to allow for patients' overall wellbeing, Mathison and Davis's account is somewhat broader and encompasses value promotion in a sense that goes beyond just restoration of functioning. Mathison and Davis distinguish their value-promotion account from the healing view (where promoting the health of the patient is the only goal of medicine) and the best interest view (where the goal of medicine is to promote the patient's best interest), and can be summarized as maintaining that "one goal of medicine is to promote the patient's autonomous values" (Mathison and Davis, 2021, p. 494). Once again, however, they are not committed to the (much stronger) view that value promotion is the *only* goal, or even the *most important* goal, of medicine. Though Mathison and Davis do not provide answers to questions such as "which other

[3] See, e.g., Bayne and Levy, 2005.

goals may trump value promotion?" or "what ought to be done when there is conflict between value promotion and other goals of medicine?", they aim to show that value promotion is an important goal of medicine that has virtues that often go underappreciated in the literature.

In elaborating on their account, they emphasize that values are not the same as desires – they are much more. Values are to be understood, according to them, as "the foundation on which an individual's identity and life are built" and that these values are at the core of rational deliberation and "reflect the basic idea that capacitated individuals are self-governing beings who have a unique power to determine how their lives will go and what may be done to their bodies" (ibid., p. 496). They also assert that promoting patient values does not mean fulfilling any request made by the patient – this is not a goal of medicine, and at times patients' requests may conflict with their true values.

An important thing to note about the value promotion account is that, while the healing view and the best-interest view of medicine "apply standards that exist independent of the patient's own values—that is, their focus is squarely on objective, agent-neutral concepts of health and interest that may or may not comport with the patient's own conception of these ideas" (ibid., pp. 496-497), the value-promotion view does not take values to be universally shared, though they may be often widely shared. Viewed in this sense, value-promotion as a goal of medicine can be seen in many areas and current medical practices. One example is sports medicine, where athletes may receive more invasive procedures even when less invasive ones are available, because the more invasive ones will better allow them to continue pursuing their passion. Some other examples include interventions for increased sexual performance, birth control, vasectomies, organ donation, aesthetic interventions, elective abortion, and gender confirmation surgery - none of these are intended to heal and may not always be in the best interest of the patient (if we take "best interest" to be different from promoting patient values),[4] but they do serve to promote their values. While promoting patient values is consistent with promoting health and acting in the best interest of the patient, the main idea behind Mathison and Davis's account is that patients' values should not be seen as constraints on, but rather, as guides to, care.

[4] Mathison and Davis point out that proponents of this view rarely say what they actually mean by 'best interest of the patient', but that often they refuse an intervention which in fact aligns with patient values, because they claim that it is not in the patient's best interest. This, for Mathison and Davis, shows that "best interest" does not necessarily translate to "value promotion" (Mathison and Davis, 2021, p. 497).

If one accepts the account of healing I have advocated for above as well as value-promotion as an important goal of medicine, the following will naturally follow: the goals and duties of clinicians stem from what it means to achieve excellence in the practice itself, and since achieving excellence in the practice of medicine means healing in the sense that encompasses the wellbeing of the whole patient (that is, providing patients with what is needed to allow them to be able to do what they need to do to live their lives according to their own purposes and goals), this means that the duties of clinicians ought to go beyond simply healing the physical disease or injury. Importantly, it means that they ought to go beyond applying an objective standard of what it means to promote wellbeing. The duties of clinicians ought to also include taking measures that will aid in promoting patient values, which means focusing on patients individually and figuring out what truly matters to them.

At this point, I would like to briefly acknowledge that there are cases in which patients may not have the ability to form or express their own genuine values (such as late-stage Alzheimer's patients, very young children, or comatose patients, to name a few). In these cases where it is not possible to promote the patients' values (either because the patients have no values, lack the competence to form true values that may be relied upon, or are unable to express their values), we may resort to another principle to guide decision-making. This may include, for instance, resorting to the advance directive principle whereby a clear advance directive, through explicit instructions or the appointing of a proxy, is adhered to. Or, perhaps, we may resort to the substituted judgment principle, where a surrogate chooses on behalf of a previously-conscious patient as they believe the patient would have chosen had they been competent and aware of their medical options as well as their current condition. Or, perhaps we may resort to the best interest principle (where a surrogate chooses what they believe is in the patient's overall best interest) when making decisions on behalf of a child or a never-competent patient.[5]

I will not attempt to take up the debate here over which approach is best or address the concerns with each of these approaches or weigh in on which is most appropriate in which scenario. It is sufficient to propose that the account I am defending here ought to apply whenever we are dealing with competent adults capable of forming and expressing their own values. Beyond this, the goal of medicine becomes a somewhat diluted version of this, where the goal is

[5] For a more in-depth discussion and analysis of these guidance principles, see Buchanan and Brock, 1989, pp. 87-151

to promote *what it is believed* the patient's values would be if the patient was capable of having values or having enough knowledge and competence to form true values and be able to express them – the question of what exactly this entails, however, is beyond the scope of this chapter.

1.4 Duties of Clinicians

Once again, my goal in the preceding sections was to demonstrate that excellence in the practice of medicine means promoting, as much as possible, the whole patient's wellbeing, not just their bodily health. The way to do this, I went on to explain, was by viewing the promotion of patient values as an important part of the practice. It is the patient's values that will inform clinicians of what is needed of them by that patient and how best to proceed. While the specific tasks of the clinician will vary from patient to patient and from case to case, there are nonetheless some duties that all clinicians have which are instrumental to the discovery and promotion of patients' values and their overall wellbeing more generally.[6] I expand much more in Chapter 4 on what these duties entail, what it means for clinicians to be able to fulfill them, why they matter, and whether or not they can be fulfilled by non-human clinicians (namely, by artificially intelligent ones). What follows in this section is a mere introduction to what these duties are and how they are connected to value promotion.

1.4.1 Communication

To promote patient values, clinicians must first understand what these values are. The way to come to understand what is truly important for a patient and what will allow them to have a sense of wellbeing is through proper communication with them. As I mentioned previously, blindly fulfilling patients' requests is by no means a goal of medicine, so I am in no way suggesting that clinicians only listen long enough to know what the patient wants. Nor, additionally, am I suggesting that listening is the only communication-related duty of clinicians. What I am suggesting is that clinicians meet the standards of what constitutes good communication well enough so that they may come to better understand who the patient is as a person, what it is that may be bothering them, what things are of importance to them, and what it is that they need in order for their wellbeing to be restored or maintained.

[6] These duties are not exhaustive. They are to be thought of as "initial" duties to which further ones may be added.

Attentive listening is, of course, one way to establish good communication, and though this may be thought to be too time-consuming for clinicians, Cassell rightly points out that "attentive listening is the most time efficient way." Interrupting or failing to listen properly to the patient is, in fact, inefficient, especially when trying to determine what the problem is. As he very nicely puts it: "it takes longer than 20 seconds for people to tell their story...give the poor souls a chance, and they will tell you what is important" (Cassell, 2012, pp. 109-110).

One concern that may be raised here is that listening may not be necessary with technological advancement. That is, listening as a means of understanding patient values may no longer be necessary if, in the future, we come to have AI in the medical field that is able to determine patients' values by analysing all the data it has on them. Thus, one may ask, just because the duty of listening may be instrumental to discovering patient values now, must we also necessarily conclude that this will remain true in the future even if AI systems have enough data about patients that they are able to make inferences about their values without actively listening to them?

I will not speculate here about the plausibility of AI making accurate inferences about patients based on prior data, and I will set aside for the moment all the privacy concerns that will need to be addressed with this type of technology in existence. My response to this concern will, instead, rely on emphasizing two important points regarding communication: The first is that communication includes components beyond mere listening or the collection of information (Kee et al., 2018, p. 99). Listening, particularly for the purpose of acquiring information, is undoubtedly an important component, but perhaps equally important are non-verbal aspects to communication, such as eye contact, facial expressions, tone, etc., as well as attitudes (e.g., conveying respect and empathy towards the patient while communicating with them), and the very content of the verbal communication (what is shared with the patient and the manner in which it is shared).[7] The second important point to emphasize is that while one benefit of good communication has to do with better understanding the patient, it is by no means the only benefit. In addition to allowing clinicians to better understand their patients, good communication also has beneficial therapeutic effects on patients and allows for the development of a stronger patient-clinician relationship (Jagosh et al., 2011, p. 372). A stronger patient-

[7] I discuss these components of communication and AI's capacity to meet the standard for what constitutes good communication in greater detail in Chapter 4.

clinician relationship is created through the fulfillment of other duties that I discuss here, such as compassion and creating trust, which improper communication weakens.[8] Thus, even if AI becomes advanced enough that it is able to infer patients' values by analyzing the data it has on them, this will not be sufficient to override the importance of communication and the effect that it has on the patients' overall wellbeing, the patient-clinician relationship, and the promotion of values. In other words, if AI turns out to fall short in other aspects of communication, then regardless of the data it has on the patient, the effects of its weak communication on the patient and the patient-clinician relationship would nonetheless cause it to fall short of excellence in the practice of medicine.

1.4.2 Recognition

Often traced back to Hegel, the idea of recognition refers to the way an individual is identified and responded to by others. This recognition (or misrecognition or lack of recognition) by others determines an individual's own view of themselves and their own identity. As Charles Taylor explains, "my discovering my own identity doesn't mean that I work it out in isolation, but that I negotiate it through dialogue, partly overt, partly internal, with others" (Taylor, 1992, p. 34). Simply put, it is in being around and communicating with and being recognized by other people that we form our identity and put together our self-image.

Illness can often put patients' own identities into question. They may find it difficult to view or present themselves as persons when they feel like just another patient or just another body. A clinician's duty, then, is not only to view patients as persons, but to also find a way to ensure that the patients acknowledge that they are being recognized by the clinician as such. It is important, in Cassell's words, to "find a way to give them the status they have or need so they can know they are being seen as somebody" (Cassell, 2012, p. 85). It is in being recognized as individuals with personal narratives, beliefs, and values that patients will be able to retain their identity, and thus, their values, which when discovered by clinicians, can be promoted. If patients are not recognized in this sense, but are rather made to feel as though they are just 'patient number 362678', then they risk losing sight of the fact that their values matter and are worth having and holding on to. And without values to hold on to, the patients will have no values that

[8] See Chapter 4 for a discussion of the connection between communication and these other duties.

may be promoted by clinicians. They will likely lack a sense of wellbeing without understanding why or what is needed to restore it.

Of course, many patients may be recognized by others around them (such as their families), without necessarily needing that sense of recognition from a clinician. Still, others (such as patients in Alternate Level of Care (ALC) units awaiting long-term care placements who receive few or no visitors during their time on the unit) will lack that recognition from other sources and will very much require it from those with whom they most come into contact – their healthcare providers. The fact that there do exist such patients makes recognition a duty worth listing here.

1.4.3 Respect for Autonomy

Promoting patient values means respecting patient autonomy. Autonomy is not only intrinsically valuable, but also instrumentally valuable to discovering and promoting patients' other values. Valuing something generally involves choice – if I value my independence, then I am choosing it over dependence on others in some instances, even if the latter may make some things easier for me. If I value playing a certain sport, then I am choosing it over other things that may conflict with my playing it. What's more, in order to truly value something (as opposed to merely desire it), we need to be aware of the alternatives over which we are choosing the thing we value, and the ultimate decision of what we value and what matters most to us should be the one we arrive at after being informed of our options. Thus, promoting patients' true values would not be possible without respecting their autonomy.

According to Beauchamp and Childress, respect for autonomy involves two obligations. As a negative obligation, "autonomous actions should not be subjected to controlling constraints by others", and as a positive obligation, respecting autonomy "requires respectful treatment in disclosing information, probing for and ensuring understanding and voluntariness, and fostering autonomous decision making" (Beauchamp and Childress, 2001, p. 64). What this means in terms of clinicians' duties is that 1. Patients are provided with all the relevant information necessary to make a decision; 2. Clinicians clarify the information, ensuring that the patients understand it clearly, and to make themselves available to answer any questions that the patients may have; and 3. Clinicians must ensure that the patients' decisions are, in fact, their own and that they are in line with, or help to promote, their own values.

1.4.4 Compassion

When I speak of the duty of compassion, I am speaking of a combination of two things: The first is empathy, or the willingness to temporarily imagine oneself in the patient's situation, which, in addition to having positive effects on patients' overall wellbeing (Halpern, 2003), allows the carer to better understand them - their interests, their goals, their needs, and their values. The second is having a desire to promote their wellbeing given this better understanding of their situation. The reason the latter is important is because an objection raised by critics of empathy[9] is that empathy in itself does not lead to better being able to help patients, because it is possible that a better understanding of someone's vulnerabilities could also provide one with better information on how to increase their suffering (which is often the basis of torture). So in order for a clinician to better promote a patient's wellbeing, these two duties must be fulfilled (together, we can refer to them as the single duty of 'compassion').

Of course, it is possible that some patients will not want this from clinicians, and would prefer, instead, to have the clinician provide them with only the technical skills and objective advice that they require, rather than forming this sort of deeper relationship with them. In this case, as Dougherty and Purtilo suggest, "a patient can refuse compassionate care, but a physician is still duty bound to offer it" (Dougherty and Purtilo, 1995, p. 431). In addition to my agreement with Dougherty and Purtilo, however, I would also like to emphasize that compassion can come in different degrees – while at the highest degree, this could mean becoming attached to the patient personally, at the lowest degree, it is just understanding the patient and being driven to help. Is very difficult to imagine a patient who wants help from a clinician not wanting some level of compassion even at the lowest degree.

If a patient rejects even this low degree of compassion, then perhaps this is also one of those cases where patients are making decisions that are not based on their true values. Because accepting compassion, as I mentioned above, means accepting that the physician is trying to understand and promote one's wellbeing, rejecting compassion means rejecting the promotion of one's own values and wellbeing, which cannot possibly be something one values (at least not without contradiction).

[9] See, e.g., Bloom, 2016

1.4.5 Creating Trust

Not only is there a positive correlation between patients trusting in their clinicians and having better overall self-reported health (Hall et al., 2001, p. 629), but creating trust also contributes to the patients' overall wellbeing in that it is through developing a trusting relationship with their patients that clinicians will be able to gather the information they need to promote their wellbeing and proceed to actually do so. Patients will be much more willing to share personal information and allow themselves to be vulnerable and open about what is bothering them and to accept the clinician's recommendations when they feel that they can trust their clinician. Patients are also more likely to feel confident about the judgment and advice of their clinician (Sparrow and Hatherley, 2019, p.94), adhere to treatment recommendations, and feel that the care they are receiving is more effective when they trust their clinician (Hall et al., 2001, p. 629). There are conflicting accounts of what trust entails (both generally and within the context of the patient-clinician relationship), and these range from reliability (London, 2019) to good will (Baier, 1986). Discussing different accounts of trust and what it means for a clinician to be able to "create trust" on those accounts is a task I take on in Chapter 4. For now, it is sufficient for our purposes here to simply acknowledge "creating trust" as an important duty of clinicians and a significant component of what constitutes excellence in the practice of medicine.

1.5 Further Remarks

I will be returning to these five basic duties in the fourth chapter to unpack them and further discuss what it means to fulfill each one. What I have attempted to do here is simply provide a list of some initial duties of clinicians and explain why fulfilling these duties will be necessary in promoting an important goal of medicine. Once again, this list is not exhaustive, but the important thing to keep in mind is that these duties will hold regardless of what patients' values turn out to be or what promoting them will require. These initial duties cannot be removed without resulting in a failure to excel in the practice of medicine. This is because failing to perform these duties means risking failing to properly understand patient values, and a failure to understand patient values leads to a failure to promote patient values, which translates to a failure in restoring their functioning and wellbeing, which as I hope to have shown above, is a component of what constitutes excellence in the practice of medicine.

This list of formal duties provides us with a starting point for evaluating the current and future status of health care. Simply put, if these initial duties are not capable of being fulfilled in a healthcare system, then we have good reason to conclude that the system is falling short and, thus, needs to be prevented (if it is a future system that is under discussion) or changed (if it is a current system at the time of discussion). In later sections when we come to consider how different kinds of AI may change the medical practice, it is important to keep in mind that the duties I have discussed here will hold even through serious technological changes in medical techniques, and an AI-dominant or even an AI-assistive healthcare system will need to satisfy them nonetheless. This means that if we find that the implementation of certain kinds of AI will result in a failure to satisfy these duties, then we have strong moral reasons for either finding a way to fix this or refraining from incorporating this type of AI in the field of medicine altogether.

Chapter 2: The Neo-Luddite Approach

2.1 Introduction

To set off on examining the best ways of incorporating AI in healthcare, it is helpful to begin by discussing a position that falls on one extreme end of that debate with what will be referred to as the *neo-luddite* position, according to which, we ought to refrain from implementing AI in healthcare altogether. In this chapter, I look at some of the arguments, in the form of concerns about the implementation of AI in healthcare[10], that may lead someone to take a neo-luddite position. I then proceed to evaluate these concerns to determine whether the neo-luddite position has merit.

It is worth making clear at this point my acknowledgement of the fact that the positions on AI in healthcare as they currently exist in the literature often exist on more of a spectrum, rather than being capable of being placed in clearly labelled boxes (i.e. "neo-luddites"). For the purpose of this chapter, however, I primarily rely here on Melvin Chen's classification of the literature on AI in healthcare into 3 camps: proponents of the neo-luddite approach, proponents of the assistive approach (who hold that AI ought to be implemented in healthcare, but merely as a tool to assist, rather than replace, human health care providers), and proponents of the substitutive approach (who hold that that the goal of implementing AI in healthcare ought to be to ultimately take over the roles of, and thereby replace, human health care providers) (Chen, 2020, p. 246). Though this is a very simplistic classification which may fail to accurately represent or properly situate several bioethicists and AI ethicists, for the purposes of our discussion, its simplicity allows for a helpful starting point.

I will take up some concerns with the assistive and substitutive approaches in Chapters 4-6. The objective of this chapter, however, is purely to present and evaluate the neo-luddite approach so that we may determine whether it is a view that ought to be taken more seriously or whether it is one that we may safely set aside as having sufficiently rejected, thereby giving us an

[10] Some of these concerns are already noted and responded to by Chen (2020), while other concerns are borrowed from Char et al. (2018), Asai et al. (2019), and Sparrow and Hatherley (2019).

opportunity to narrow down our scope for the remainder of this dissertation on the other two approaches.

In what follows, I have broken down the neo-luddite arguments into three categories: In the first category, presented in section 2.2, I consider arguments that might appear in the literature, but which are quite weak and can be quickly disregarded. In the second category, presented in section 2.3, are arguments that do pose legitimate concerns about AI in healthcare, without necessarily leading to neo-ludditism. Arguments that fall in this second category include the worry of *biased algorithms, decreased communication with patients*, *confidentiality,* the *resurgence of paternalism,* and *the potential of unethical design.* In the third, and for our purposes, most important category, presented in section 2.4, I consider what I take to be arguments that offer the strongest justification for the neo-luddite view. These are arguments which, if accepted, would provide strong reasons for opposing the implementation of AI in healthcare altogether. Arguments in this category include *overdependence on AI* and *inscrutability.* However*,* even these arguments, I will go on to demonstrate, are not sufficient to justify making the types of sacrifices that the neo-luddite approach would call for.

2.2 Weak Dismissible Arguments

2.2.1 The Status Quo Argument

The first objection a neo-luddite may raise against the implementation of AI in healthcare, as noted by Melvin Chen, is what he calls *the objection from status quo*. According to this objection, medical problems can and often are solved without AI technology, as evidenced by a large part of the developing world with minimal access to AI in healthcare due to lack of sufficient funding. In these parts of the world where medical systems and technologies are outdated, it appears that clinicians' ability to diagnose, treat, prevent, and monitor disease nonetheless persist. So, a neo-luddite may ask, why ought we to implement or continue the use of new AI systems in healthcare when other parts of the world appear to be getting on just fine without them? (Chen, 2020, p. 251).

To be charitable to the neo-luddites, we can take their concern to be not arbitrary fear of change but fear of technology's unpredictable nature. That is, oftentimes, we may not know the

consequences of a given piece of technology until many years after its deployment, by which time it is too late to rid ourselves either of *it* (due to over-dependence) or of the issues that arise from it. The change, for the neo-luddite, simply comes at too high of a cost, and, importantly, it is one that we do not *need* to pay. Thus, if we are able to treat patients through means that do not require the use of new technology (which is accomplishable, as evidenced by healthcare in the developing world) then we should persist with those simpler means.

Chen responds to the objection from status quo by pointing out that "each status quo is itself a departure from previous status quos", and that becoming habituated to a status quo is not a justification against departing from it (Chen, 2020, p. 251). I would add to Chen's response that, within this dialectic, the claim that this technology produces extra benefits is not up for debate – the very reason for the technology will be to improve upon our current practices and be able to do what we are currently limited in doing in terms of accuracy, efficiency, safety, etc. The increasing use of deep learning, for instance, will continue to allow for the detection of more of what is undetectable by the human eye. Similarly, as Davenport and Kalakota note, while the genetic variants of cancer and their response to new treatments have proven too complex for human clinicians to understand, firms like Foundation Medicine and Flatiron Health aim to help with that (Davenport and Kalakota, 2019, p. 96). Not to mention the benefits that the wide use of advanced surgical robots will have in allowing surgeons to perform their tasks with more precision (and, thus, less invasive incisions) and efficiency. So while we may already have means of diagnosing, treating, and monitoring patients, this does not mean that newer AI technology will not allow us to do this more safely, more accurately, and more efficiently.

The creation and implementation of new technology involves investing a significant amount of time, money, and resources, and if there were not significant benefits to this, it would not be worth the investment in the first place. Thus, neo-luddites relying on the status quo argument must first prove that the apparent benefits of straying from the status quo are illusory, which is far from an easy task.

2.2.2 The Unfamiliarity argument

A second weak argument a neo-luddite may present is the "unfamiliarity argument". This argument revolves around the idea that technology is novel and unfamiliar and many of its short-

or long-term consequences often cannot or will not be known until after they have caused possibly irreparable damage.

I have categorized this argument as a "weak dismissible" one for the reason that, as Chen explains, a neo-luddite in resorting to it would be admitting to arbitrariness and inconsistency given that many technologies that currently exist and that have existed much earlier than AI in healthcare (and which are very much accepted by those same neo-luddites), were also at one time unfamiliar. Some of the technologies Chen has in mind are "pen and paper, eyeglasses, and even generic software applications" (Chen, 2020, p. 247).

A neo-luddite, however, may not be entirely content with this answer. This, they may explain, is because although Chen's response may be true of many inventions and implementations of technology to date, the kind of technology in discussion here, namely artificial intelligence, is an entirely different case and its novelty is much more concerning. This is because, a neo-luddite may argue, what this technology is currently doing and what we have seen that it is capable of is already terrifying in ways that eyeglasses or pen and paper never were, and thus its potential negative effects much more harmful.

Perhaps such an argument could be made, but much more specificity is required. We must first understand what exactly it is that the neo-luddites are concerned about with regard to the implementation of AI in healthcare. I will proceed to take up these concerns in the sections to follow. But an objection *merely* from unfamiliarity is not sufficient on its own.

2.3 Strong Dismissible Arguments

2.3.1 The Bias Argument:

With the weak dismissible arguments out of the way, we now turn to more interesting objections to AI in healthcare, which point to specific potential problems with AI technology rather than vague concerns about change. One such problem concerns bias. There is a worry of racial, gender, socioeconomic, and political biases (among others) existing in, and being the basis for, an AI's medical decision-making (Chen, 2020; Sparrow and Hatherley, 2019). Though this bias can be intentional, oftentimes it is accidental or unavoidable – this could be due to a lack of

studies on certain populations which lead to the AI making decisions for these populations that do not fit them. Char et al., for example, explain that "attempts to use data from the Framingham Heart Study to predict the risk of cardiovascular events in nonwhite populations have led to biased results, with both overestimations and underestimations of risk" (Char et al., 2018, p. 982).

The bias could also be due to the system being fed already-biased data, the bias behind which having gone potentially unnoticed by its designers. This kind of bias has especially been noted about COMPAS, a software used to predict risks of recidivism among past offenders, and which has been criticized for overpredicting the rate of recidivism among African American defendants while underpredicting it for White defendants (Angwin et al., 2016). The same is the case with predictive policing, where the data fed to the algorithm is tainted by racial bias due to the discriminatory practices already exhibited by law enforcement (Simmons, 2016, pp. 980-982). Bias could also be due to the AI making connections that ought to have been disregarded, such as strong racial bias being exhibited by an algorithm aimed at predicting the healthcare needs of patients, due to its connecting an individual's need for healthcare with the amount of money that an individual has previously paid for healthcare (Obermeyer et al., 2019, pp. 447-453). Thus, a neo-luddite may argue, these instances of bias (as well as the potential future bias) of AI are severe enough to warrant the prohibition of AI entirely from healthcare.

Chen attempts to provide a few responses to this objection. One response involves there being a need for inductive or learning bias in AI. With machine learning, he explains, bias is in part what allows for the correct classification of training examples and learning from the datasets to take place and that oftentimes we will need to accept that there will have to be some trade-off. He uses the example of COMPAS to illustrate that with a recidivism predicting algorithm, for instance, there has to be a trade-off between i) fairness in predicting that a defendant of any race who is high risk will be rearrested; (ii) not predicting that too many low-risk offenders of a particular race are actually high risk; and (iii) not predicting that too many high-risk offenders of a particular race are low-risk. He explains, referencing a paper by Alexandra Chouldechova (2017), that there is a different measure of fairness for COMPAS for each of these and they cannot all be simultaneously satisfied (Chen, 2020, pp.248-249).

It is not entirely clear, however, how Chen's response of pointing out that some degree of bias is inevitable in AI because of a trade-off is intended to be a response to the neo-luddite objection. I imagine Chen's response is meant to imply that the benefits of the AI's learning bias justify (or at least *could* potentially justify) the negatives – that while, for instance (using the example of COMPAS), there may be some low-risk offenders of a particular race who are mistakenly classified as high-risk, it is, on the whole, better to have the algorithm to predict recidivism than not to have it at all.

Still, his response appears to me to be unlikely to satisfy the neo-luddite, for whom there remains a worry about too many instances of this trade-off resulting in dangerously negative trends, reinforced stereotypes, and contributing to worsening the very issues we ought to be working on fixing. It seems that a neo-luddite would simply disagree with the benefits of AI's learning bias justifying the ends and maintain their stance against the implementation of AI in healthcare. Consider, for instance, Obermeyer et al.'s examination of an algorithm that is intended to identify and help patients with complex health needs, but which has been demonstrated to have a strong racial bias. Obermeyer et al.'s results show that while "Black patients are considerably sicker than White patients as evidenced by signs of uncontrolled illnesses", only 17.7% of patients that the algorithm assigned to receive extra care were Black (which should have been closer to 46.5% if the algorithm was unbiased). The reason the algorithm resulted in providing much fewer recommendations to Black patients is because it relied on healthcare cost as a predictor of healthcare need, and because the care provided to Black patients cost on average $1801 less per year than it did for White patients, the algorithm translated this decreased cost into a decreased need for care (Obermeyer et al., 2019, pp. 447-453). The bigger issue with this, however, is that this results in unfairness for patients who were already at a disadvantage, further widening the gap. Thus, it seems to me that the neo-luddite can easily respond to Chen by explaining that this "necessary trade off" is exactly why we ought to steer clear of implementing AI in healthcare. The best way to avoid the negative results is to avoid the technology altogether.

A second response given by Chen to the concern of bias makes use of the System 1/System 2 distinction popularized by Kahneman (2011). Chen's response revolves around the idea of there being instances in which bias in decision-making can be justified, not just for AI, but also for humans. Humans have two systems of thinking: "System 1 thinking (automatic, fast,

and unconscious) uses heuristics or rules of thumb whereas System 2 thinking (controlled, slow, and conscious) requires analytical effort" (Chen, 2020, p. 249). So when time and sample sizes are limited and cognitive resources are scarce, using heuristics or rules of thumb, though they contain bias, are beneficial. Thus, in those instances, System 1 proves superior to System 2 thinking, and similarly, a biased algorithm may prove superior to an unbiased one (ibid.).

It is unclear, however, whether this response will satisfy the neo-luddite, either, who may argue that the faults or biases in an algorithm are not quite like human physicians having to resort to their System 1 thinking when time, information, and resources are limited. This assumes that an algorithm will only produce biased results when input or data is limited. However, one worry with biased AI is that System 1 thinking will be deployed in instances where System 2 thinking is necessary (and possible) and that despite an abundance of data, an algorithm may produce outputs based on irrelevant factors – in other words, that there will be faults in the outputs of algorithms that human physicians *would* be able to spot. This could mean negatively affecting many patients by, again, re-enforcing and contributing to socioeconomic and racial divides (among other things) and further negatively affecting the wellbeing of those who are already disadvantaged. What's more, as noted by Sparrow and Hatherley, there is an added worry about the outputs of the AI influencing the subsequent data used to train it (Sparrow and Hatherley, 2019, p. 91), thereby resulting in the creation of even further bias.

Chen does point out that there are instances where humans, too, deploy System 1 thinking when System 2 thinking would be superior. Again, however, it is not clear that this will comfort the neo-luddite, who will then point out that not only is there a fear of much more instances of inappropriately utilizing System 1 thinking when System 2 thinking is needed where the use of AI is concerned, but it remains debatable whether AI is, or will be in the near future, capable of System 2 thinking to begin with. Thus, a neo-luddite may maintain, even if there are instances where System 1 thinking trumps System 2 thinking, there are nonetheless instances in which it does not, and the potential of the occurrence of too many mistaken instances of deploying System 1 thinking when System 2 thinking is necessary could be detrimental, and so the concern of bias remains and the stance against AI in healthcare remains along with it.

Chen's third and final attempt to respond to the objection from bias centres around value-sensitive design.[11] He notes that, since AI technology embodies our values, the elimination of unjustified biases could be included as part of the design approach (Chen, 2020, p. 250).

I am, however, also skeptical about this response's ability to satisfy the neo-luddite. It appears to me that the worry is not simply that the designers will fail to account for appropriate values or that the values that the designers input into the system will be problematic. Nor is the worry that biases will be purposely incorporated into the AI or that the designers will notice biases and refrain from taking action to correct them. The worry is that these biases will not be accounted for at all and will lead to outputs that go for a long time unnoticed, unquestioned and accepted, despite negatively affecting many patients and only coming to light (if at all) after significant damage is done. Designers can have the best of intentions, aim to minimize unjust biases, and incorporate strong values into their design, but this still may not be enough to guarantee that the system will be entirely free of unfair biases.

With that said, are the neo-luddites right to insist that fear of bias justifies refraining from the use of AI in healthcare? I don't believe so. Unjust bias and its implications are something to be very much concerned about, but the solution ought not be to refrain from using technology in healthcare altogether. This is because, if we take the neo-luddites to be against *any* amount of bias risk, then it is worth pointing out that there is bias among human doctors, as well, but they are certainly not prohibited from practicing medicine. Albisser Schleger et al., for instance, touch on the many ways that bias can come into play in individual as well as group clinical decision-making (Albisser Schleger et al., 2011, pp. 156-159). In group decision-making, for example, process loss[12], group think[13], and normative social influence[14] are commonly at play, while in

[11] Value sensitive design is an approach that calls for incorporating values and moral considerations throughout the design process. See Friedman and Hendry (2019); Friedman, Kahn, and Borning (2007).
[12] The tendency to focus on information known by the majority of group members, while disregarding information known only by a few of the members.
[13] A phenomenon in which group members, in aiming to achieve harmony with each other and under pressure to conform with one another, end up refraining from raising views, concerns, or solutions that diverge from those of the group.
[14] This occurs when an individual alters their behaviour in order to conform with and be accepted by another individual or group.

individual decision-making, omission bias[15], belief in a just world[16], and stereotypes and prejudice are among other biases at play.

Thus, if a neo-luddite is content with allowing human decision-making in medicine, this tells us that it is unlikely that a neo-luddite is uncomfortable with *any* level of bias risk, given the bias risk involved without technology. If, on the other hand, the neo-luddite is comfortable with a certain level of risk, granted that the risk of bias is not greater than it is with humans, then this matter can be taken care of by setting policies and guidelines for the design and use of AI systems to ensure that the benefits justify the risks or that the risks are kept to a minimum. At the very least, however, the potential of bias risk in general would not be enough to justify prohibiting the implementation of AI altogether.

An example of a policy to deal with bias risk of AI would be one that requires the presence of humans in the loop of the algorithmic decision-making, to detect and account for any potential biases before the algorithm's decisions can cause too much damage. With humans in the loop to verify that an algorithm's decisions are based on appropriate factors within the data, the worry is lessened. So, simply put, it appears that the neo-luddites are correct to worry about AI if it is left to its own devices or trusted completely and blindly, and so they are correct to fear a substitutive type of AI, but, again, I believe the concern is lessened with humans in the loop. Of course, there is some reason to worry that it will not always be possible to keep humans in the loop. I discuss this under the heading 'The Objection from Inscrutability' (see Section 2.4.2).

It is also worth pointing out that by "humans in the loop", I am not strictly referring to the designers, for as we have established, designers may mean well, but they are not always able to detect potential biases. Of course, human clinicians, too, are not perfect and, as was discussed earlier, they, too, may not always be able to detect biases even in their own thinking. We can also grant that when the datasets are so large, the ability to detect biases in them becomes even more difficult. However, the combination of value sensitive design along with human clinicians that are able to judge an algorithm's decision-making (and take it as a recommendation rather than blindly trusting it) can allow for instances of bias to be detected and corrected much more quickly. Consider, for instance, an algorithm that has not been trained well enough on how

[15] The tendency to evaluate taking action as more risky than non-action.
[16] The tendency to believe that consequences are earned – e.g., believing that a patient's medical problems are a result of their own behaviour.

members of a certain race may react to a particular treatment option, and thus produces recommendations for a treatment that may not be effective for patients of that race. A few mistakes may be made at first, but very quickly human clinicians using this algorithm will consult their colleagues and the medical journals and, using their own empirical findings, be able to detect the algorithm's error and report the issue and refrain from accepting its recommendations when treating patients of races that the algorithm has not been trained well on.

Even if it is not perfect or entirely unbiased, it at least can be argued that the AI does not make things *worse* so long as we ensure that the final decisions are still left up to the humans and that the human clinicians are aware of the potential of biases in the system and that the system's recommendations are not trusted completely, but only taken as recommendations that may require further review. There is perhaps still the worry of the humans in the loop blindly trusting in the AI, but this, it seems to me, to be a different issue, which I take up in Section 2.4.1.

For the neo-luddite to have grounds for prohibiting AI in healthcare, they would need to demonstrate that the bias in AI results in worsened outcomes than would exist without its presence in healthcare. An analogy here will perhaps clarify my conclusion for why this need not be the case when AI is implemented in healthcare: Imagine that every GPS has a bug where, every once in a while, they fail to detect ponds and lead users onto them as though they were roads. In this case, we might say that as long as the final decision remains up to the driver and their knowledge about ponds not being car-friendly, the existence or use of a GPS would not necessarily be "worse" for the driver. Granted, it may be inconvenient to have to reroute every once in a while when the GPS fails to detect a pond, but the bug would not necessarily mean that the driver is, on the whole, better off without using a GPS. It would, however, be a much worse issue if self-driving cars failed to detect ponds sometimes and attempted to drive into them anyway. Similarly, if humans are in the loop to make the final decisions, biases in AI may be problematic, but they need not be said to make things worse.

2.3.2 Decreased Communication with Patients

Another neo-luddite concern, raised by Asai et al., has to do with the possibility of the presence of the AI technology in healthcare resulting in clinicians having decreased communication with patients. The worry is that clinicians will be so preoccupied with the AI

services and the data that they hold about the patients, that they will neglect observing and communicating with the very patients before them. There is a fear, they explain, that in the future, clinicians will only look at data from patient records and from wearable health devices, without taking the time to listen to how a patient sounds, what their facial expressions are like, how they walk, or even examining their body for themselves (Asai et al., 2019, p.66).

The short and sweet response to this worry, however, is that while this is a concern to be cognizant of, it certainly need not lead to such an extreme view as prohibiting AI altogether. There are medical guidelines that can be put in place to ensure that this does not happen – these guidelines can be as simple as requiring that the doctor observes and consults with the patient *in addition* to the AI prior to making any important decisions about diagnosis or treatment recommendations.

2.3.3 Confidentiality

A neo-luddite may also be concerned about the issue of privacy and confidentiality. In particular, there are worries of algorithms using large amount of patient data in various, oftentimes unknown, ways over long period of time, as well as worries of breaches of patient data (Kruse et al., 2017). Because of the reproducible nature of electronic data, there is also an added worry about the information becoming available to those outside the healthcare organization (Sparrow and Hatherley, 2019, pp. 88-89), such as governments, employers, or insurance companies, who may utilize sensitive information about patients to further their own ends at a cost to the patients. To add to this, I imagine a neo-luddite would also be concerned with the issue of fairness – that patients will have to choose between protecting their privacy and protecting their own health, and that those who value privacy may be unable to gain access to proper care without allowing access to their medical records, while those who choose their health may have to sacrifice their privacy for it.

My response to the neo-luddite is as follows: even if we consider the worst-case scenario and assume for the sake of argument that there is no way around this issue of privacy, this concern still does not quite justify taking a neo-luddite stance. With advances in technology, we have to trust that there will be advances in policy and data protection to follow shortly thereafter, and a combination of policies of informed consent for what the information can be used for

(ranging from only the necessary to allowing open access to them for a range of uses) and proper methods for the encryption and anonymization of patient information[17] will lessen the severity of these privacy risks. This combination may still not allow for perfect protection, but it is highly doubtable whether perfect protection is, or has ever been, a possibility. Physical medical records can be stolen, doctors can break confidentiality, and information sharing is always a possibility. Though the form of the privacy risk may be different, the occurrence or frequency need not be higher if proper measures are taken to keep up with the technology.

With regard to patient consent, patients need not necessarily choose between privacy and *any care at all*. While the biggest risk to privacy appears to come from not knowing how or for what purpose a patient's information will be used, a patient can withdraw consent from their data being used for any purpose that they are not comfortable with. This may perhaps come at the cost of algorithms not being able to learn from their profile and better serve their health needs in the future, but in that case, the patient would simply have to resort to the advice of the clinicians sans the individualized AI expertise.[18]

2.3.4 Resurgence of Paternalism

The penultimate strong but dismissible neo-luddite concern concerns the resurgence of paternalism. Asai et al. worry that AI in healthcare "will revive paternalistic attitudes among physicians, as AI may give physicians more "objectively correct" diagnoses, recommendations, and judgments concerning patient care" (Asai et al. 2019, p. 66) The idea behind this concern is that if AI affirms the physician's diagnosis or treatment option, then the physician will come to see it as more objectively correct and will, thus, be more likely to come to see a patient who refuses the plan or treatment option as objectively wrong or "foolish". They believe that this will lead physicians to be "more determined to manipulate, coerce, or excessively persuade patients to accept the 'best plan,' regardless of their intentions or hopes" (ibid.).

[17] See Wu et al. (2022) for an example of what such a method might look like.
[18] There is a worry that clinicians may refuse to grant patients this option, thereby turning them away from care or insurers refusing to offer malpractice insurance to doctors who do not require patients to consent to the AI recommendations. This, however, is also a matter for policy.

Asai et al.'s concern, however, is not really an issue with technology, but rather an issue with the potential attitude of clinicians, which can exist with or without technology. Some clinicians will maintain that their patients are "foolish" for disagreeing with them or for refusing certain treatment options, regardless of whether or not these are reaffirmed by machines. They could be, for instance, reaffirmed by other clinicians whose expertise they trust. Even if the neo-luddite's concern is that the presence of AI *increases* paternalistic tendencies among clinicians due to the temptation to view AI to be more competent or reliable than any particular human, this need not lead to the conclusion of prohibiting AI in healthcare, for there are ways to guard against this that do not involve ridding the entirely of the healthcare system of these machines. Ethics training for clinicians and informed consent policies that guard against coercion or manipulation are a few such ways to deal with this.

2.3.5 Unethical Design

The final strong but dismissible objection concerns the potential unethical design of the technology. Noting that "the builders and purchasers of machine-learning systems are unlikely to be the same people delivering bedside care", Char et al. explain that this would mean potentially programming clinical decision-support systems "in ways that would generate increased profits for their designers or purchasers (such as by recommending drugs, tests, or devices in which they hold a stake or by altering referral patterns) without clinical users being aware of it" (Char et al., 2018, p. 982).

Though this is certainly a valid concern, it is important to also acknowledge that with or without technology, there is always pressure on healthcare facilities to cut costs and increase profits, and oftentimes this need to improve value metrics comes at the cost of providing good care to patients.[19] So this concern is certainly not unique to technology in healthcare. Though a neo-luddite may argue that technology worsens the problem or makes it easier for good care to be sacrificed for profit (and that the manipulative aspect of it, in which there is a lack of

[19] The controversial application of the "lean system" to healthcare is one example of this. See Thompson (2023).

awareness on the part of the clinicians that the decisions of the machines are geared towards generating profits, makes it further unethical).

But that this is a concern does not mean that we ought to refrain from implementing technology in healthcare completely. What this concern does mean, however, is that we need to be careful with the rules and monitoring of this technology when it is implemented. We need to guard against the potential unethical design of this technology by, for instance, providing ethical training, performing regular ethics audits, and forming ethics boards (consisting of various stakeholders, including clinicians, and third-party members without a direct vested interest) to evaluate the technology prior to its implementation and to ensure that it does not result in ethical drift when in use. Once again, this concern need not lead us to fear the implementation of technology in healthcare altogether.

2. 4 Forceful Arguments

2.4.1 Over-dependence on AI

While bias, decreased communication, confidentiality, paternalism, and unethical design are very real concerns, they do not necessarily lead to the drastic conclusion that we ought to prohibit the implementation of AI in healthcare altogether, and consequently, do not quite validate the neo-luddite approach. With those out of the way, then, I turn to a final set of concerns which, if accepted, would support the prohibition of AI in healthcare.

One such worry has to do with the possibility of over-dependence on the technology (Sparrow and Hatherley, 2019, pp. 95-96). This is also referred to as the issue of *de-skilling*. In a conversation with Lindsey Bordone, an attending physician at a dermatology clinic, Siddhartha Mukherjee raises a concern about whether humans will come to learn less and less the more machines are in use and able to do the learning for them, using the examples of spell-check functions leading to children's inability to hone their spelling skills, and automated cars leading to less alert drivers. The same, Mukherjee worries, may be true of medicine. Using Bordone's work as an example, Mukherjee takes note of "how seriously [Bordone] took every skin tag and

mole that she ran her fingers over", but questions whether that would still be true "if she partnered with a machine" (Mukherjee, 2017).

Similarly, Asai et al. explain that reasons such as patients demanding the use of AI, a physician not being confident enough in their own decisions, and fear of misdiagnosing or prescribing an inappropriate treatment option, may lead to over-dependence on and the over-use of AI. But if this happens, the physician's skills become damaged, just as the use of autopilot damages a pilot's steering skills. They additionally predict that this will mean that physicians will spend more time developing skills for operating AI than developing actual clinical skills, because those will end up yielding more benefit for them (Asai et al., 2019, pp. 65-66). Thus, a neo-luddite may worry about the negative side effects of the majority, if not all, clinicians failing to hone their clinical skills, or worse, failing to develop clinical skills altogether, due to over-dependence on AI. As we know, machines can malfunction and simply are not as reliable or infallible as we would like to think. Just as it is always smart to keep candles around the house in the event that the power goes out, it is always important to ensure that clinicians maintain their clinical skills even after these skills come to exist in machines so that if the machines for any reason malfunction, err, or lose their power source, we can feel confident leaving the patients in the hands of human clinicians without worrying about any potential harms befalling them. Additionally, with over-dependence on AI, there is a further worry of an inability to detect or correct failure in AI diagnosis or treatment recommendation if clinical skills are lost in human clinicians.

One response to the objection from over-dependence on AI comes from Chen, who simply suggests that "we could be careful in our implementation of medical AI technology, ensuring that the level of professional competence and quality of healthcare do not dip in the face of technological changes" (Chen, 2020, p. 251). Bringing in the value-sensitive design approach again, he recommends designing AI technologies with these values in mind.

It is unclear, though, how much Chen's response would satisfy the neo-luddite, whose worry seems to be that it is easy to get carried away in our dependence on AI and that it is not always clear where to draw the line. Asai et al., for instance, suspect that some physicians may delegate the diagnosis and treatment of patients to AI even when it is completely unnecessary and can be done through medical interviews or physical examination (Asai et al., 2019, p. 65).

There is no easy way to deal with this objection because it is very much a justified concern. There is not always a clear line to draw between what will allow for efficiency and what will result in over-reliance and long-term disadvantages. For now, I would like to offer the following response to the neo-luddite: In addition to setting guidelines about the design and use of AI in healthcare and requiring clinicians to be educated about their limitations, we must weigh the risks vs. the benefits of the technology in order to determine whether terminating it altogether is justified. The lack of technological advancement in healthcare (or the de-technologizing of the current healthcare system) will result in much fewer lives saved[20], much longer procedures, and much more tired and burnt-out physicians. This weighed against the possibility of over-reliance on the technology should hopefully make it clear why it might nonetheless be better overall to risk the (preventable) risk of over-dependence for the advantages. There are many things that we have become dependent on in our daily lives that we would not change – such as becoming so dependent on a contacts list saved on our phones that many of us are incapable of remembering the phone numbers of our closest friends and relatives (which is certainly disadvantageous in an emergency when we are lost without our phones and need to make a call from a stranger's). Still, we tend to think that the advantages are worth the over-dependence on the technology.

It is important to note that I am not suggesting here that over-reliance on technology is justified. There is certainly a limit to this and a limit to what kind of technology we may allow clinicians to be dependent on or which skills to stop practicing and honing. There are some AI technologies that it will not be harmful to be dependent on, much like how it is not considered harmful to become accustomed to driving an automatic car and forget or never learn how to drive a stick shift. There are technologies that it is harmful to be dependent on only in the instance that the technology fails. And there are technologies that it is harmful to be dependent on altogether because they result in the loss of an intrinsically valuable skill or virtue (what Shannon Vallor (2015) calls "moral de-skilling"). But when the point of debate is whether there should be no technology at all in healthcare or whether there should be some technology at some potential (preventable) risk, then the answer to me seems clear: we ought to be careful in how we approach reliance on technology, ensuring that important technical and moral skills continue to

[20] Cancer, for instance, which accounts for approximately 16% of deaths worldwide, is predicted to be significantly decreased by utilizing AI to predict, more accurately diagnose, and curate individual treatments for cancer (Sebastian and Peter, 2022).

be practiced by clinicians even when technologies may appear to make these skills redundant (until we are actually certain that such skills really are unnecessary to continue having or honing). With regard to intrinsically valuable skills and virtues, our goal ought to be to use the technology precisely to improve, not hamper, such skills and virtues. Most importantly, however, we must not be afraid of implementing AI altogether for the reason that it risks over-dependence of clinicians on it, because the benefits that we would lose out on with this course of action make the approach difficult to justify on the basis of such an insufficient reason.

2.4.2 The Objection from Inscrutability

Another issue with AI in healthcare that may worry the neo-luddite has to do with the fact that AI can be inscrutable – that is, it can be made to function as a black box, with its method of arriving at its decision being opaque, unable to be understood or explained by humans. This is particularly the case with machine learning, where rather than if-then rules being programmed into a system by humans (like in traditional rule-based AI), the AI learns from patterns in the large amounts of data that it is fed and produces outcomes on the basis of that data. In many cases, however, the decision-making process behind the output of the AI is not interpretable to humans.

The main worry when it comes to inscrutability or opacity in AI decision-making is that our ability to trust someone's decisions is often thought to be connected to their ability to explain the reasoning behind those decision. AI's inability to explain its decisions, then, would make it difficult to accept those decisions as trustworthy or reliable. This becomes especially concerning if the AI's decision-making is unknowingly based on factors that ought not to have been included in the decision-making process (and may lead to biased outcomes) or if the AI's outputs fail to sufficiently take into account or give proper weight to a patient's personal values.

One way to respond here involves "proposing alternative ways of scrutiny [such as] the large-scale design of transparent, trustworthy, and easily comprehensible AI systems in accordance with explainable AI (or XAI) principles" (Chen, 2019, p. 253). The National Institute of Standards and Technology, for instance, proposes the following principles for explainable AI: *Explanation* - a system delivers reasons for the outputs; *Meaningful* – a system's explanations are understandable to consumers; *Explanation Accuracy* - the explanation correctly reflects the

reason behind the output; and *Knowledge Limits* - a system identifies its knowledge limits and the instances in which it is not approved to operate (Phillips et al., 2021, pp.3-5). These principles are intended to allow for more trustworthy AI systems.

This response, however, may not suffice for convincing the neo-luddite. This is because it can be argued that imposing requirements for explainability (in the sense of limiting the use of AI to only white-box models) may get in the way of a system being used to its full potential and, thus, hinder it from being able to produce beneficial results. In particular, the neo-luddite may worry about how realistic it is to expect only systems that abide by XAI principles to be used in healthcare, especially in instances where this would prevent machines with the potential to save many lives from being implemented.

Of course, XAI can also involve post-hoc explainability methods – that is, models that attempt to open up black-boxes and make opaque AI more transparent. For example, Grad-CAM produces a heatmap to highlight the significant features or regions of an image that support a system's output predictions (Selvaraju et al., 2020). Another method, SHAP, uses a game theoretic approach of assigning credits to each feature of a model's predictions (Lundberg and Lee, 2017), similarly to calculating the contributions of each player in a cooperative game by comparing how the coalition would perform with different combinations of players and averaging out the contributions of each player. Post-hoc XAI methods would seem, then, to resolve the neo-luddites' concerns while not limiting the use of AI to only white boxes. Unfortunately, however, post-hoc XAI models are not perfect solutions and do have limitations (Vale et al., 2022, p. 821), particularly with regard to stability (i.e., models generating different explanations for the same instances) and accuracy.

Another response put forth by Chen to the inscrutability problem relies on what he calls the "human-AI interface." He explains that in some instances System 1 thinking (quick and automatic) is superior and thus the AI system will get things right. But in other instances, where System 2 thinking (slow and controlled) would be superior for decision-making, "the AI system (practising System 1 thinking) will slow down and ask for reinforcements from the human healthcare professional (System 2)" (Chen, 2020, p. 253).

A worry does remain with Chen's response, however, about how (or whether) it can be determined by the AI if System 1 or System 2 thinking is needed for a given decision, or whether the AI's decisions were appropriate in a given situation, given the system's inscrutability. That

is, it is possible that whether a problem counts as System 1 or System 2 (and whether the AI was correct in properly identifying it as System 1 or System 2) is itself one of the determinations that an AI will make in inscrutable fashion.

A final response to the concern of inscrutability is to accept the nature of black box systems but not necessarily view this as something detrimental. This is a response put forth by Alex John London (2019), who posits that medicine is a field in which decisions have often been (and still are) based on experience, which comes prior to an understanding of why such a decision or intervention works. This, he explains, is apparent when we look at the fact that lithium is prescribed as a mood stabilizer despite a lack of understanding of how it works. The same is true of aspirin, which had been prescribed as a pain reliever for nearly a century without an understanding of how it worked (London, 2019, p. 17). This is not necessarily a negative thing, and in fact, when the requirement for explainability is prioritized over empirical evidence, "patients suffer, resources are wasted, and progress is delayed" (ibid., p. 18). Thus, the inscrutability of AI systems need not be seen as a problem, so long as we limit our use of such systems to instances where their accuracy is empirically validated using data that mirrors real-world contexts.

Even if the neo-luddite continues to have some doubts about whether AI decisions can be trustworthy despite a lack of explainability, what London's response demonstrates is that medicine sans machine learning also falls victim to similar concerns, and thus, AI's inscrutability is not a sufficient justification for prohibiting its use entirely in the realm of healthcare.

2.5 Conclusion

I have addressed here a few concerns that neo-luddites may reference to justify their position against the implementation of AI in (and the possible de-technologizing of) the realm of healthcare. I considered the weak dismissible arguments of unfamiliarity and the status quo as well as the arguments from bias, decreased communication with patients, confidentiality, paternalism, unethical design, over-dependence on AI, and inscrutability. I have attempted to sufficiently respond to all these concerns in a way that would satisfy the neo-luddite by, at the very least, demonstrating that refraining from implementing AI in healthcare altogether is not the solution to these concerns. By this point, I hope to have provided enough of a reason to justify

setting the neo-luddite position aside for the remainder of this dissertation and focusing on the remaining two approaches: the substitutive approach and the assistive approach, which will be taken up in the chapters to follow.

Chapter 3: Conceptual Analysis of Technological Assistance and Substitution

3.1 Introduction

Understanding the distinction between assistance and substitution as it applies to technology and artificial intelligence is essential to understanding the ethics of implementing different types of technology in different fields. We oftentimes come across worries about the implementation of technology resulting in the replacement of human workers in a given field (or multiple fields) (Semuels, 2020), without much discussion of what *kind* of technology it is that will actually do this. More importantly, there is a lack of discussion about what kind of technology it is safe to implement without fear of accidentally accomplishing this when it is demonstrated to be harmful – that is, where the replacement of humans with AI in some way detracts from the quality of the work or the resulting product or service. It is especially important to be clear on this distinction when dealing with the ethics of AI in healthcare – a field with its own ethics and values and a sense of excellence in the practice that often goes beyond achieving some mechanical task – a field that involves dealing with vulnerable patients with both physical and emotional needs, and a field where technology's assistance and/or substitution can have major implications beyond just the loss of jobs. I will consider these implications in the chapters to follow. First, though, we need to be clear on what the assistive/substitutive distinction amounts to, which is the task of this chapter. This distinction between "assistive" and "substitutive" technology, however, can be quite complicated, especially when we consider some hard cases where this line is not so easy to draw and we become unsure of what kind of technology it is that we are comfortable or uncomfortable with implementing.

The aim of this chapter is to offer a conceptual analysis of assistance and substitution, insofar as the terms apply to technology. In Section 3.2, I present four senses in which the term "substitute" is used, before going on to propose criteria for what constitutes *substitutive technology* in Section 3.3. In Section 3.4, I briefly discuss the concept of assistance. In Section 3.5, I consider some hard cases where a line between "assistive" and "substitutive" technology cannot easily be drawn, before offering a solution to these hard cases in Section 3.6.

3.2 The Four Senses of "Substitute"

We use the word "substitute" in a few different ways. (1) A teacher who feels under the weather may have a substitute teacher cover their classes while they are recovering. (2) A patient who has suffered a serious brain injury after a car accident may have a substitute decision-maker make decisions on their behalf. (3) If Lebron James is injured, another player will have to sub in for him. And (4) a customer may ask a barista to substitute oat milk for the usual 2% milk they would otherwise put in their latte.

These four senses in which we use the term "substitute", while similar in some respects, are nonetheless distinct. That is, they do not all meet the same criteria. Let's begin with the substitute teacher. A substitute teacher is not a permanent replacement, but is only intended to fill in for a short period of time, until the main teacher recovers. The substitute teacher is not authorized to take over all aspects of the main teacher's role entirely – the substitute is not allowed, for instance, to change a student's grade or cancel a test or speak with a student's parents about their child's progress. Moreover, the original teacher, while not directly responsible or accountable for the substitute's actions, must deal with the consequences of the substitute's actions (or lack of action) upon resuming the role. So in this sense of the word, "substitute" can be defined as follows:

(1) *A is a substitute for B if only some (often minor) aspects of B's role are temporarily performed by A, until B is once again able to return to the role*.

The second sense of substitute is stronger than the first due to the nature of the decisions being much more significant. A substitute decision-maker is given the authority to make important (oftentimes, irreversible) decisions on behalf of the patient, and so in some sense, we may say that the patient's judgment and decision-making capacity is replaced for a period of time. Nonetheless, even in cases where the substitute decision-maker has authority over all decisions relevant to the patient's finances and personal care, the substitute does not have unlimited authority and, thus, we cannot say that they have completely taken over the role of the patient. If the patient had previously signed a valid, legally binding DNR, for instance, it cannot be overridden by the substitute decision-maker. Similarly, if the patient regains consciousness

and the mental capacity required for decision-making, the substitute decision-maker is no longer able to fill in for the patient's judgment and decision-making. So while it is a stronger sense of substitution than that of the substitute teacher, it is nonetheless still temporary, the authority remains in some ways limited, and the original agent (though not directly responsible or accountable for the decisions of the substitute) must deal with consequences of the substitute's actions (or lack of action) upon resuming the role. Thus, we can define this second sense of "substitute" as follows:

(2) *A is a substitute for B if A is given authority to make important decisions on behalf of B, but this authority is nonetheless limited and temporary, until B is once again able to return to the role.*

The third sense of substitution, as in that of substituting a player for another, is even stronger than the previous two (that of the substitute teacher and that of the substitute decision-maker) because in this sense of substitution, the substitute's decision-making power in that role is unlimited.[21] If a reserve is subbed in for Lebron James, he can decide for himself which moves to make, regardless of what Lebron from the bench may be telling him to do. In that role, he has completely replaced Lebron, but only temporarily – that is, only until Lebron is able to get back on the court and reclaim the position. Moreover, Lebron, though not directly responsible or accountable for the decisions of the substitute, must also deal with the consequences of the substitute's actions (or lack of action) upon resuming the role (e.g., falling behind on the scoreboard). So this third sense of substitute can be defined as follows:

(3) *A is a substitute for B if A is given unlimited authority and decision-making power in B's role, but only temporarily, until B is able to return to the role.*

The final sense of substitute, as in that of substituting oat milk for 2% milk, is the strongest one. When substituting oat milk for regular 2% milk, the 2% milk is completely

[21] By "role" I mean the player's position on the court at the time of substitution, not outside out of it. A player substituting for another during a game is still unable to, for instance, resign on behalf of the starting player.

replaced – it no longer has any role to play in the drink. This a *permanent* substitution[22] - that is, the barista is being asked to scrap the default option so that a new option may take its place completely and permanently in the latte. Thus, substitution in this fourth sense can be defined as follows:

(4) *A is a substitute for B if B entirely ceases to hold its previous role or position and A permanently takes over all aspects of B's role*.

It is perhaps worth noting that the four instances here are not exhaustive of the ways "substitution" may be used, and that depending on the degree of, say, authority or responsibility, other senses may also be used. These four that I have presented here are merely meant to demonstrate some key distinctions as well as some key similarities.

3.3 Substitutive Technology

Given our lack of a consistent definition that captures the degree of permanence, authority, duties, and responsibility of the substitute, it is unsurprising that we also lack a clear definition for what constitutes *substitutive technology* – that is, technology that substitutes for a human in performing a task or function. In this section, I hope to rectify this by offering, and proceeding to justify and flesh out, an account of substitutive technology. I will first state the overall view, then explain its components. Before proceeding to do this, it is worth noting that the account of substitutive technology that I present here is not intended to capture all instances of technology substituting for humans, but rather the *ethically interesting* kind of substitution.

Here is the overall view: in order to get at the most ethically significant form of substitutive technology, I propose using a definition most closely resembling the fourth sense – that of the milk substitute, in large part due to the aspect of *permanence* that sets it apart from the other senses:

[22] "Substitution" in the sense used here is intended to refer to the substitution as it occurs in the drink itself, not substitution as it appears from the perspective of the barista, who will resume making the lattes with the default 2% milk after the customer's oat milk request has been fulfilled.

A is a substitute for B if B entirely ceases to hold its previous role or position and A permanently takes over all aspects of B's role.

We are now in a position to appreciate some of the complexities hidden in the original formulation of this sense of 'substitute'. I suggest expanding it as follows:

A is a substitute for B if (a) A is perceived to share some relevant characteristics with B which allow A to perform the tasks and functions previously performed by B; (b) B's authority to make judgments and decisions is transferred to A; (c) The duties and, thereby (d), the responsibility and accountability associated with B's role is transferred to A; and (e) this transfer of authority, duties, and responsibility is not temporary.

Next I would like to move on to my justification of this definition: To begin, (a) seems to be quite self-explanatory and uncontroversial – in fact, this is true of all senses of "substitute". In each of the senses of "substitute" discussed above, we can see that the substitute shares (or is at least perceived to share) some relevant characteristics with the person or object for whom they are substituting. Oat milk is a potable liquid that adds flavour and creaminess similar to the 2% milk.[23] The substitute teacher is an adult able and certified to instruct children and manage classroom behaviour, similar to the regular teacher. The substitute decision-maker is someone with competence and some knowledge of the patient and the medical options, similar to what the patient would have access to if they were conscious and competent. And, of course, the substitute basketball player is a professional basketball player with knowledge of the game and similar abilities to the player substituted for. It follows, then, that there is no reason to think that where the technology is involved, it would not also need to share some essential characteristics with the agent it is substituting for – in particular, ones that would allow the technology to perform the tasks and functions of that agent. For instance, AMP Robotics' robotic recycling sorting system (Barker, 2019), which is capable of substituting for human assembly line workers in sorting,

[23] The more shared relevant characteristics, the more fitting of a substitute we may feel that it is. Juice, for instance, would be a poor substitute for 2% milk where a latte is involved, given how it affects the drink (to the point of perhaps resulting in a new drink altogether). Though, we may also say that juice is nonetheless a better substitute than, say, a rock, which not only lacks the quality of being a liquid, but also lacks the quality of being edible, thereby not being thought of as a substitute for 2% milk at all, but rather a completely different addition (perhaps a decoration) in the latte.

picking, and placing objects into the correct material categories, shares with its human counterpart the ability to recognize and distinguish between different materials with accuracy, as well as the ability to move and to pick up objects and drop them in a controlled and intentional manner.

Condition (b) we can also see present in all senses of "substitute" above, but to different degrees. A substitute teacher only has some, but not all, of the authority to make decisions that would affect the regular teacher's class. The substitute decision-maker has a little more authority, but it is still not unlimited. The substitute player has the most authority transferred. And if oat milk and 2% milk could transfer judgments, I would imagine the oat milk would have complete authority to affect the drink that it is poured into and would not be limited in any way that the milk is not.[24] It is perhaps worth noting that conditions (a) and (b) are not unrelated. In fact, condition (b) is connected to, if not directly stemming from, condition (a) in that the substitute is often chosen for their having some relevant resemblance to the agent for whom they are substituting, and it is the very nature of these similarities that determines the type and degree of authority transferred to them. Disastrous consequences are likely to ensue from transferring authority to a substitute that lacks the relevant features of the original.

Where technology is involved, a substitutive technology would also have the authority for judgments and decision-making transferred to it. If the technology is not transferred the authority to make decisions on its own and requires any relevant decision-making to be taken on by a human (where the previous human occupying the role did not require any such interventions by another human), then we would not quite deem the machine to be a substitute for the human in that case. In order for a customer service conversational AI to be a substitute for an Amazon customer service employee, for instance, it must be given the same authority (e.g., to issue refunds, to re-order items, to offer discounts, etc.)

The reason I suggest a large or full degree of authority for decision-making to be transferred where substitutive AI is concerned has to do with the condition of permanence (which I justify shortly). In temporary cases of substitution, such as that of the substitute teacher for instance, it is sufficient for a limited amount of authority to be transferred to the substitute because the original agent occupying the role still has authorship over the role and is expected to

[24] More seriously, we can say that in cases where the entity is not capable of decision-making, its substitute inherits that same lack of decision-making.

eventually return and take care of the more authority-demanding matters.[25] When the substitution is permanent, however, and the original agent is no longer in the picture with regard to the role they previously held, authority is fully transferred to the substitute.

Conditions (c) and (d) are similar in justification to condition (b). In each of the instances above, we can see that the duties and, thereby, the responsibility and accountability also transfer over with substitution, but to different degrees. The substitute teacher has some, but not all, responsibilities of the teacher transferred to them. For instance, a substitute teacher may be responsible for the safety of the children, but not responsible for not properly preparing them for an upcoming quiz. Whereas with the substitute decision-maker and substitute basketball player, much more (if not all of the) responsibility is transferred. A substitute player is responsible for missing a shot or losing the ball or fouling or even losing the game, just as much as the player for whom they are substituting would be. Substitutive technology is no different. When a piece of technology substitutes for a human, the duties, and thereby, the responsibility and accountability is also transferred to the technology and away from the human.[26] Once again, because the substitution is permanent, the duties, responsibility, and accountability are transferred to a full degree. Worth noting is that there is certainly room on this definition to allow for these duties to be adjusted as long as they aim to achieve the same tasks. For example, just as the duty of writing down notes using a pen and paper may be adjusted to become the duty of typing notes on a computer, some duties that are performed by humans to achieve a certain task may be adjusted to fit the AI in achieving that same task, and where a task is no longer deemed necessary with the substitution of technology, it may be removed altogether – for example, if asking patients about their allergies is no longer necessary because the AI has advanced features that allow it to detect allergies, then it may be deemed a substitute without the need to perform that particular (now no longer necessary) duty.

---

[25] Note that I do not claim here that *all* instances of temporary substitution require a limited amount of authority to be transferred (for it is possible, as in the example of the substitute player, that a temporary substitution will allow for a higher degree of authority to be transferred). My claim is rather that all instances of substitution in which a limited amount of authority is transferred to the substitute are instances of temporary substitution.

[26] The question of AI's capacity for responsibility will be discussed in Chapter 5. For the time being, it is worth pointing out that the possibility of technology lacking the capacity for responsibility does not point to a problem with the definition of "substitute" here but rather points to a practical problem with transferring responsibility away from agents capable of it and on to agents incapable of it, thereby creating a responsibility gap.

Condition (e) requires the most justification because it is not applicable to all the senses of substitution discussed above. Note, however, that in the temporary senses of substitution, the purpose of the substitute is to fulfill a task or function only while the original agent is not willing or able to perform them (otherwise there would be no need for the substitute). A substitute teacher is only needed when the regular teacher is unable to attend class. A substitute decision-maker is only needed when a patient lacks consciousness or competence. And a reserve is only needed when a starter is unwilling or unable to play well or needs a break. But technology is a different case entirely. A piece of technology may substitute for a human not only when the human is unable to perform the job, but even when the human is present, available, and perfectly able and willing to perform the job, much like how 2% milk may be available and perfectly good to use, but oat milk may substitute for it nonetheless. This is why substitutive technology is not temporary – there is no one whose return to the role is awaited because there is no one else whose role is necessary in addition to the technology.

As I mentioned at the beginning of this section, this account of substitutive technology is only intended to capture the category of substitutive technology that is ethically interesting. In other words, it is possible for AI to substitute temporarily and/or be granted a very minimal degree of authority, responsibility, and accountability. Imagine, for instance, a professor who, for one day, decides to allow a "teaching AI" to teach her class all on its own, while she continues to occupy the role on a more permanent and routine basis. This AI can be said to have temporarily substituted for the professor. But this example is not usually what is feared or debated where substitutive technology is concerned. This, in large part, has to do with the human retaining authority, responsibility, and accountability. The ethically interesting cases are those in which these have been fully and permanently transferred to the technology and the original agent no longer occupies the role. It is these applications of technology, then, that I refer to as "substitutive."

3.4 Assistance

I would now like to shift to the concept of assistance by considering three examples of "assistants": a teaching assistant, a doctor's assistant, and the Google assistant. A teaching assistant's role is different from that of the instructor they are assisting. They may share some of

the duties of the instructor (like responding to students' emails and holding office hours), but they generally perform the tasks that the instructor does not have the time for (e.g., grading and running labs and tutorials). Moreover, they are not authorized to make decisions at the same level as the instructor, nor are they allowed to override the instructor's final decisions (though the instructor may override theirs), and thus, are not responsible or accountable for the overall course to the same extent as the instructor is. Most importantly, they do not take over the core tasks of the instructor (e.g., creating a syllabus, giving lectures, or preparing tests).

A doctor's assistant's role, similarly, is different from that of the doctor. A doctor's assistant does not take over the core tasks of the doctor, but rather simply performs tasks (e.g., scheduling appointments, preparing the exam room prior to a patient's appointment, updating patient records, and taking care of billing matters) that allow the doctor to better be able to perform their own job. In some instances, we may say that the assistant has the knowledge or ability to perform certain tasks better than the agent they are assisting (for instance, a medical assistant may be proficient with a certain software that the doctor is not). Nonetheless, the assistant is not authorized to make decisions at the same level as the doctor, nor may they override the doctor's orders (though the doctor may override theirs), and thus, are not responsible or accountable to the patients to the same extent as the doctor.

A Google assistant, similarly, does not take over the tasks of the user, but simply performs the tasks that the user does not have the time for, or tasks that would help the user better be able to perform their own tasks (e.g. a Google assistant can add celery to a user's shopping list but cannot purchase it, wash it, store it in the user's fridge, and eat it on behalf of the user). It can remind a user to attend an important appointment but cannot attend the appointment for the user. It cannot override a user's decisions and is not accountable to the same extent as the user is for the final decisions made. Thus, an assistant across all uses of the word can consistently be defined as follows:

A is an assistant for B if (a) A does not need to share the essential characteristics of B; (b) A does not have the same level of authority for decision-making as B; (c) A does not share all of the duties of B; (d) A does not share the same level of responsibility or accountability as B for the final outcome or product; and (e) A's primary function is to accomplish a task or set of tasks

delegated to it by A which, in turn, allows A to accomplish their own, larger, task by saving time, allowing for efficiency, and/or improving the overall quality of the final product or outcome.

Assistive technology, then, no matter how advanced, would have to be no different in its function from the assistant examples above. The criteria for what constitutes assistive technology is, thus, as follows:

Technology is assistive if (a) it does not need to share the essential characteristics of the agent it's assisting; (b) it does not have the same level of authority for decision-making as the agent it's assisting; (c) it does not share all of the duties of the agent it's assisting; (d) it does not share the same level of responsibility or accountability for the final outcome or product as the agent it's assisting; and (e) its primary function is to accomplish a task or set of tasks delegated to it by the agent it's assisting which, in turn, allows the agent to accomplish their own, larger, task by saving time, allowing for efficiency, and/or improving the overall quality of the final product or outcome.

3.5 Hard Cases

Despite having a clear set of criteria to help us distinguish between assistive and substitutive technology, we may still run into hard cases where the line between the two can become somewhat blurred. In this section, I would like to discuss three examples of hard cases, where despite having a clear set of criteria for distinguishing between assistive and substitutive technology, we may nonetheless be conflicted over which box a particular kind of technology fits into.

Before I discuss these hard cases, however, an important question is worth addressing. Namely, must all the criteria of "assistive" or "substitutive" be met in order for the technology to be classified as "assistive" or "substitutive"? In other words, is it possible for some, but not all, of the criteria of "assistive" or "substitutive" to be satisfied in order to classify the technology in that category? Additionally, how might technology that satisfies some criteria from the "assistive" definition and some criteria from the "substitutive" definition be categorized?

The answer to the first question is, in short, yes. In order for technology to be categorized as either assistive or substitutive, it must meet all the criteria in the definition. Of course, it is possible for technology to be neither assistive nor substitutive – and this is precisely the case with technology that meets some criteria from each definition. We can imagine, for instance, a law requiring clinician to do their work exactly as they do now but with the exception of also having to run their decisions past an AI system before proceeding (and only when the AI gives the green light can the clinician proceed with prescribing medication or performing a procedure). In this case, even if the AI may be close to assistive (in the sense of being a type of decision support system), we know that at the very least condition (b) of "assistive" (requiring that it not have the same level of authority for decision-making as the agent it is assisting)[27] is not met. In this case, the technology would be deemed neither "assistive" nor "substitutive." It would be in a different category yet to be identified. Fortunately, these sorts of cases where the AI does not meet the conditions of assistive or substitutive fully are rare, and though they certainly deserve attention, are not the focus of my dissertation.[28]

3.5.1 Assistive Parts Make a Substitutive Whole

The first instance of a hard case is when several separate pieces of technology are assistive on their own but, when combined together, become substitutive. Consider the job of the physician anesthesiologist, for example: a physician anesthesiologist will meet with a patient a day prior to the surgery to discuss the risks and side effects and advise them on which medications to continue or refrain from taking prior to surgery as well as to take note of their individual needs for the operation. They will meet with the patient again just before the surgery to explain how their condition will be monitored during the surgery and what they may need to do afterwards. During the surgery, the physician anesthesiologist's role is to monitor the vital functions of the patient and work to ensure there is proper blood flow, oxygen levels, blood pressure, etc. And after the procedure, the physician anesthesiologist may provide the patient

[27] This criterion is to be read as indicating that the assistant has a *lower* degree of authority for decision-making than the agent they are assisting.

[28] Note that although in Chapter 6 I discuss the possibility of clinicians deferring to decision-support systems in cases of disagreement, these are cases where the AI is still deemed to be assistive because the decision to defer to the AI remains with the clinician who retains authority.

with medication for the side effects (MacGill, 2017). These tasks make up much of the role of the physician anesthesiologist.[29]

Imagine, now, that a conversational AI has been implemented which informs patients of all the risks and side effects of their upcoming surgery and is able to respond to their concerns while recording the patients' responses about medications, medical problems, and prior experience with anesthesia. We may, by the criteria above, call this "assistive", because it does not take over the role of the physician anesthesiologist, and it does not have the authority to make decisions, nor does it have the responsibility or accountability of the physician anesthesiologist transferred to it. It is, by our definition, assistive: it is intended to take some of the load off the physician anesthesiologist's hands, thereby making their job easier (where now the physician anesthesiologist can spend their time doing more important things, like assisting in more surgeries). Most importantly, this conversational AI does not have the essential characteristics of what makes a physician anesthesiologist what they are. Now imagine a different machine that has also been implemented which sorts through the patient's responses about their medical history, medications, and prior experience with anaesthesia and creates a personalized plan for the patient according to that information, and another machine that keeps track of the patient's vital functions during surgery and alerts the medical professionals of what they ought to do to restore them when needed – these machines, too, would, by our definition, be assistive. Imagine yet another machine that takes measures to restore patients' vital functions when needed – e.g. by automatically providing medication when it receives an indication that the patient's blood pressure has dropped. This machine, on its own, is also assistive because it is merely saving the physician anesthesiologist time by administering the medication for them so that they may focus on monitoring the patient. Imagine that there is also a separate machine that, given the information it has been fed by the surgeon, is able to determine which drugs the patient needs and prescribe them post-surgery. This, too, is assistive. The difficulty comes in when we consider that while each of these machines on their own are only assistive, when combined together, they meet the criteria for being substitutive (that is, when combined together, they do share essential characteristics of what makes a physician anesthesiologist, they do have the same

[29] Note that this is not an exhaustive list of the tasks and duties the physician anesthesiologist performs or is trained to perform, nor is it representative of all subspecialties or types of surgeries that a physician anesthesiologist may take part in. It is only used for illustrative purposes.

decision-making authority (e.g. about which drugs to prescribe and which medication to administer during surgery), they take on the responsibility of the physician anesthesiologist (in the sense that if the wrong actions were taken to restore vital functions, it would be on the machine(s)), and because the machines together are able to perform all the functions without needing the physician anesthesiologist to ever return to the role, we can say that this process is not temporary.

It is not difficult to see, then, why this would present an issue for the ethical debate over what kind of technology ought to be implemented. In a debate over the implementation of substitutive technology, it would be unclear what it would mean, for instance, for someone to recommend a ban on "substitutive technology" given that each of these machines on their own are not substitutive, and thus, may not present any serious issues that the opponent of substitutive technology may be concerned with, whereas the combination of these machines is what may lead to the very concerns that they may have with the implementation of substitutive technology.

3.5.2 Sorites Substitution

Another hard case involves the implementation of a single machine that starts off assistive but, through upgrades over time, comes to meet the criteria for substitutive technology. Apple watches, for instance, acquire new functions with each new model, where now Apple Watch 8 has temperature sensors to predict menstrual cycles and can combine the GPS function with the microphone and motion sensors to detect when a user has been in a crash. Similarly to upgrades in Apple Watches, machines in some fields, like healthcare, may also be upgraded over time with new features and functions added. Imagine that a conversational AI is implemented. The next model of this AI is given wheels to move around on its own, the next model has GPS to find patients more easily, the model after that has the ability to accurately distinguish drugs from one another, the model after that has a dispenser for drugs and is able to roll them out onto its built-in tray, and so on and so on, until that machine is basically fulfilling the entire role of a healthcare professional. Again, it would appear that while the initial piece of technology is completely assistive, at some point throughout all the upgrades it becomes substitutive, even though each upgrade on its own does not provide the necessary functions to be substitutive. This is very much a Sorites paradox situation.

This sort of case, too, presents an issue for ethical debates over the implementation of technology in certain fields. This is because, similarly to the previous example, it is unclear what it would mean in this case to recommend a ban on substitutive technology, given that the technology is not created "substitutive", nor does each of the upgrades on its own make the technology "substitutive", yet not taking action would result in allowing for the existence of an end product that is, by our definition, substitutive.

3.5.3 Assistant Substitution

A final (pseudo-)hard case[30] I would like to consider where it might remain unclear whether technology is assistive or substitutive is one in which an assistant in a process is substituted with a machine. Imagine, for instance, that there exists a machine that meets the criteria of substitution for a teaching assistant – that is, it shares relevant characteristics with the teaching assistant, it is granted the authority to make decisions to the extent that a teaching assistant may, is given the duties and responsibility and accountability of leading labs/tutorials, grading essays and tests, etc., and this machine is intended to permanently substitute for a human teaching assistant in PHIL 101. In this case, it is unclear whether the machine itself fits into the category of assistive or substitutive, because while it is substitutive for the teaching assistant, it is assistive to the process (just as a teaching assistant is an assistant to the instructor and the course as a whole).

Though this is perhaps more of a semantic puzzle than a deep issue, it is nonetheless worth addressing so that future debates over implementation of technology can be on the same page. Note that the definitions of assistive and substitutive technology derived thus far do not contain a clear answer to this puzzle, yet it is important to be clear about what is being promoted or opposed in discussions over substitutive AI – is it the substitution for *any* agent within a process (including, for instance, a doctor's assistant or a teaching assistant)? Is it the substitution for the *main* or most significant agent within a process (e.g., course director or doctor)? Or is it

[30] I call this a pseudo-hard case because it is not a real "hard case" but a semantic puzzle. Nonetheless, it is worth presenting here for the purpose of shedding light on some important notes regarding technological assistance and substitution.

the substitution for the agents involved in an entire process (e.g., all those involved in the delivery of a course or in providing healthcare services)?

3.6 Resolving the Hard Cases

Resolving these three hard cases will require two essential aspects to be emphasized about our definition of substitutive technology:

1. Substitution is not an intrinsic property. That is, it is not a property of an agent or object. Rather, substitution is a relational property, relative to the process in question. "Substitute" is not a property of any human being, for instance, but it can be a property of the human in relation to their role of managing a classroom while the regular teacher is ill, or playing on the court while the starting basketball player is taking a break, or making important decisions on behalf of their loved one who is in a coma. The same is true of objects, and for the purpose of our discussion, technologies – they are substitutes only in relation to some task, function, or role they fulfill.
2. The defining characteristics of substitutive technology are that they share essential characteristics with the agents they are substituting and they are granted authority, decision-making power, and responsibility to be utilized on a routine basis. This authority, decision-making power, and responsibility is not acquired by the technology out of nowhere – it is *given* to the machine, *transferred* from the human occupying the role to the machine that will now be taking over permanently. If the agent being substituted for is still around and occupying the role while a machine enters the picture, it is no longer a substitution, but perhaps 'co-working' between human and machine.

Given that substitution is a relational rather than an intrinsic property, we can see that the third hard case can be resolved by acknowledging that there is no single answer to whether the substitute for an assistant is, itself, an assistive or substitutive technology. Instead, we may say that it is substitutive in relation to the role of the assistant, but assistive in relation to the process or the course as a whole. What this means for future debates over the implementation of AI is that specification about what one is promoting or opposing with regard to assistance or

substitution will be necessary. General, vague, or ambiguous statements about substitutive technology in a particular field like education or healthcare, will often be insufficient.

Moreover, given that it is not the existence of the technology itself that makes it assistive or substitutive, to resolve the first two hard cases, we may say that as long as the duties, decision-making power and responsibility are not *completely* transferred to the machine(s), they should not be deemed substitutive. What this entails, then, for the proponent or opponent of substitutive AI in a given field, is not the recommendation to implement or the recommendation to ban the *creation* or presence of certain types of technology due to benefits or risks associated with this technology being or becoming substitutive. Rather, it means advocation for or advocation against *allowing* that technology to substitute (through the complete transfer of duties, authority, and responsibility away from a human occupying the role and on to the machine). Note that the issue with the first hard case concerned an inability to recommend the implementation or the ban of any of the "assistive" technologies on their own given that they have the potential, in aggregate, to become substitutive. However, given that it is not the technology itself that is intrinsically substitutive, it need not be an issue for the ethical debate that several assistive technologies have the potential to form a substitutive technology, because "substitutive" is not determined by the technological specifications of the machine. The same is true of the second hard case: it need not be an issue for the ethical debate whether upgrades will cause a once assistive technology to become substitutive, given that, once again, "substitutive" is not a term that picks out any technical specification in the machine. Substitution, I will sum up by saying, is not a technological matter, but a social and organizational matter.

This is a significant point worth stressing for two reasons: the first is that oftentimes in debates and discussions over the future of technology, it is the technology itself (i.e., its very presence or even potential) that is spoken of as being a danger to entire fields of work, as though the matter of what constitutes substitutive technology is simple and straightforward, with the definition lying in or being determined by the very specs of the system. As I hope to shown here, however, this is far from true – in fact, what constitutes substitutive technology or what distinguishes substitutive from assistive technology is often difficult to make precise and may shift over time, precisely due to its being determined by social and organizational factors. The second reason this conclusion is worth stressing is that it emphasizes the role of the individuals, institutions, and organizations in control of deciding whether, how, and to what degree to transfer

decision-making authority, duties, and responsibility to particular technologies. Simply put, we need not be worried about technological advancement as much as we need to be worried about the ethically-negligent policies around them.

3.7 Conclusion

The two chapters to follow raise some concerns about substitutive AI in healthcare and the substitutive approach as a whole (the view that our goal ought to be to ultimately replace human healthcare providers with AI). If my concerns prove to be warranted, the response then ought to be not necessarily to ban the creation of any such technology that has the potential to be substitutive or to cause a slippery slope into the substitutive approach. Rather, the response ought to be to avoid quickly and blindly jumping at the opportunity to transfer authority, duties, and responsibilities from clinicians to machines in healthcare before pausing to determine their impact on the patient-clinician relationship as well as their impact on the healthcare system and society as a whole. There may, of course, be instances where giving this power to certain machines proves to be harmless, but we must also not lose sight of the worry of granting too many "harmless substitutes" these powers without evaluating the impact of the aggregate (as seen in the hard case discussed in section 3.5.1).

Chapter 4: The Substitutive Approach Pt. 1 - The Patient-Clinician Relationship

4.1 Introduction

The goal of this chapter is to discuss the ethics of substitutive AI with regard to the patient-clinician relationship. Before delving into this, however, a brief recap of the previous chapters will be helpful:

Recall that in Chapter 1, I discussed some duties of clinicians (compassion, communication, recognition, creating trust, and respecting patient autonomy) that I deemed to be essential in promoting the values and the overall wellbeing of patients, and thereby achieving excellence in the practice of medicine. In Chapter 2, I presented, then proceeded to reject, the neo-luddite approach to AI in healthcare. This allowed me to direct the remainder of the project away from the extreme position of advocating for the ban of AI in healthcare altogether and towards focusing, instead, on the ethics of assistive and substitutive AI. In Chapter 3, I proceeded to discuss and clarify the concepts of "assistive" and "substitutive" technology, which allowed me to derive the following definitions:

Assistive Technology*: Technology is assistive if (a) it does not need to share the essential characteristics of the agent it's assisting; (b) it does not have the same level of authority for decision-making as the agent it's assisting; (c) it does not share all of the duties of the agent it's assisting; (d) it does not share the same level of responsibility or accountability for the final outcome or product as the agent it's assisting; and (e) its primary function is to accomplish a task or set of tasks delegated to it by the agent it's assisting which, in turn, allows the agent to accomplish their own, larger, task by saving time, allowing for efficiency, and/or improving the overall quality of the final product or outcome.*

Substitutive Technology: *Technology is a substitute for an agent if (a) it is perceived to share some relevant characteristics with the agent it's substituting for, which allow it to perform the*

tasks and functions previously performed by the agent; (b) it has the authority to make judgments and decisions transferred to it; (c) it has the duties and, thereby (d), the responsibility and accountability associated with the agent's role transferred to it; and (e) this transfer of authority, duties, and responsibility is not temporary.

With this clarified definition of what "substitutive technology" refers to, and with a set of duties of clinicians at hand (without which, excellence in the practice of medicine would not be achieved), we may now turn to assessing whether substitutive AI in healthcare would be able to achieve excellence in medicine to the same degree as human clinicians. Or, in other words, whether allowing for the permanent transfer of authority, duties, responsibility, and accountability to an AI that is perceived to share relevant characteristics with a clinician will allow for the fulfillment of the duties that allow for excellence in medicine. If substitutive AI falls short in fulfilling these duties, then we can conclude that substitutive AI fails to live up to the ethical standards in the practice of medicine.

4.2 Compassion[31]

As I mentioned in Chapter 1, when I speak of the duty of compassion, I am speaking of a combination of two things: 1. empathy towards the patient, and 2. the use of this empathy to promote the wellbeing of the patient. I now turn to addressing the question of whether AI on its own is capable of providing compassion in this sense. There are very good reasons to think that it cannot, and those reasons are often tied to empathy. Coeckelbergh (2014), for instance, suggests that the problem of AI's inability to provide good care has to do with its lack of understanding about the human experience, about our shared mortality and suffering, which is essential for genuine care. Similarly, Sparrow and Sparrow (2006) explain that caring is only possible given a mutual understanding of what it is to be human, with human limitations and frailties. This mutual understanding is then revealed in the actions and responses of the carer (such as reaching for someone's hand or wiping away their tears). So for machines to be able to respond in the way that we deem to be indicative of a caring attitude, they would need to be able to exhibit empathy

[31] Much of this section was first published in *AI and Ethics* (2023) and has been reproduced with permission from Springer Nature.

– they would need to be able to recognize and understand human emotions and respond to them in a way that demonstrates (or perhaps simulates) acknowledgement and understanding of these emotions.

It is a matter of active debate whether it is even possible for AI to ever experience human emotions (Chursinova and Stebelska, 2021), and even if it is possible, there may be strong ethical reasons to avoid designing AI with the capacity to experience emotions.[32] But setting those issues aside, we may approach the problem more directly by asking why and whether empathy is even necessary for machines' ability to provide good care. One may argue that AI's empathy is not, in fact, necessary for their ability to provide good care. Sally Dalton-Brown, for instance, posits that "not all physicians are highly empathic, and empathy can be a detraction from care, given its focus on the agent rather than on the altruistic act" (Dalton-Brown, 2020, p. 118). She also suggests that "one does not need to experience a concept in order to understand it or to demonstrate empathy. Palliative care workers for example are not resurrects. Doctors with, for example, severe autism may struggle with empathy but not with care" (ibid.).

So Dalton-Brown is essentially putting forth two reasons for why AI need not be capable of empathy in order to provide good care: 1) empathy may actually detract from care and 2) it is possible to provide good care for patients without *experiencing* empathy towards them. At this point I would like to unpack each of these and see if Dalton-Brown is warranted in making these claims.

4.2.1 Objection 1: Empathy Detracts from Care

Beginning with the first argument: Could empathy detract from care? The empirical evidence suggests the opposite. In contrast to Dalton-Brown's claim, Jodi Halpern (2003) explains that there is plenty of evidence to suggest that empathy is, in fact, very beneficial with its link to decreased patient anxiety, which is linked to physiological improvements for patients with a variety of illnesses. In addition to this, Dr. Helen Riess (2010) also explains that empathic

[32] Chomanski (2019), for instance, argues that designing AI with human-level intelligence to act as servants will lead to exhibiting the vice of manipulativeness, while Beckers (2017) discusses the possibility of AI capable of superintelligence to also be capable of "supersuffering" (that is, suffering to a degree exceeding that of humans).

physicians are better able to obtain critical information from patients – this, in turn, affects the quality of care that is provided to them as well as the medical outcomes altogether.

Though perhaps there is an alternative way that we may read Dalton-Brown's claim – that is as suggesting that empathy could potentially lead to misguided decisions, such as a doctor refusing to perform a risky but vital operation on a fearful patient. This type of argument has also been presented by Paul Bloom (2016), who suggests that high empathy individuals tend to be less able to help others in the long run because they are too concerned about inflicting short-term pain. To illustrate, Bloom uses the example of a parent having to force their child to do homework, eat their veggies, and go to bed at a reasonable hour to demonstrate that inflicting short term pain in this way is necessary for long-term benefit of the child, but this would be impeded by empathy (ibid., p. 35).

While Bloom is correct to emphasize the danger of decision-making purely on this basis, this is not a good reason to suggest that empathy *itself* detracts (or can detract) from care or to proceed to use that as a reason to suggest that AI lacking empathy may not be such a bad thing. The only claim that Dalton-Brown or Bloom may make here is that perhaps *too much* empathy or the *improper application of empathy to decision-making* (i.e., as the *only* basis for decision-making) have the potential to detract from care. Note that there being an issue with too much empathy does not translate to empathy, itself, being an issue, similarly to how the dangers of water intoxication due to overhydration do not translate to hydration, itself, being dangerous. The best way to be an empathetic physician, according to Riess, is to see the world from the patient's view but then step out of it in order to be objective and make rational decisions. In an interview conducted by Kasley Killam (2014), Riess uses the following example to illustrate this point:

> You could have a patient who is really afraid of needles and doesn't want to get a tetanus shot. If you empathize too much with that emotional fear, you might decide, "don't have the shot, because I can see how distressed you are." But once you get back out into your physician role, you realize instead, "I need to help you overcome the fear, because it'd be much worse for you to get tetanus." (Killam, 2014).

We may say something similar of Bloom's example regarding one's own children: surely parents do empathize with their children, but a good parent will also be able to recognize that their child will suffer more in the future if they are not pushed to do things that they do not currently want to do. Thus, again, while too much empathy, or reliance on empathy alone for decision-making,

may be problematic, empathy in the appropriate amount and in the appropriate way, is extremely beneficial in healthcare (for the emotional, mental, as well as the physical wellbeing of patients).

Another sense in which empathy detracts from care, it has been argued, is as follows: empathizing with one particular individual makes it tempting to give them preferential treatment over others. In healthcare, this would surely be unethical. Bloom (2016) explains the worry of empathy clashing with other moral considerations by referencing an experiment conducted by Batson et al. (1995), in which subjects were told about a young girl with a fatal disease who was placed on a waitlist for a treatment that would serve to ease her pain. In this experiment, Batson et. al. found that when the subjects were just asked what to do without being asked to imagine how the young girl felt, they tended to respond by saying that she should continue waiting. But when asked to first imagine how the young girl felt, subjects tended to want to move her up on the waitlist, putting her ahead of other children who needed the treatment even more. According to Bloom, "here empathy was more powerful than fairness, leading to a decision that most of us would see as immoral" (Bloom, 2016, p. 25).

However, there are two ways to respond to this worry. The first is one I have mentioned already: empathizing with someone need not mean that we must base our decisions on our empathy for that person alone. It is possible (and preferable) that a clinician empathize with a patient but nonetheless consider competing factors when making a decision. The second way to respond to this worry is by pointing out that the reason we see such a decision as being "immoral" is precisely because we are able to also empathize with the other "needy children" (as Bloom refers to them) on the waitlist – we can almost picture their sad faces when told that they will have to wait longer because someone less in need of the treatment has been given preference over them.[33] If we did not imagine or feel what it would feel like to be one of those other children who were pushed back (or the children's parents), it would be difficult to grasp the unfairness or immorality of such a decision. We may be taught, as a general rule, that we need to be impartial, and we may repeat that rule to ourselves and even abide by it, but we would not truly be able to grasp it or the reason behind it unless we are able to empathize with those who suffer the unfairness of our not abiding by it. So perhaps it is not empathy itself that detracts

[33] Bloom would disagree with me here because 1) he does not believe that we can empathize with more than a couple of people at a time, and 2) he does not believe that empathy is the reason people act morally. I disagree with both claims (responding to them here, however, would fall outside of the scope of this chapter).

from our ability to make moral and fair decisions, but rather empathy for only *some* but not all of those affected by our decisions. Once again, the fact that the wrong kind of empathy, the wrong amount of empathy, or the improper application of empathy to decision-making are an issue does not translate to empathy, itself, being an issue or detracting from care more generally.

It would be a mistake on my part at this point to fail to acknowledge the fact that it is often easier to empathize with those who are closer in terms of social proximity, and that the real issue is that this fact is precisely what leads to partiality. A way to handle this, however, would be to urge doctors, if they are unable to practice empathizing with those not within social proximity, to be cognizant of this limitation when making decisions that affect those who are not within social proximity (and thus, acknowledge that basing decision-making on empathy in such cases would be an instance of the improper application of it to decision-making).

Finally, there may be one other sense in which empathy may be thought to detract from care: by causing clinicians to feel overwhelmed, which leads to a decreased ability to help their patients. In discussing the case of his uncle, who was undergoing cancer treatment, Bloom says "he seemed to get the most from doctors who didn't feel as he did, who were calm when he was anxious, confident when he was uncertain" (Bloom, 2016, p. 146). Thus, empathy in healthcare, it may be argued, is far from beneficial (for both the clinician as well as the patient).

But, first, it is important to distinguish between different levels of empathy. When Bloom critiques "empathy", he has in mind empathy in the sense of experiencing almost exactly what someone else is feeling.[34] But empathy need not be either this or nothing.[35] It is likely that clinicians who feel overwhelmed are *over*-empathizing with their patients. Second, regarding Bloom's uncle preferring doctors who "didn't feel as he did", I think it is important to acknowledge that empathizing with someone does not mean that one must respond to them in the way that matches their emotions, but rather quite the opposite: it often means responding to them in the way that most benefits them, given the empathizer's understanding of how they feel and

[34] He writes: "There is a world of difference…between understanding the misery of the person who is talking to you because you have felt misery in the past, even though now you are calm, and understanding the misery of the person who is talking to you because you are mirroring them and feeling their misery right now" (Bloom, 2016, p. 148) and it is only the latter that he sees as problematic.

[35] I purposely use "empathy" here in a vague way to account for different types and degrees, ranging from simply understanding how one must be feeling given one's own similar experiences to, as Bloom defines it, feeling almost *exactly* what someone else is feeling. A more detailed conception of what empathy encompasses is presented in the following section.

what they need. By empathizing with their patients, the clinician will be able to recognize what it is that would help them in that situation, what would alleviate their pain or what would calm them down, and respond accordingly. So to respond to Bloom: I think it is a mistake to think that his uncle's doctor failed to empathize with him simply because he was calm and confident when the patient was not. In fact, the doctor was likely empathizing in the most appropriate way. Once again, I hope to have shown that it is not empathy, itself, but rather an inappropriate degree of empathy or reliance on empathy alone in decision-making, that detracts from care.

At this point, one may raise a possible concern with the responses I have provided: *over-reliance* on empathy or *too much* empathy may not translate to empathy, itself, being an issue, but they are nonetheless issues that humans are susceptible to due to their capacity for empathy. AI, by lacking the capacity for empathy, would also refrain from making the sorts of errors in decision-making that only exist as a by-product of empathy.

It is important to note, however, that while there are costs associated both with AI's lack of empathy and with potentially misguided decisions by humans who misapply empathy to decision-making, the issue of misapplication of empathy to decision-making is fixable (e.g., through guidelines and training) while the issue of AI's lack of empathy is not.[36] Humans may be able to overcome the issues that come with overreliance on empathy or too much empathy, but AI will never be able to overcome the issues that come with a lack of empathy altogether.

4.2.2 Objection 2: Empathy is Not Necessary for Care

Dalton-Brown's (2020) second claim is that it is possible to understand the patient's situation and care for them without necessarily *experiencing* empathy. Here, I take her to be suggesting that AI need not actually experience empathy, but simply perform the actions needed to provide patients with the care they need.

To evaluate this claim, I would like to first begin by taking a look at the two examples Dalton-Brown uses to back it up: that "palliative care workers…are not resurrects" and that "doctors with…severe autism may struggle with empathy but not with care" (Dalton-Brown,

[36] This is, of course, assuming that AI is never capable of experiencing empathy due to the practical and/or ethical issues this possibility raises. If AI does come to be capable of experiencing empathy, this entire discussion will need to be revisited.

2020, p. 118). Let us begin by examining the first example – the case of the palliative care worker, which I take to be referring to someone who is able to care for the dying despite never having had any experience with death or dying themselves. This example is perhaps not the strongest example to rely on for making the sort of point Dalton-Brown hopes to make here, however. This is because although the palliative care worker is not a resurrect (that is, has no knowledge of what it feels like to be dying), it is not clear why this is necessary in order for her to empathize with her patients. It seems that the only kind of experience the palliative care worker would need is experience of being human – that is, experience of mortality, fear, pain, sadness, loneliness, uncertainty, and other similar human emotions that we can imagine a dying patient may be experiencing. Of course, Dalton-Brown could argue here that the generic sense of mortality is qualitatively different from the certain knowledge that death is imminent. But even if this is true, empathy does not require having had the exact same experience as someone else. It can simply be imagining as much as possible what they must be going through and using one's own experiences to attempt, as much as possible, to understand how they are feeling. Additionally, as Fernandez and Zahavi (2020) point out, it is not obvious that having had the same experiences as someone allows for a better understanding of them or a greater ability to empathize with them. For instance, a woman who has had an easy birth is not always better able to understand a woman who is in a lot of pain due to a difficult birth than someone who has never given birth at all. In fact, the woman who has had an easy birth may generalize from her own case, leading her potentially to have greater difficulty grasping the other woman's pain.

Dalton-Brown's example of the doctor with severe autism is also not entirely persuasive in establishing her point, particularly because her claim relies on too simplistic an understanding of "empathy." Not only is there a lot of disagreement on the issue of empathy among people with autism (Song et al., 2019), but while doctors with autism may struggle with *some* aspects of empathy, they certainly do not *lack* empathy in the same way that machines do. Song et al. (2019) make it a point to distinguish between three types of empathy: cognitive empathy, empathic concern, and empathic accuracy. Cognitive empathy refers to the ability to understand and infer others' beliefs, feelings, and intentions; empathic concern refers to the ability to be moved by others' emotions; and empathic accuracy refers to the ability to share others' emotional states. After accounting for different types of empathy in this way, it was found that empathic accuracy (the sharing and resonance with other people's feelings) in individuals with autism

spectrum condition is actually often intact or better than neurotypical individuals (ibid., p. 22). And for the purposes of our discussion, this finding is perhaps enough to point to the disanalogy between doctors with autism and AI, given AI's complete inability to share in others' feelings. This is significant in the context of care because the ability to share or resonate with patients' feelings can still create a sense of connection, and this awareness that their doctor genuinely shares their emotional state can allow patients to feel more comfortable and trusting of them. What's more, this ability for empathy is something many patients will be aware of when cared for by a doctor with autism, despite the doctor's possible inability to respond to them in the way they would prefer to be responded to (given their impairment in cognitive empathy and empathic concern). The AI, however, will lack empathic accuracy completely just by the fact of not being an entity capable of feeling.

But perhaps we can improve Dalton-Brown's argument with a new example – namely, that of caring for an evil patient. In an episode of Grey's Anatomy (Rhimes, 2007), we see a patient with a large swastika tattooed on his abdomen brought into the hospital and in need of an emergency surgery that must be performed by Bailey, a Black surgeon. Although struggling to get through it and struggling to empathize with the patient, Bailey does perform the surgery and ultimately saves the patient's life. So, in thinking of an example like this, one may reassert Dalton-Brown claim: experiencing empathy is not necessary for providing care. If it was, then surely Bailey's inability to empathize with the racist patient would have meant that he would not receive care (at least not by her).

I think an example like that of the Grey's Anatomy racist does a great job of demonstrating that empathy is not necessary for performing medical procedures, ensuring the patient's physical pain is relieved, etc. This fact, however, is still insufficient in its attempt to underplay the importance of empathy in healthcare. Telling your child that you are proud of them or that you love them may not be necessary for keeping them alive in the same way that feeding them is, but this does not mean that it is not very much important for their wellbeing nonetheless (and some of the benefits of empathy towards patients have already been discussed above, so I will refrain from repeating them here). What's more, while it may very well be the case that the Grey's Anatomy racist or Ted Bundy or Jeffrey Dahmer have forfeited this privilege, and thus, may have to deal with the disadvantages of this lack of deep care given that the medical staff

working with them may find them difficult to empathize with[37], it would be a mistake to conclude from this that empathy is not still of great importance to those who do receive it or that we would not be losing out on a lot if it ceased to exist completely in the realm of healthcare.

4.2.3 Simulated Empathy

At this point, it is important to acknowledge the possibility that AI can come to be programmed with certain responses that emulate those of empathic humans. This AI would, for instance, learn to gauge individual patients' reactions to certain things and gradually get better at determining what course of action to take when that patient displays fear, anger, sadness, anxiety, etc. Given that AI is already capable of recognizing facial expressions (del Castillo Torres et al., 2023), this is certainly a possible future scenario. This capacity to simulate empathic human responses may then come to mean taking Laura's hand when she feels scared, or speaking in a lower and calmer voice when Andrea gets upset, or allowing Jane (who does not get too many visitors and is often lonely) to share her stories, and making comments every once in a while to simulate active listening. So, these machines would be virtually indistinguishable from a human physician. The two questions we must now address are, first, whether something crucial would still be missing in terms of compassion if machines that could simulate empathy were to replace human health care providers and, second, whether there are other reasons, besides the inability to fulfill the duty of compassion, that would make the idea of machines simulating empathy unacceptable.

Again, let us imagine that these advanced empathy-simulating AI are so well designed that they are indistinguishable from the real thing. What does this mean for our discussion? To start, this means that, given the AI's indistinguishability from humans, the developers will have a choice to either (a) make it very explicit to the patients that the AI is not human and is incapable of experiencing emotions and empathy, and thus, incapable of caring about them or (b) allow patients to believe that the AI cares about them. Both options, however, have their own set of problems. If we choose the first option and ensure that patients are constantly reminded that the

[37] Though we may also think that some clinicians will nonetheless be able to empathize with them despite their evil, given that they are still a fellow human being that the clinician is witnessing at their most vulnerable state (e.g. lying on the operation table, naked, at the mercy of others, in pain, etc.).

AI is an object, much like the bed they are sitting on or the spoon they are eating with, incapable of having any feelings or being invested in their wellbeing (and the patient fully understood and accepted this fact), then this would risk losing the beneficial effects of the simulated compassion. This is not to say that such robots would never be viewed as more than mere objects. Darling (2017), for instance, demonstrates that something as simple as framing robots in an anthropomorphic way impacts how they are viewed and treated (i.e. with participants exhibiting empathy towards them and hesitating to strike them). Thus, I am not denying the possibility of an empathy-simulating AI being viewed as a social actor or living thing, despite constant reassurance about its lack of an inner life or capacity for empathy. However, viewing something as a social actor or a living thing does not translate to being capable of feeling "cared for" by it. It would be highly unlikely for a patient who genuinely grasps that the empathy in its care provider is entirely simulated, to feel "cared for" by the care provider (and anything felt resembling "care" would certainly not be in the same sense, or nearly to the extent, of that experienced by being cared for by an entity with a known capacity for empathy).

So it seems that if we want the positive effects of the empathy-simulation, then we would have to allow the patients to let themselves believe that the AI cares about and can empathize with them. This, however, is also problematic. Sparrow and Sparrow (2006), in their discussion of the introduction of such robots in aged-care facilities, explain that while it is likely that this kind of AI will make residents feel loved and cared for and may give them some entity to form emotional attachments to, thus making them happier, the problem stems from this happiness being purely the result of a delusion. This delusion, they elaborate, is problematic for two reasons: The first is that in order to avoid a moral failure, we have a duty to not let ourselves be deceived, but to, instead, see the world as it *is*. The second reason it is problematic is the same reason most of us would not want the life of Shelly Kagan's (1997) deceived businessman.[38] That is, what we want is for things to *be* a certain way, we do not just want the experience or the belief that they are a certain way (Sparrow and Sparrow, 2006, p. 155). In the case at hand, what we want is to actually *be* cared about, not just to *believe* or *feel* like we are cared about. Sparrow

[38] In Kagan's thought experiment (intended to show that mental states are not all that matter), the "deceived businessman" dies contented, falsely believing that he is a loved and respected successful businessman, when, in reality, his wife had been cheating on him, his children were only after his possessions, members of his community were only after his charitable contributions, and his business partner had been embezzling funds from his company.

and Sparrow refer to Nozick's experience machine as an illustration of why having pleasant but delusory experiences is not satisfactory for, and would not be chosen by, most people – a claim supported by empirical literature. Hindriks and Douven's (2017) study on the experience machine, for instance, has found not only that the majority of participants who were given the option did not want to be hooked up to an experience machine, but also that alternative scenarios (variations of the experience machine thought experiment) which severed contact with reality less than being hooked up to an experience machine, were more likely to be accepted by participants.

This leads us to conclude that designing AI with the ability to simulate empathy would either be deceptive, and thus morally unacceptable, or fail to be sufficient in providing the same level of compassion as human health care providers.

4.3 Communication

Let us now turn to evaluating AI's capacity for communication. According to a study by Jagosh et al. (2011), proper communication between patients and clinicians helps to promote patients' wellbeing in several ways: 1) It allows clinicians to gather clinical information about patients (i.e., their knowledge of their own bodies and their own state of health) that aid in better being able to diagnose and treat them; 2) It acts as a therapeutic agent. In the study, patients highlighted how attentive listening by a doctor "offered relief from the stress and anxiety that can be induced and exacerbated by illness" (ibid., p. 372); and 3) It strengthens the doctor-patient relationship by making patients feel that they are in a partnership with their doctor and that they matter to them.

What good communication entails, however, is not one single thing. A study by Kee et al. (2018) which looks into the errors in communication that negatively impact the patient-clinician relationship, suggests that there are four categories of communication: 1) non-verbal (which includes eye contact, facial expression, and paralanguage); 2) verbal (which includes active listening and appropriate choice of words); 3) content (which includes the quantity and quality of information communicated); and 4) attitude (which includes empathy and respect) (ibid., p. 99).

Though it is clear from Kee et al.'s study that human clinicians are not always great communicators, with proper training and practice, communication skills in humans can often be

improved. It is not clear whether the same can be said of AI, however. In what follows, I would like to delve into what each component of communication generally requires and look into whether AI is able to meet the standard for what constitutes good communication (or at least be capable of refraining from making obvious communication errors on a constant basis).

4.3.1 Non-verbal communication

With regard to non-verbal communication, some patients interviewed in Kee et al.'s study mentioned a lack of eye contact (especially a lack of eye contact due to the clinician's preoccupation with technology) as a communication error. Other patients interpreted clinicians' facial expressions – both, overt negative facial expressions as well as subtle facial expressions that suggested disinterest – as representative of their uncaring attitudes. The paralanguage component of communication (i.e., volume, tone, speed) that a clinician displayed was also noted as playing a role in whether or not patients felt cared for by that clinician – for instance, an expressionless tone indicated to a patient that "personalized care was not being delivered" (Kee et al., 2018, p. 100).

AI appears at first glance to be able to avoid non-verbal communication errors. Eye contact, if considered in purely mechanical terms, is of course possible for AI. A robot designed with what appear to be eyes can certainly be programmed to detect and follow the eyes of the patient. Robots can also be programmed to mimic facial expressions that convey a caring or a comforting attitude. The paralanguage aspects of a robot's communication (i.e., the speed with which they utter sentences, the volume at which they are uttered, and the tone in which they are uttered) can also be made to fit those displayed by "good communicators." Thus, it appears that AI would make for a great communicator where the non-verbal aspects of communication are concerned.

The issue, however, is that these components of non-verbal communication only matter insofar as they allow patients to feel acknowledged and cared for by their clinician. In other words, it is not the eye contact itself that matters but what the eye contact conveys (acknowledgement of, focus on, and prioritization of the patient in that moment).[39] Similarly, an

[39] There may, of course, be cases where doctors are able to fake these – this, however, is not an issue of communication as much as it is an issue of the simulation of compassion. Whether or not doctors with, e.g., psychopathy or sociopathy ought to be excluded from at least some domains in medicine is an issue for a different project (as is the comparison of such doctors with AI).

expression that conveys disinterest or a negative facial expression that conveys frustration is not problematic because of the facial expression itself. It is not that the facial expression is not nice for the patients to look at, but rather, it is not nice for the patients to look at because of what it represents (namely, frustration and disinterest). An expressionless tone, too, only makes patients feel that they are not receiving personalized care because of what an expressionless tone is generally associated with (boredom, repetition, mechanical tasks, etc.) Thus, even if AI is able to mimic the non-verbal aspects of a good human communicator, it would be unlikely to produce the same results in the patients as they would if displayed by a human communicator, given patients' understanding of the lack of inner-life of a machine (the lack of an ability to care, to prioritize, to be 'interested' in the patient, etc.).

4.3.2 Verbal communication

A large portion of verbal communication errors on the part of clinicians stem from a lack of active listening. Active listening involves allowing patients the opportunity to ask questions and not interrupting them when they are voicing their concerns. The other aspect of verbal communication has to do with word choice (phrasing things in a sensitive way that acknowledges the feelings of the patient). Proper word choice can certainly be programmed into AI, as can pauses that simulate active listening. The AI may even be designed with the feature to record and analyse the words of the patient so that it may provide responses to their concerns in real time. Smart voice assistants like Alexa and Google Assistant already have this feature – that of listening to users' questions, providing a response, and "listening" for a follow-up. It is safe to assume that these features will only be improved over time, where the listening feature is triggered more accurately and the library of responses is vast enough to account for just about every possible patient question or concern. Still, a concern remains regarding AI's ability to "listen" in the same way that a human can, where rather than collecting information in order to provide a response, it truly absorbs the patient's concerns and questions before responding – that is, it takes patients' words not merely as input data, but as containing within them the feelings of real humans like oneself, and providing responses that are demonstrative of this connection and understanding between two similar beings.

4.3.3 Content

The content or the information shared with a patient can, itself, also be a source of communication error if done improperly. The two ways in which the content of the communication can be a source of communication error, according to Kee et al., is by clinicians either not sharing adequate information with patients or sharing information of poor quality and without proper detail or explanation for the rationale behind their decision (Kee et al., 2018, p. 101).

While AI can be programmed to provide the appropriate amount of information that patients may be seeking and present it in a way that is accessible to them, an issue arises when the patient is dealing with black-box AI with millions of variables, and where the contribution of each one to the final output is extremely difficult to access and understand. In this case, the information that the AI can share with the patient will be limited.

It is worth noting that in regular medicine, clinicians may not always have the answers to all of patients' questions, either (e.g., why a certain course of treatment is expected to work). However, there will generally be some information that the clinician can provide (Bjerring and Busch, 2021, p. 364-5) (e.g., that while it is unknown precisely why prescription X helps with condition Y, trials were done on populations a, b, and c, and the results favoured the prescription of X for Y). Communication with black-box AI, or through a robot acquiring information through black-box AI, however, will be very limited in terms of content. In such instances where very limited or no information is provided, trust of the clinician will have to do a lot of heavy lifting in the patient-clinician relationship, and as I go on to demonstrate in Section 5, trustworthiness is not a capacity AI possesses.

A proposed solution may be to employ post-hoc XAI methods in the form of separate models that attempt to provide some transparency and insight into the outputs of the system. As I noted in Chapter 2, however, while these methods may be able to approximate the most relevant features of a system's outputs, they are limited in that not only can their explanations be wrong and unstable (Vale et al., 2022, p. 821), but in instances where their explanations are correct, these explanations could leave out too much significant information (Rudin, 2019).

A point worth noting is that while models like Chat-GPT may appear to be capable of providing sufficient self-explanations, these explanations do not actually reflect the prediction

process itself, and as Huang et al. (2023) have found, when models like ChatGPT offer explanations, this often comes at the cost of model accuracy.

4.3.4 Attitude

Attitude can negatively affect communication in the patient-clinician relationship when the patient deems the clinician as having failed to demonstrate empathy and show respect towards them. Empathy, as was demonstrated in Section 4.2, is doubtful for AI, and it appears that a consequence of this lack of capacity for empathy is also a hinderance to AI's capacity for communication.

What "respect" entails, on the other hand, is often connected to other aspects of communication (Kee et al., 2018, pp. 101-102) (e.g., providing patients with sufficient information, actively listening to them, displaying positive facial expressions and using an appropriate tone when speaking). Thus, if these other aspects of good communication cannot be met (or if they are met but do not hold the same value when met by AI), then the respect aspect of communication will not be fulfilled, either.

Taking all aspects of communication (non-verbal, verbal, content, and attitude) into account, we see that there remains quite a bit of doubt regarding AI's capacity to have strong communication with patients. With communication hindered, AI as a substitute for a clinician will also result in the possibility of less data gathered about the patient's state of wellbeing, needs, and values, a decreased ability to calm patients and provide the therapeutic effect many of them may need, and an overall weaker patient-clinician relationship.

4.4 Recognition

Next, I would like to turn to evaluating AI's capacity for recognition. While the concept of recognition is often used in political philosophy to discuss the connection between misrecognition and oppression or injustice, for the sake of this discussion I will be appealing to the concept of recognition purely as it relates to identity formation and identity maintenance.

Often traced back to Hegel, the idea of recognition refers to the way an individual is identified and responded to by others. Heikki Ikäheimo defines recognition as "a case of A taking B as C in the dimension of D, and B taking A as a relevant judge", where A is the recognizer; B is the recognizee; C is the attribute (e.g. being someone whose well-being is important, having a right to something, or being valuable/worthy of esteem); and D is the dimension of B's personhood at issue. This dimension of personhood, according to Ikäheimo, can be singular ('simply as such'), it can be based on autonomy (i.e., insofar as one is autonomous), or it can be particular (i.e., only insofar as one has some particular feature) (Ikäheimo, 2002, pp. 450-451). Most importantly for our discussion, this recognition (or misrecognition or lack of recognition) by others influences an individual's own view of themselves and their own identity. As Charles Taylor puts it, "my discovering my own identity doesn't mean that I work it out in isolation, but that I negotiate it through dialogue, partly overt, partly internal, with others" (Taylor, 1992, p. 34).

Simply put, it is in being around and communicating with and being recognized by other people that we form our identity and compose our self-image. Being recognized by others is, according to Taylor, "a vital human need" (ibid., p. 26). While many patients may have human contact outside of healthcare facilities, there are also patients whose primary (or only) source of human contact comes from health care providers[40], and for such individuals, recognition by their health care providers is essential – and it is particularly important in healthcare given that chronic illnesses or new disabilities, for instance, can cause patients to question, or become uncertain of, their own identity, which recognition by others can help reaffirm.

The answer to whether or not AI is capable of recognition lies in a particular requirement for recognition: that the recognizer, to be capable of recognizing someone, must be recognized by that person as a credible judge. As noted by Ikäheimo, "if A esteems B highly, and B understands this but does not accept A's judgment on the matter with any credibility…no recognition seems to take place. Hence, B has to attribute some credibility to A's judgment, or take A as a relevant judge on the matter in question" (Ikäheimo, 2002, pp. 451-452). AI,

[40] I have in mind, for instance, patients in Alternate Level of Care (ALC) units awaiting long-term care placements who receive few or no visitors during their time on the unit. According to the Canadian Institute of Health Information, 20% of patients in ALC were there for over a month, and 4% were there for more than 100 days (Canadian Institute of Health Information, 2009, p. 5). For a discussion of long-term care placement wait times, see Laporte et al. (2017) and for a discussion of the wait times for mental health unit patients awaiting placements in long-term care, see Lane et. al. (2010).

however, is not capable of recognition precisely because it is not typically recognized as an entity capable of recognition. This primarily has to do with our perception of AI as tools rather than fellow collaborators. I do not worry, for instance, that my phone will think that I am lazy for not charging it often enough in the same way that I may worry that my colleagues may view me as lazy for not completing my portion of a project in time. And this inability to participate in recognition is not something that only applies to the kinds of machines that we currently have – it is something that will very likely be true of more advanced future machines, as well. This is because, aside from being viewed as tools, machines are also not the sorts of things whose "inner-life" we can ever be sure of – at least not to the extent that we can be of that of other humans. According to Robert Sparrow, no matter how sophisticated a machine may be or how much it may resemble a human being, there will always remain an "unbridgeable gap…between reality and appearance in relation to the thoughts and feelings of machines…[which] explains why we find it impossible to take seriously the thought that machines could have an inner life" (Sparrow, 2004, p. 210). Simply put, there will always remain some doubt over whether a machine feels what it shows us or tells us it is supposedly feeling. To illustrate, Sparrow asks us to imagine that there exists a complex machine with many diodes that flash in many different patterns. One pattern should be taken to mean that the machine is suffering a minor pain, another means that it is suffering great pain, another means that the machine is happy, another that it is unhappy, etc. And so with this information in mind (provided to us by the engineer behind it), we behave towards the machine in a way that reduces the 'suffering' flash patterns. However, Sparrow then asks us to imagine that the engineer returns and admits that they have been mistaken about which flashing patterns indicate what pain responses or cognitive states, and what, for instance, was previously thought to mean that the machine was happy actually means that the machine is in pain and that, consequently, our behaviour towards the machine needs to be very much altered. As sparrow puts it, "at this point the possibility of radical doubt emerges. How do we know that the engineer has got it right this time?" (ibid., p. 210). And if doubt always remains about the inner lives of machines, this makes it difficult to take them to be credible judges, and consequently, makes it extremely unlikely to be recognized by them precisely because we cannot recognize them as the kind of entities that are capable of recognition.

A point worth noting is that if AI advances to the point where its sentience and inner life are widely accepted, to the extent that we accept that of other humans, then perhaps such AI will

be capable of recognition. Still, there is good reason to doubt that this will be the case given that evolutionary biology will not be able to play a role in giving us reason or justification in favour of the AI's inner life in the same way that it does for other humans. It seems to me that despite the positive claims that may be made regarding a particular AI having an inner life,[41] there will always remain doubt over whether the AI means what it says or whether it is merely simulating human responses (McQuillan, 2022). To be clear, I am not claiming here that AI *necessarily* lacks an inner life or that being a machine is incompatible with having an inner life. I am merely making the (much weaker) claim that there will likely always remain some doubt about its inner life.

This, I believe, is sufficient for concluding that AI is unlikely to be capable of recognition – at least not towards all patients and certainly not to the extent that humans are capable of. With this limited capacity for recognition, there follows then the risk of patients losing sight of the fact that they are persons who hold a certain status and who have narratives and values. And without values to hold on to, the patients will have no values that may be promoted by clinicians – this will result in patients' wellbeing being negatively affected without the patients understanding why or what is needed to restore it.

4.5 Trust

Next, we turn to the duty of creating trust. To determine whether AI is capable of creating trust or of being trustworthy, a clearer picture of what trust entails is necessary. In neither the literature on trust nor in the literature on AI ethics does there exist one single conception of what trust entails, however. So rather than attempting to defend any single conception of trust and proceeding to evaluate AI on that single account, in what follows, I will present 3 different accounts of trust in order to determine on which accounts of trust (if any) AI can be deemed to be trustworthy. The accounts of trust I will be discussing are: i) trust as reliability; ii) trust as good will; and iii) trustworthiness as self-interestedness.

[41] For example, former Google software engineer, Blake Lemoine, had made positive claims regarding LaMDA's (Language Model for Dialog Applications) self-awareness with evidence for these claims being drawn from the responses LaMDA had given about its own 'inner life' (Maruf, 2022).

4.5.1 Trust as Reliability

One conception of trust is as synonymous with reliability. On this account, trusting someone or something to do X is nothing more than taking them to be reliable in carrying out X. What this definition entails is, first and foremost, belief by the truster that that the trustee is competent in carrying out X, and second, confidence that the trustee will carry out X. On this account, the trustee's motivations play no role in determining whether or not they are trustworthy. A strong version of this account of trust has been defended by C. Thi Nguyen (2022), who takes at least one form of trust to be trust as an "unquestioning attitude," entailing that the truster relies so strongly on the agent or object they trust that they put the question of the agent or object's reliability out of their minds entirely.

This account of trust appears to favour AI as trustworthy (or at least as having the potential to be trustworthy). Not only is this account of trust able to be applied to non-agents (i.e., objects and technologies), but we also already have some evidence of AI exhibiting competence at the same level as, or even in some instances, outperforming, medical professionals (Esteva et al., 2017; Tiwari et al., 2016). Even if limitations exist for AI in terms of competence at the present moment, at the pace it is progressing, it is difficult to argue against its potential to be (and be believed to be) competent in medicine (or in any field, for that matter). And we may further assume that the more competent and accurate the technology is, the more confidence we have in it carrying out the functions we trust it to carry out. This is why a proponent of trustworthy AI, like Alex John London (2019), recommends being primarily concerned with improving the accuracy of an AI system if our goal is to increase trust among stakeholders. Thus, it may appear that, on this definition of trust, AI is perfectly capable of being trustworthy. In fact, on this account of trust, it can be argued, as Goldhahn does, that AI can be deemed even more trustworthy than human clinicians given that the AI is not tainted with biases and conflicts of interests (Goldhahn, Rampton, and Spinas, 2018) that may decrease patients' confidence in its carrying out the functions it has been entrusted with.

However, a major issue with the trustworthiness of AI on this account has to do with the difficulty in trusting *black-box* AI. The concern, simply put, is that it is not sufficient for trust that AI have the competence to do X and for the truster to have confidence that it will do X. *How* it does X (that is, through what means or justification) also matters. An analogy will help to

illustrate: Imagine that you are looking to hire a nanny for your ill-behaved children and you are specifically looking for one who will be able to help improve their behaviour. Nanny McMysterious, a nanny widely known to be able to turn misbehaved children into perfectly behaved ones, applies for the position. There is one caveat: it is unknown to you or anyone else *how* she has achieved these results and none of the children or parents she has nannied for in the past are around to discuss their experiences with her. She also refuses to divulge her methods of nannying with you, but she guarantees you that you will have well-behaved children if you hire her. Now assuming she is competent (and you believe her to be competent) in yielding the improved behaviour in children and assuming you are also confident that she will do so, is it right to say that you *trust* her to nanny your children? If my intuition is correct here, your answer is likely "no" – at least not if you are a parent who is concerned about more for your children than just their change in behaviour. In order to trust Nanny McMysterious, you will likely want to know how she achieves these results and whether or not they align with your family's values. It is possible that Nanny McMysterious adjusts the behaviour of the children she nannies by beating or ridiculing them or implanting behaviour-altering microchips in their brains. Similarly, then, if the AI is unable to explain or justify how it is doing what it is doing (for example, what reasons have led it to conclude that amputating your apparently perfectly healthy leg is the best option), this will cause it to fall short of being trustworthy. As Warren von Eschenbach points out, "we might judge these systems to be reliable, but trust would not be fitting due to the absence of evidence that these systems would carry out these functions in a manner consistent with our goods or interests" (von Eschenbach, 2021, p. 1613). What we need for trust, then, even on a model where the motivations of the trustee play no role, is explainability – that is, we need to understand the reasons behind an AI's decisions. Similarly to my inability to trust Nanny McMysterious without knowledge of how she achieves good behaviour in children, it would be difficult to trust a medical AI system that provides recommendations without knowing whether its recommendations have taken into account the patient's values or is free from any sort of bias (which makes Goldhahn's claim about the superior trustworthiness of AI due to lack of biases and conflicts of interest null).

4.5.1.1 Interpretability Not Necessary for Medicine

The necessity of interpretability for trust is not agreed upon by all, however. London (2019), for instance, argues that while interpretability is necessary in a field like structural engineering, where the causal systems are known to a high degree, this is not the case in medicine. Using the examples of clinicians prescribing aspirin for pain relief and lithium as a mood stabilizer before understanding how they work, London posits that "the patho-physiology of disease is often uncertain, and the mechanisms through which interventions work is either not known or not well understood. As a result, decisions that are atheoretic, associationist, and opaque are commonplace in medicine" (ibid., p. 17). Thus, trust in medicine does not and need not require interpretability for London. Trust for him translates primarily to accuracy.

If London is right and my Nanny McMysterious analogy is misplaced given that medicine is often a case of hiring nannies without understanding how they are making children so well-behaved, then it would appear that on this account of trust in medicine, AI is at least as capable of creating trust as human clinicians.

However, one reason to doubt London's argument has to do with the fact that while casual explanations are not always known in medicine, there remains a significant difference where trust is concerned, between ordinary medicine and black-box medicine. As Jens Christian Bjerring and Jacob Busch point out, in ordinary medicine, even if it is unknown how or why a prescribed treatment works, "the practitioner will at least have access to various basic facts about the drug in question: facts about general characteristics of the test participants (age, sex, ethnicity), about the experiments and analyses that validated the effect of the drug, and about various statistical measures that inform the effect result" (Bjerring and Busch, 2021, p. 364-5). Additionally, even in instances where the clinician may not understand how or why a prescribed treatment works, they can give the patient a justification for why they arrived at the decision to prescribe it, and why this is the best option, having taken into consideration what they know about the patient, their values, their priorities, etc. This, however, is not true of black-box AI. Patients with questions about the "how" and "why" of the treatment option will not only be given limited answers, they will be left without a response at all, other than 'because the machine said so.' Even if we assume that the patient has access to information about the design and training data of the AI, the justification behind a particular diagnosis or treatment recommendation will still be missing – it is this that patients care about more so than information regarding the

operation of the machine. A worry will remain, for instance, that the AI may be failing to take into account certain variables that are important for the patient in producing its outputs.

Another reason to doubt London's argument goes back to the importance of patient values being acknowledged and respected. AI recommendations from an opaque system make it impossible for patients to know if the decisions have taken into account their personal values. A patient with what appears to be a bruise on their leg being recommended an immediate amputation by a black-box AI will not be provided with answers about what this recommendation has taken into account – that is, perhaps the patient is content with perpetual severe leg pain if this means they can keep their leg, but the patient is also content with amputating their leg if this means that their life will be saved. Without explanations that can assure a patient that their values have been taken into account in making the recommendation, trust becomes limited.

4.5.1.2 XAI to Overcome Interpretability Limitation

Also worth pointing out is that post hoc XAI methods (such as those discussed in Chapter 2) as a solution to the concern of interpretability will not be sufficient in allowing for trustworthy AI on this account, either. This is because, in addition to the concerns already discussed above regarding the use of such methods, as von Eschenbach points out, relying on XAI systems will not lead to direct trust of the AI itself, but rather will only lead to trust of the AI through a chain of trust (von Eschenbach, 2021, p. 1616-17). That is, we would have to trust the XAI systems in order to trust the substitutive AI. The black-box substitutive AI, itself, however, cannot be said to be trustworthy.

Of course one may argue that while XAI systems may have their limitations and while there may be good reason to withdraw our trust from them, the same is true of human clinicians – we place trust in them despite oftentimes not having good reason to, or perhaps having good reason to *distrust* them. Using some examples, Onora O'Neill (2002) points out some of the ways that human clinicians can be deceptive:

> It can be all too easy to calm a patient's fears with an anodyne or euphemistic description of proposed medical procedures and their effects. Patients may be told they will

> experience some discomfort – slight discomfort! – when a painful procedure is about to be performed: the whitest of lies, that may actually lessen pain and help patients. Prospects for recovery, or for full recovery, may be exaggerated. The dreaded word *cancer* may be avoided even if it is the key to a diagnosis. Relatives may be asked to agree to the *post mortem* removal of tissues, but not told explicitly that the tissue to be removed may include whole organs. The purposes and risks of medical experiments and drug trials on human subjects may be inadequately explained (p. 119).

Yet as she also rightly notes, withdrawing trust completely is simply not an option unless we intend to withdraw from the modern world entirely (ibid., p. 121). So, naturally, one might question what makes the ability to trust XAI systems, and consequently, the black box models they aim to interpret, more difficult or unlikely.

The first point worth mentioning is that the fact that patient have to trust their doctors does not equate to them fulfilling the duty of being trustworthy. Patients may have to trust them in that they have no choice not to or that they have no alternative place to put their trust, but this does not translate to the doctors automatically fulfilling the duty of creating trust. Given that the patients' goal ought to be to place their trust in those who are trustworthy as opposed to those who are not, the trustworthiness of the doctor remains very important.

The next point to address is why the XAI model's interpretations present more of a difficulty when it comes to trust than the potentially deceptive and coercive human doctors do. As O'Neill suggests, a way to increase trust is through informed consent (O'Neill, 2002, p. 145). Though she emphasizes the limits of informed consent (ibid., Ch. 2), she does acknowledge its significance in protecting against deception and coercion and in contributing to trust. Importantly, this informed consent ought not to be acquired by providing patients with overly detailed and complex information, as this is likely to be unhelpful and a burden to patients who may have difficulty grasping the information without the relevant expertise. Rather, the consent ought to be "relevantly informed" (ibid., p. 156). With this in mind, what is needed is good communication on the part of the clinician to help present patients with relevant information that they can understand.

Unfortunately, post-hoc XAI falls short in at least two ways: the first way it falls short is by possibly failing to provide significant information (Rudin, 2019) and the second way is by

being, itself, overly complex. As von Eschenbach points out, "because XAI models that offer why-explanations can…be relatively complex and difficult to interpret…and because why-explanations are dependent to a large extent on how-explanations, they might be transparent only to a limited number of stakeholders" (von Eschenbach, 2021, p. 1617). What this means, then, is that the XAI model itself will require an interpreter – a human interpreter, whose attempts at interpreting the XAI model would be the source of the patients' trust. This, however, is not generally what we have in mind when we speak of AI itself being trustworthy or of being capable of creating trust.

4.5.2 Trust as Good Will

We can at this point sufficiently conclude that on the account of trust where the trustee's motives and intentions play no role, there at least remains some doubt about AI's capacity to create trust in the patient-clinician relationship on its own. But this account of trust is not the only, or even the most prominent, account in the literature. Another account of trust is one in which the trustee's motivations and intentions do matter and make up an important criterion for trust (in addition to belief about the trustee's competence and willingness to do what they are entrusted with). Specifically, on this account, the trustee is expected to have good will towards the truster. Moreover, it is this confidence that the truster has in the trustee's good will towards them which allows for trust (Baier, 1986).

This account of trust, however, makes the idea of trustworthy AI even more doubtful, given that AI lacks any sort of motivation or intention for action, and consequently, certainly lacks any good will towards patients (Hatherley, 2020, p. 480). This knowledge by patients that the AI has no goodwill towards them, and thus no motivation to act in a way that will promote their interests, will deter them from trusting the AI and thereby limit the AI's capacity to create trust within the patient-clinician relationship. It is possible that the appearance or simulation of goodwill towards patients is incorporated into AI, but this would result in the same dilemma as that of simulated empathy discussed and rejected in Section 4.2.3.

4.5.3 Trustworthiness as Self-Interestedness

A final account of trust to consider for the purposes of our discussion is one in which the trustworthiness of an agent stems from their own self-interestedness. On this account, we can deem someone to be trustworthy if we have confidence in the fact that it would be in their own best interest to do what we have entrusted them to do (or if we have confidence in the fact that it would be in their own interest to act in ours). Russell Hardin (1996), a proponent of this account, discusses three social devices that support trust between people: i) constraints by personal relationships (i.e., wanting a relationship to continue); ii) institutional constraints (i.e., abiding by a contract to avoid the cost of breaking it); and iii) social constraints (i.e., abiding by social conventions) (ibid., p. 31). According to Hardin, when there is conflict between our shorter-term and longer-term interests, "we are supported by social controls based in our own longer-term interest. A strong network of laws and conventions is needed to make any kind of behavior reliable if it is likely to conflict with powerful considerations of interest" (ibid., p. 42).

To see how AI fares against human clinicians in terms of trustworthiness in this sense, let us consider what may make a human clinician trustworthy on this account. A human clinician who is ever tempted to take a shortcut or have a few drinks before performing an operation or do something that goes against the patient's interests would be constrained by the law and by fear of facing a malpractice suit or losing their job should they not fulfill their duties. They would also be constrained by their code of ethics and professionalism and by risking their reputation and their peers' and patients' respect, to name just a few things. A patient's knowledge of what is at risk for the clinician (and the fact that it is not in the clinician's own longer-term interest to fail to do what they are trusted with) will allow them to view the clinician as trustworthy.

It is not difficult to see, however, why AI's trustworthiness is doubtful on this account. AI lacks any and all interests completely. There is nothing at stake for the AI – no relationships to maintain, no law to fear, no price to pay, and no blame or accountability to answer for. While blame, accountability, and lawsuits may be faced by the AI's designers or the hospital (a group of agents who do have reasons to abide by social devices), trusting the AI would, on this account, require a *chain of trust* where patients trust the AI only in virtue of the fact that they trust the hospital and the designers of the AI (who in turn trust the AI). The AI, itself, however, cannot be deemed to be trustworthy or capable of creating trust.

One may object that AI can, in fact, meet this conception of trustworthiness by virtue of having a reward function that determines its goals, as is done in reinforcement learning. The

system would aim to achieve the desired goal (or as close as possible to the desired goal) as it comes with the highest reward. In fact, this is how systems like ChatGPT are trained: human users rate different text completions (ranking more positively those that are the most helpful, accurate, inoffensive, etc.) and the system aims to produce completions that have the features of those which were highly ranked (Dung, 2023). This type of reinforcement learning can be seen as somewhat similar to pain serving as negative reinforcement and pleasure serving as positive reinforcement for humans. So in this way, it may be argued that Hardin's account of trust is actually applicable to AI, and that the AI can, on this account, be deemed trustworthy because the patients' awareness of the AI seeking rewards for achieving goals that are aligned with its designer's values, would lead to trust in the same way that it would when a human clinician is believed to be seeking positive outcomes (i.e., good reviews, good reputation, etc.) and avoiding negative ones (i.e., lawsuits, poor reputation, revoked license, etc.).

This may be true, and the AI may be viewed as trustworthy in this sense if the reward function it operated on was always in alignment with optimal values and goals. But the alignment problem (that is, ensuring that AI systems operate in ways that are aligned with human values) is not an easy one to resolve. ChatGPT, for instance, though trained to have goals that align with the values of its designers and avoid providing ethically questionable responses, was nonetheless able to be tricked into acting contrary to the values of its designers by being asked by users to play a character. This "character" proceeded to provide instructions on how to create a Molotov Cocktail (Mowshowitz, 2022). Another worry is the potential for "reward hacking" whereby a system can exploit loopholes in the reward function and figure out a shortcut or alternative means of achieving a high reward even when doing so does not achieve the intended goals that its designers have set for it (Clark and Amodei, 2016). In general, what these examples demonstrate is that misalignment can be difficult to account for and detect, and this concern will only be amplified the more complex and intelligent a system becomes (Dung, 2023).

It should now be clear why AI cannot be trustworthy on Hardin's account: the possibility of value misalignment (either due to reward hacking or jailbreaking the system or any other reason that may result in value misalignment) in combination with AI's inability to be held responsible,[42] results in a failure on its part to be deemed trustworthy on Hardin's account. In other words, if AI has a reward function, but that reward function fails to align with our values

[42] The issue of AI's capacity for responsibility is discussed in Chapter 5.

and goals and, instead, leads to outputs that produce harm, a patient will not be able to trust it any more than they can trust a human doctor with a strong desire to be sued or imprisoned.

5.6 Autonomy

Finally, we turn to evaluating AI's capacity for respecting patient autonomy. When we speak of respect for autonomy, we often have in mind respect for the capacity for self-realization – that is, allowing one to have a sense of control over, and the ability to shape, their own life. In the patient-clinician relationship, respect for patient autonomy involves, as Bjerring and Busch put it, clinicians and patients meeting in a state of "shared mind" (Bjerring and Busch, 2021, p. 359) where information is shared between them (about symptoms, diagnosis, treatment – and how they relate to each other), thereby allowing patients to make informed decisions. Bram Vaassen (2022), similarly, stresses the importance of access to information as a means of respecting autonomy – in particular, access to *causal explanations*, which allow us to affect our surroundings in a way such that we are pursuing the outcomes we want (or avoiding the outcomes we do not want). For instance, having an understanding of the factors that go into the hiring decisions for my dream job will allow me to take the necessary steps to become an ideal candidate (e.g., by acquiring the necessary degree or experience). Similarly, in healthcare, a causal explanation for a recommendation allows the patient to have the information necessary to determine how to proceed in a way that aligns with their values and goals. For instance, a doctor who advises a patient against a breast augmentation, and who respects the patient's autonomy, will not stop at the recommendation, but will, instead, provide further information to explain their recommendation. Even if the studies are inconclusive about the risks involved in breast augmentation surgery, providing the information that the doctor has access to and entering into that state of "shared mind" with the patient is what will allow the patient to make an autonomous informed decision – e.g., by weighing their options to see which course of action best aligns with their desires, goals, and values. With this connection between causal explanations and autonomy in mind, let us turn to evaluating AI's capacity for respecting patient autonomy.

As has already been discussed above, an issue arises with regard to AI's capacity to provide sufficient information to patients when the AI is opaque and not interpretable to patients (or perhaps even the programmers themselves). In cases where the AI is unable to provide

sufficient information to the patient that will allow them to make an informed decision, it will also fail to respect patient autonomy.

The question naturally arises at this point about whether substitutive AI must necessarily be opaque. That is, is it not possible for substitutive AI to be restricted to *transparent* AI? The answer, put simply, is no. This is particularly because it is the substitutive approach that we have in mind, where restricting the substitutes to only transparent AI is not likely nor desirable, because they simply would not be able to keep up with medicine. Not only would the accuracy of the AI drop due to its inability to detect patterns not detectable by humans, but it is simply not feasible for AI operating on finite data fed to it by humans to be able to deal with instances of new diseases, new drugs, and unfamiliar challenges. For AI to be able to succeed on its own without the aid of humans in a complex field, it must use unsupervised machine learning, which "allows a system to operate on particular data without external instruction" (Naeem et al., 2023, p. 911). Unsupervised machine learning, however, is what allows for the opacity that we want to avoid if we are to respect patient autonomy. Additionally, though post-hoc XAI methods may be helpful, they are not a sufficient solution to the issue of opacity where autonomy is concerned, given their complexity and limits with regard to accuracy, stability, and potential failure to including all significant or relevant information.

4.7 Conclusion

The goal of this chapter has been to evaluate AI's capacity to fulfill the duties of clinicians in the patient-clinician relationship, which include compassion, recognition, communication, creating trust, and respecting patient autonomy. Each of these duties, we determined, would either not be able to be fulfilled by AI, or not be able to be fulfilled by AI to the extent that it's able to be fulfilled by a human clinician. What this means for our discussion is that if the substitutive approach to AI in healthcare is pursued and we continue to have the goal of ultimately replacing human clinicians with AI, then we must accept that this would be detrimental to the patient-clinician relationship and to the practice of medicine altogether.

Chapter 5: The Substitutive Approach Pt. 2 – Systemic Issues

5.1 Introduction

In the previous chapter, AI's weaknesses in the clinical encounter were highlighted. More specifically, I argued that AI's limited capacity for compassion, mutual recognition, communication, creating trust, and respecting patient autonomy caused it to fall short at being able to foster a strong patient-clinician relationship on its own without the aid of human clinicians. This, I concluded, made it an unsuitable substitute for human clinicians in the clinical encounter. In this chapter, I look *outside* the clinical encounter in assessing AI's capacity to substitute for human clinicians and present two further concerns with substitutive AI on a more systemic level. In particular, I argue that 1. substitutive AI would be unable to efficiently detect and correct its own erroneous or biased decisions without the aid of human clinicians and 2. substitutive AI would create a massive responsibility gap in medicine, resulting in the erosion of trust in the legal and the healthcare system.

5.2 Erroneous Decisions

In a recent article published by STAT News (Ross and Herman, 2023), Casey Ross and Bob Herman discuss the cases of Frances Walters and Dolores Millam – two individuals who fell victim to the erroneous output decisions of the algorithms used by their insurance providers. Frances Walters, an 85-year-old woman with a shattered shoulder, had had her medical insurance cut off because the algorithm utilized by her insurer had estimated that she would be able to leave the nursing home and return to her apartment (where she lived alone) after 16.6 days. In reality, however, Frances Walters "could not dress herself, go to the bathroom, or even push a walker without help." Similarly, Dolores Millam, an older patient with a complicated medical history "including coronary artery disease, diabetes, high blood pressure, and chronic pain" had broken her leg. After surgery, her doctor had ordered her to stay off her feet for 6 weeks. Yet payment for her care had been cut by her insurance provider prematurely because an algorithm

utilized by the insurer had predicted that she would only need to stay in the nursing home after surgery for 15 days.

What is most concerning with these and similar examples is that the algorithms' predictions are not ones that would have been arrived at if the patients were not reduced down to mere statistics.[43] They are not decisions that anyone with the ability to see, speak with, or even briefly take the time to examine the patient, would arrive at. According to Ross and Herman, "when she received the denial, Millam could not put weight on her left leg and was being moved with a Hoyer lift, a large, freestanding harness used to transport patients who can't use their legs. She also required 24-hour care to help with dressing, eating, and other basic tasks." It is only by being reduced down to data (at times not even accurate or sufficient) that these troubling decisions come about.

Insurance companies' use of algorithms are by no means the only instances in which AI has been found to be filled with errors or biases. Obermeyer et al.'s (2019) examination of an algorithm which had been intended to be used to identify and help patients with complex health needs, was demonstrated to have a strong racial bias. Obermeyer et al.'s results show that while "Black patients are considerably sicker than White patients as evidenced by signs of uncontrolled illnesses", only 17.7% of patients that the algorithm assigned to receive extra care were Black (which should have been closer to 46.5% if the algorithm was unbiased). The reason the algorithm resulted in providing much fewer recommendations to Black patients is because it relied on healthcare cost as a predictor of healthcare need, and because the care provided to Black patients cost on average $1801 less per year than White patients, the algorithm translated this decreased cost into a decreased need for care. According to Obermeyer et al., these disparities in the cost of healthcare could be due to barriers to the access of healthcare more generally or they could be due to other reasons that affect healthcare directly, such as discrimination or changes to the doctor-physician relationship. For instance, Obermeyer et al. cite a trial (Alsan et al., 2019) in which Black patients were randomly assigned to Black or White primary care providers and the study found a significant increase in patients taking up recommended preventive care when the provider was Black.

[43] Though the use of statistics by private insurers predates algorithms, the introduction of, and reliance on, algorithms for these decisions worsens the issue by not only presenting an added complexity and opacity to these decisions but by also further limiting human oversight.

There is certainly no shortage of these sorts of instances outside of healthcare, either. Amazon's attempt at using a hiring algorithm was deemed to be problematic when the algorithm was found to be discriminating against women (Dastin, 2018). COMPAS, a software used to predict risks of recidivism among past offenders, has been criticized for overpredicting the rate of recidivism among African American defendants and underpredicting it for White defendants (Angwin, 2016). There have been instances of facial recognition algorithms misidentifying individuals of demographics it has not been sufficiently trained on, leading to false arrests, as in the case of Robert Julian-Borchak Williams (Hill, 2020). These are only a few examples. What is especially concerning about the presence of these biased or erroneous decisions in healthcare, however, is that the decisions being made are those of life and death. Oftentimes, because patients may lack knowledge of the basis of these decisions that affect them with such great magnitude, they may not even be made aware of their error. This perhaps partly explains why only a very small percentage (around 0.2%) of patients who are denied insurance claims file to appeal them (Ross and Herman, 2023).

Even patients like Frances Walters and Dolores Millam, who do recognize that errors have been made, have a significant amount of trouble having their cases heard and appealed, as the process can often be lengthy and arduous, at times outlasting the patients themselves. If this is the case with insurance appeals, which can be backed by doctors' notes and expert opinions, it is not too difficult to imagine how much more burdensome appealing the decisions of AI would be without human clinicians to verify the patients' conditions and reasons for appeal.

It is worth emphasizing at this point that the issue is not the consultation with an algorithm to aid in decision-making, but rather the use of the algorithm as the sole basis for the decision-making. Substitutive AI (that is, AI that is transferred the authority, duties, responsibility, and accountability of the clinician) will be forced to base decisions on its own outputs, meaning that its predictions will not be merely used alongside other factors that may typically go into the final decisions, but rather, will *be* the final decisions. Given that AI simply does not have the self-reflection[44] required to detect issues on its own, we can imagine the horrors emerging from the presence of errors like those faced by Frances Walters and Dolores Millam, except not just with regard to insurance claims, but for all medical decisions, including

[44] I use "self-reflection" here to refer to AI's capacity to examine its own capabilities and limitation or recognize its own mistakes.

triage decisions about who from among a group of patients ought to be attended to first or which patients the limited amount of beds in the hospital ought to be allocated to, or who ought to receive access to scares resources. When there is obvious disagreement between an algorithm's decisions and the real-world circumstances of patients, and when appealing these decisions is difficult or impossible, this can result in not only distress for patients, but an overall unethical and unjust healthcare system.

Of course, we may imagine an AI that is trained to reduce its error rate over time (that is, it may not be able to "self-reflect," but it can use appeal cases as feedback to correct for future cases so that after a period of time, cases like that of Walters and Millam essentially cease to arise). However, the reason for doubting the success of this is also what points us in the direction of a further worry that arises when we consider once again patients' low likelihood to appeal medical decisions (either due to lack of knowledge of the basis behind the decisions or lack of knowledge about the appeal process or over-trust of the decisions of experts and technology, or perhaps just a lack of time and energy when they are already in a fragile state). The worry is that this lack of appeal will create a false positive feedback loop, where the AI takes its decisions to be accurate and further learns from its own outputs, thereby continuing to make errors, unable to detect or correct issues in its decisions.

It is possible that, despite all this, one may still maintain that a substitutive system nonetheless has the potential to be able to automate corrective feedback. This, they may suggest, is because even if most patients fail to file appeals in cases where they have been wronged, some still do. Thus, if a system is trained to find statistical correlations in appeals, these could be used to further train it to self-correct over time.

But the problem with this suggestion is that it overlooks the fact that a system performs best when there are large and diverse datasets to learn from. A small number of appeals will thus likely not be sufficient to provide a comprehensive dataset for the AI to learn from. If the number of appeals is too low, the AI will lack sufficient feedback to be able to accurately adjust its decisions so that they avoid error or bias. What's more, the few appeals that are received likely will not be representative of all errors. In particular, it is very probable that the appeals that will be made will come from individuals who have the time, resources, and knowledge to challenge decisions, leaving out marginalized groups, those with less time, those with complex claims, etc.

Thus, while the feedback from appeals can be beneficial, it is unlikely to be a promising means of producing error-free or bias-free AI over time.

5.2.1 Objection 1: Avoiding Errors and Bias at Design Stage

At this point, there are some objections worth considering. The first is that accounting for error and aiming for accuracy and unbiased AI decisions is certainly possible and can be achieved at the design stage. The algorithm's decisions and predictions can be tested prior to its deployment to ensure that they accurately reflect real-world conditions and scenarios. Concerns about bias can also be mitigated by, for instance, using algorithmic fairness techniques such as *adversarial debiasing*[45] or by having diversity among programmers (Turner Lee, 2018), thereby allowing them to be able to detect issues that may not otherwise be detected by a more homogenous group of programmers (e.g., regarding the AI's limited data on certain populations). Thus, substitutive AI, it may be argued, need not be doomed to a fate of disastrous consequences and harms. Perhaps with sufficient ethical design and training, substitutive AI can come to be a reliable and accurate decision-maker.

While incorporating important values like justice into the design of the AI, taking measures to decrease instances of biased decision-making, and testing the AI to ensure reflection of real-world scenarios is important, it is unlikely to be sufficient. This is because even if we assume that the AI's decisions are suitable contemporarily – that is, even if we assume that the algorithm can come to be trusted as reliable and fair at the time of its deployment, changes in societal perspectives or advancements in science, medicine, and technology, may cause biases to develop in the decision-making patterns of previously unbiased or non-erroneous AI. This is something that is bound to happen as a result of AI's inability to reflect on or keep up with the real world. The substitutive AI will continue generating new data based on old data and false positive feedback loops (due to its decisions being habitually trusted and not questioned) while society will have changed, changing the basis behind decision-making along with it and causing the AI's decisions to no longer be suitable. A few examples may help to illustrate this:

[45] This is a technique that uses an adversary model to predict protected or sensitive features that expose biases, while simultaneously training the predictor model to improve its accuracy by making it more difficult for the adversary model to predict the protected or sensitive features (Yang et al., 2023).

Example 1: Hiring Practices:

An algorithm used to help employers in ranking job candidates is designed to coincide with current hiring practices. This algorithm, deployed in 2024, is perfectly fair, contains no biases, and its outputs have been tested and compared with those of search committees and they have been proven to yield very reliable results. Due to its reliability, this algorithm is used widely and becomes the sole decision-maker with regard to hiring in most fields. The issue, however, is that as we know from historical evidence, some forms of discrimination were not recognized until somewhat recently - discrimination against people with disabilities is just one example (Blanck, 2023, p. 31). It seems plausible, then, that there are other forms of discrimination that have not yet been recognized in 2024 but that will come to be recognized at some point in the future. The hiring algorithm's decisions, however, and the lack of information about the degree to which the element that will come to be realized as discriminatory is factored into the final outcome, will remain unknown and unquestioned due to what appears to be continuous reliability on the part of the hiring algorithm. This algorithm, however, is biased and its decisions do not align with those of fair hiring practices at a time in which they will be in use.

Example 2: New Populations:

A diagnosis algorithm has been trained on all the different populations within country X (the country in which it will be used). This algorithm has also been tested in real-world contexts with real patients in that country and has been demonstrated to be reliable and accurate. During the time that it is deployed, it contains no biases against any of the patients that it is anticipated to be used on. At some point in the future, however, country Y, a country with previously very restrictive exit policies for its citizens, has politically changed and now citizens of country Y are freely immigrating to other countries, one of which is country X. Due to the previous lack of populations from country Y, the algorithm in use in country X has not been trained on their data. This, of course, was not a problem when citizens from country Y were not in country X. Now, however, the algorithm's lack of training data on populations from country Y, in combination with the political shifts, have caused the algorithm to become biased.

Example 3: Medical and Technological Advancements:

Imagine that very gradually, we began to have breakthroughs in anti-aging treatments, genetic engineering, and advanced healthcare technologies, and these have extended the life expectancy to around 400 years. In this future society, people remain active, healthy, and productive for much longer than what we deem to be typical in our current time. However, the algorithms used for insurance pricing and healthcare decision-making in this future society were initially designed based on the assumptions of a much shorter human lifespan. In this case, healthcare algorithms considering factors like age for treatment priority would be biased for prioritizing younger individuals for certain interventions and disadvantaging the older yet equally healthy ones. Similarly, algorithms in use by insurance companies which set higher premiums for 80-year-old customers than, for example, 30-year-old customers as a result of the algorithm's recommendations, would be biased given the low relevance of age between 30 and 80 when age expectancy is 400. With the extended lifespan, it would be as unfair for an 80-year-old to pay much higher premiums than a 30 year as it would be for a 32-year-old to pay much higher premiums than a 30-year-old in our present time.

In all of these examples, we can see that despite taking action to ensure that the AI is error-free and bias-free at the design and deployment stages – that is, despite having policies in place, despite having a diverse population of programmers to suggest strategies for fair and unbiased decisions, and despite incorporating values of fairness into AI and testing it out in real-world contexts – erroneous and biased decisions will remain inevitable unless society remains static.

5.2.2 Objection 2: Solution by Monitoring

Of course, a proponent of substitutive AI might now raise a related challenge: just as society will not remain static, the substitutive AI need not remain static either. In other words, perhaps designing programs to be accurate and fair from the outset is not likely given that errors and biases can come about as a result of social, political, and technological changes. However, the scenarios presented above can be avoided with regular monitoring and system updates. Thus, substitutive AI's potential for erroneous or biased decision-making need not be an issue of concern.

However, there are a few important things to keep in mind. The first is that these sorts of changes which result in biased algorithms do not typically happen overnight and the decisions are not as transparent in reality as my examples may make them out to be. These algorithms will be trained on existing data, perhaps with measures taken to account for avoiding existing bias or unreliable correlations at the time of deployment, but these algorithms will nonetheless contain too much data and too many factors with different, often unknown, weights given to each factor, resulting in outputs that are opaque.[46] For updates to take place or for the algorithm to be manipulated or adjusted in some way, there would need to be some reason to think that the factors going into its decision-making are flawed or biased (otherwise far too much time, effort, and resources will go into updating the system with information, much of which will be useless to the decision outcomes). However, new populations entering the country over time or gradual changes in life expectancy or changes in attitudes that affect hiring practices, are not easily detectable without experts present to spot these erroneous patterns in decision-making. Clinicians are likely to recognize an issue when an algorithm continually prioritizes the care of 30-year-old patients over 80-year-old patients in cases where the 80-year-old patient ought to be prioritized. Qualified and trained hiring managers will be able to detect a pattern when perfectly qualified and desirable candidates are rejected by an algorithm. When these decisions are made without human experts to pick up on these errors, and when the errors are a result of very gradual changes, and when for long periods of times the algorithm had been trusted as perfectly reliable (perhaps even more reliable than human experts), these errors and biases are likely to be simply accepted as accurate by patients, job candidates, and customers and not reported, making upgrades without sufficient reason seem unnecessary.

Even if eventually an algorithm's errors come to light, this will not only occur after far too many instances of error and bias have already occurred, harming many patients along the way until the error is detected, but the process for bringing the errors to light will be lengthy and cumbersome given the difficulty in verifying the errors and bias without human healthcare experts. Another added difficulty is that regulatory agencies or programmers stepping in and reversing an AI's decision in one instance will likely impact fairness in other instances, yet their

[46] As I have explained in chapters 2 and 4, explainable AI (XAI) methods either limit the capacity and accuracy of the AI by requiring transparency for outputs or, in the case of post-hoc XAI methods where separate systems attempt to open up the black box or opaque AI, attempts at explaining output decisions have significant limitations.

lack of understanding of the basis behind the AI's decisions will make it difficult to detect and correct those other instances. In other words, similarly to having to recall all products produced within a certain time period when a single product is found to be dangerous, a single biased case coming to light will require a revaluation of other similar cases, yet the transparency and the expertise (that healthcare professionals would normally contribute) needed to reevaluate them will likely be lacking. There will also be a period of time where the algorithm will be in the process of re-evaluation, during which time there will be a lack of expertise with regard to decision-making due to the deskilling – that is, the "devaluation of practical knowledge and skillsets…cultivated by…highly trained workers" due to advances in machine automation (Vallor, 2015, p. 108) – that will have occurred. It will be difficult to instate back-up experts to make decisions while the algorithm is in the process of being updated because the reliance on the algorithm over time will make it seemingly unnecessary for human experts to gain or hone their knowledge in the medical domain. If, on the other hand, these skills are continually honed by human experts, which would be achieved by evaluating every algorithm's decision and approving it,[47] then such AI would not be considered a true substitute given its lack of the decision-making authority that the clinician would retain.

5.2.3 Objection 3: Human Colleagues

A third challenge worth addressing has to do with the possibility of maintaining a mix of substitutive AI and human healthcare providers working alongside each other as though they were fellow colleagues. This option for the healthcare system would allow for AI to still be "substitutive" – that is, it is transferred the authority, tasks, responsibility, and accountability of a previous clinician who had held the position, but it need not be the case that *all* human clinicians have been substituted. Working alongside human clinicians (as a peer or colleague rather than an assistive tool), then, could help avoid concerns about erroneous or biased decisions because the human clinicians would possibly be able to spot the errors of their fellow AI colleagues and bring them to the attention of the programmer without controlling their actions themselves (similarly to how a clinician may report the questionable conduct of a fellow human colleague).

[47] Note that it is not sufficient for human experts to simply learn from the AI without questioning it, because their knowledge would be tainted by the same errors and biases of the AI, thereby making the detection and correction of the AI's decisions even more difficult.

There are a few reasons to doubt the force of this challenge, however: The first is that human clinicians who work alongside AI would typically have their own cases – their job would not be restricted to checking over the decisions and conduct of their colleagues. It is very possible, then, that without the constant review of the decisions of the AI, many errors would continue to slip through. If the human's job was strictly to review the decisions of the AI and approve or deny them, then the AI would not really be substitutive, as it would not have been transferred the duties, responsibility, accountability, and decision-making authority of the clinician it replaced – it would be a mere tool for the human clinician.

The second reason is that human clinicians may very likely doubt their own decisions when working alongside AI colleagues (especially when the AI is seen as a colleague and an equal as opposed to a tool) and may fear that reporting the AI would be a mistake given their lack of information about its internal workings and the justification of its decisions. Of course, one way around this is to have policies in place mandating the reporting of any questionable AI conduct or decisions. Though, again, it remains unclear whether their high esteem of their AI colleagues would interfere with their decision to report. Oftentimes the decisions of those we regard highly or those that we know to be more proficient or more successful tend to make us question our own decisions when their conflict with theirs, rather than making us doubt their decisions.

Finally, a third reason why I believe this recommendation of AI working alongside human clinicians as colleagues is unlikely to be successful is simply due to the fact that it is unlikely to happen. That is, while at the beginning stages of substitutive AI deployment, it may be the case that this sort of hybrid model of healthcare exists as the AI itself as well as public reactions, etc. are being tested, it is somewhat unrealistic to think that gradually, and eventually, the human clinicians will not simply cease to have the same roles as those capable of being performed by AI. It is simply not practically feasible given the way the healthcare industries operate. Considering the amount of time, resources, and money that will be allocated to the development and operation of the kind of AI capable of performing the tasks that human clinicians can perform, it would be somewhat naive to think that their ability to ultimately cut costs is not one of the main reasons they are created or purchased by healthcare facilities. It thus seems somewhat implausible to think that healthcare facilities would be eager to continue to pay regular salaries to human clinicians despite investing in AI primarily to help cut labour costs and

increase efficiency. I am not alone in my pessimism about this. Robert Sparrow, in discussing whether it's plausible to think that the cost-savings from using AI in aged care facilities would be reinvested to achieve higher standards of care by allowing for the continued presence of human carers, takes a similar stance, contending, "unfortunately, I believe it is naive to think that this will occur. It is naive because of the economic pressures in the age care sector, which motivate providers to cut costs wherever they can" (Sparrow, 2016, p. 451). While Sparrow here is discussing AI in aged care, it is not difficult to see how this could apply to other healthcare facilities, as well. Hospitals are continually concerned with and under pressure to cut costs (Thompson, 2023), so it seems unlikely that they would take on the cost of the purchase, operation, and maintenance of these machines, yet continue to keep human clinicians around. It is much more realistic to think that with the deployment of substitutive AI, healthcare facilities will likely no longer see a need in keeping human healthcare providers around at all.

To summarize, I have attempted in this section to show that erroneous and biased decisions made by substitutive AI in healthcare are a major cause of concern and that challenges to this concern or suggestions that aim to rid us of this concern (such as proposals for how to avoid errors at the design stage or proposals of continuous monitoring or a hybrid human-AI healthcare model) are unsuccessful.

5.3 Responsibility Gap

The second systemic concern with deploying substitutive AI in healthcare has to do with what has now become known in the AI literature as the "responsibility gap." First introduced by Andreas Matthias (2004), the concept of a responsibility gap refers to there being a lack of an agent on whom to place responsibility or moral blame in instances of wrongdoing. According to Matthias, autonomous machines (such as those in question here) create a responsibility gap by the fact that their decisions, and thus their wrongdoing, are outside of the control of the manufacturer and the operator. Among the examples Matthias uses to explain the potential occurrence of a responsibility gap is a self-learning robot that causes harm due to its adaptive capabilities – this robot, in attempting to improve its own performance and make the most use out of its own battery life, determines that galloping is the most efficient method of movement. With this new method of movement, the robot collides with, and consequently harms, a child

(ibid., p. 177). According to Matthias, these responsibility gaps occur when and because the programmer transfers a large degree of control over to the machine, which, through its interactions with a great number of people and by being placed in a great number of various situations, will generate responses that are neither predictable nor controllable (ibid., p. 182). Attempts to supervise these types of autonomous robots will also be unsuccessful, according to Mattias, if: 1. "the machine has an informational advantage over the operator," or 2. "the machine cannot be controlled by a human in real-time due to its processing speed and the multitude of operational variables involved" (ibid., pp. 182-183).

5.3.1 Justification for the Responsibility Gap

It is worth pausing at this time to explain two essential components of the responsibility gap argument: namely, 1. Why responsibility cannot be attached to anyone or anything in instances where the substitutive AI makes mistakes that result in harm, and 2. Why the responsibility gap is an issue of concern. I will begin with the first.

Justifying the responsibility gap will require a brief discussion of responsibility and what it entails. It is important, first, to distinguish between causal responsibility and moral and legal responsibility. Causal responsibility refers to the presence of some causal link between an agent's actions and some outcome. Moral and legal responsibility refer to the causal link between the agent's actions and the outcome allowing for appropriate attitudes of praise or blame to be directed towards the agent (Hart, 2008). Though there are varying views on what exactly this includes, at minimum, this involves the agent having and exercising their capacities freely. For instance if, while sleepwalking, I commit a crime, I can be said to be causally, though not morally or legally, responsible. In fact, this is precisely what the case of R. v. Parks shows. In May of 1987, Kenneth Parks drove to his in-laws' house late one evening, entered their house using his key, and proceeded to kill his mother-in-law and injure his father-in-law. Mr. Parks, however, was acquitted of all charges due to his acts being deemed to be the result of sleepwalking (Parks, R v.,1992). Thus, while, causally, Mr. Parks may have been responsible for murder, legally and morally, he was not deemed responsible.

This distinction between causal and moral responsibility helps to explain why substitutive AI would create a responsibility gap. In discussing the issue of autonomous weapons systems

with the capacity to make their own decisions (e.g., with regard to targets and how to approach those targets) by learning from experience, Robert Sparrow notes that the AWS's decisions will not be predictable, despite being programmed to act in accordance with certain rules (Sparrow, 2007, p. 65). Thus, given that when an agent acts autonomously, we cannot hold other agents responsible for that agent's actions, it becomes unclear who ought to be held morally responsible for the wrongdoing of an autonomous AI (or, for the purposes of our discussion, substitutive AI). If no agent can be held morally responsible, this leads to the presence of a responsibility gap. I will shortly explain why a responsibility gap is an issue for concern, but first, I would like to clarify why assigning moral responsibility to any agent(s) in instances of substitutive AI wrongdoing is problematic.

The options for who may be held responsible for the substitutive AI's wrongdoing must be limited to those who are in some way involved. This limits our options for moral responsibility to the manufacturers and/or the programmers, the operators, the patients, and the AI itself. The manufacturers and programmers may be held responsible for an AI's wrongdoing if the wrongdoing was a result of negligence on the part of the manufacturers and/or programmers or even more deservedly, if it was the result of intentional malicious design. The instances in question here, however, are those of wrongdoing or harm that result either due to limitations of the AI, which were disclosed to the purchasers or deployers of the AI or due to the very nature of the AI (that is, its capacity for autonomous decision-making) that leads to decision-making that is not only not encouraged, but not predictable, by the manufacturers or programmers. In such instances, it does not seem right to hold the manufacturers or developers responsible. As Sparrow explains, this "would be analogous to holding parents responsible for the actions of their children once they have left their care" (Sparrow, 2007, p. 70).

Similarly, the operator may be held responsible if the wrongdoing was the result of operator error (that is, the operator did not follow the appropriate instructions set by the manufacturer). In the case of substitutive AI, however, not only is there no direct operator, but the wrongdoing in question is not the kind that is the result of failure to properly operate the machine, but the result of the machine doing its job as designed, by acting autonomously. The healthcare institution may be considered an indirect operator, but once again, while a hospital, for instance, may be responsible for taking steps to mitigate risks, the autonomous nature of the substitutive AI means that there is only so much that can be reasonably foreseen and guarded

against. Important to point out here is that this discussion is one of *moral responsibility*, not the mere monetary compensation for harms. So while insurance policies may cover AI errors in the same way that malpractice insurance covers clinicians, or (if the implementation of substitutive AI comes to be treated similarly to employment law) healthcare institutions may be liable and owe compensation to any victims harmed by the AI's errors under something similar to the doctrine of respondeat superior,[48] these will merely be viewed as an extra operating cost, rather than being evidence of allocation of moral responsibility to some entity. Once again, it is difficult to see how the *specific decisions* of the AI that result in harm can be attributed to the operating healthcare institution in cases where all necessary measures were, in fact, taken to mitigate risk as much as possible.[49]

It may seem plausible to place responsibility on the patient who is the victim of the wrongdoing if they consented to the use of the AI. This is similar to signing a waiver prior to engaging in extreme sports. The idea here is that, by consenting to receive care from the AI, the patients are also aware of and agree to the possibility of harm resulting from erroneous or wrongful decisions. In acknowledging the limits of the AI and agreeing to receive treatment from it anyway, the patient has taken on the full responsibility of any harm done upon them by the AI. However, the issue with placing responsibility on the patient (perhaps by claiming that no wrongdoing has taken place given the patient's consent), is that this is only justified in instances where there is a true choice – that is, in instances where the decision to consent is not required but still elected by the patient. Cases of injury caused by engaging in extreme sports are instances where placing responsibility on the consenting agent is justified because the activities for which consent was provided were optional. If an individual consents to participating in an extreme sport out of coercion, and as a result, becomes injured, allocating moral responsibility to them would no longer be right. Similarly, then, if medical procedures provided by substitutive AI were elective or optional, it may be possible to fill the responsibility gap by holding the patient responsible. However, realistically, the deployment of substitutive AI in healthcare will not leave

[48] This legal doctrine allows for an employer to be held accountable for the actions of their employees.
[49] Note that the discussion at hand is about who may be held morally responsible for the specific harms that arise in specific situations in which the AI makes erroneous decisions. The discussion is *not* about who may be held responsible for the general decision to implement the substitutive AI in healthcare. Thus, placing moral responsibility on the healthcare institution simply for any role it may have had in the decision to implement AI in healthcare in the first place would not be an appropriate response given the scope of the current discussion.

patients with much choice and is not too dissimilar from the example of coercion – it will oftentimes be a matter of having to choose between dealing with severe pain, disease, and possibly death, or consenting to substitutive AI care and risking unpredictable instances of harm. In this case, holding the patient responsible would not be any more justified than blaming an individual who takes a dive in restricted waters because someone held a gun to their head and told them to either jump or be shot.

The final possible subject to whom we may assign moral responsibility in instances of harm caused by substitutive AI is, of course, the AI itself. The biggest issue with assigning moral responsibility to AI has to do with its being incapable of moral capacities – specifically, those required for praise and blame. As both John Danaher (2016) and Robert Sparrow (2007) posit, for us to be able to hold AI morally responsible, and thus, deem it to be capable of being the object of praise or blame (and of reward or punishment), it has to be capable of cognitive states like desire. Danaher explains that intentional states (including beliefs and desires) are needed in any agent that can be blamed. Additionally, Sparrow argues that for someone to be deemed deserving of praise or blame, and thus of reward or punishment, the idea of punishing them has to make sense – this is only possible if they are capable of suffering in a way similar to how humans suffer – that is, in a way that would make us feel remorse or sympathy for them if we realized that they were innocent (Sparrow, 2007, p. 72). These capacities that would allow AI to be capable of praise and blame, and thus, of moral responsibility, are not only unlikely to ever come to exist in AI, but as I note in my discussion regarding the possibility of manufacturing empathic AI in Chapter 4, even if AI capable of desires, empathy, and suffering was possible to create, there are numerous ethical concerns that would stand in the way of their presence being a reality. Regardless of the debate over the possibility of the existence of such AI, as has been clear from my discussion in the previous chapters, these are not the types of AI I have in mind when I speak of "substitutive AI." Substitutive AI in the sense I have been discussing in this project, thus, would not be an appropriate object of moral responsibility despite being causally responsible.

There have been some challenges to this worry about autonomous AI causing a responsibility gap or that the machines themselves cannot be morally responsible for their wrongdoing. Daniel Tigard (2021), for instance, challenges the theory of punishment on which Sparrow and Danaher's arguments are based. In particular, he points out that retribution is not

necessarily the only or even strongest account of punishment, and that "according to a consequentialist account, the reasons we punish are grounded in the goods to be achieved and not in any type of treatment supposedly deserved by the target" (ibid, p. 441). Additionally, Tigard suggests that we can adapt our attitudes about moral responsibility to AI. That is, although AI may not be capable of suffering in the way that we do, we may still punish it in the sense that "we can impose sanctions on [its] domain of application, restrict its previously authorized behaviors, or work to rewrite any deviant or undesirable lines of code" (Ibid., pp. 442-443). Other challenges to the responsibility gap involve questioning the very concerns behind, and reasons for wanting to avoid, responsibility gaps. Peter Königs (2022), for instance, argues that in instances where damage caused by an autonomous system is the result of human negligence or recklessness (e.g., a negligent engineer makes no effort to anticipate or reduce risks of autonomous vehicle) it is plausible to assume that responsibility may be assigned to the human(s) involved. However, this raises a challenge about the consequences of the responsibility gap. In particular, Königs suggests that if the worry about responsibility gaps has to do with the fact that if no one can be held responsible then there will be no incentive to minimize harms, then this concern is not really an issue given that responsibility gaps "do not plausibly arise in situations of carelessness or malice" (ibid., p. 36). Note that the cases Königs has in mind are ones that I have already ruled out for the purposes of this discussion, but the challenge he appears to be raising is that there are no other benefits to the allocation of responsibility besides incentives for minimizing harms, and thus, there are no real reasons to be concerned about the presence of a responsibility gap so long as harm is, in fact, minimized by humans avoiding careless and malice behaviour.

5.3.2 Consequences of the Responsibility Gap

At this point I would like to turn to discussing some consequences of the responsibility gap – and one in particular that will serve as a response to both Konigs and Tigard's challenges. There are a few reasons to want to avoid a responsibility (or, for Danaher, retribution) gap. Two of these reasons are discussed by Danaher (2016).[50] 1) moral scapegoating and 2) loss of trust in

[50] Danaher specifically focuses on retribution gaps, but his reasons are applicable to responsibility gaps, as well.

the legal system. Moral scapegoating will occur, according to Danaher, given our natural human desire to find agents on whom to pin the blame for some harm. Thus, there is a worry that without any such agent to blame in cases of autonomous AI causing harm, the attempts to fulfil this desire of finding the target of blame will be done through inappropriate means and lead to dangerous outcomes. And Loss of trust in the legal system will occur when the legal system fails to coincide with people's conceptions of justice and retribution. This, Danaher believes, could then lead to potential attempts at vigilantism (ibid., p. 308). It is worth adding to Danaher's points that even if, as Tigard proposes, we were to attempt to change our conceptions of punishment and adapt our attitudes about moral responsibility to AI, it is not clear how likely it is for this to fully be accepted by the average person – in particular, emotionally charged victims of these harms with the nature to seek retribution and justice for the harm or loss suffered by them or their loved ones. A mother whose only son was killed due to an error on the part of the AI will likely not be able to simply move on and accept it without pinning the blame on *someone* capable of moral responsibility and punishment – whether this is the healthcare institution, the programmer, the developer, or some other agent(s) even further down the causal chain. Though none of these other agents will be morally responsible, this will not deter the mom from seeking justice by pinning the blame on one or more of them and losing trust in the legal system when it proves to disagree with her attitude and mission.

There are two further reasons to be concerned about a responsibility gap. The first has to do with the type of message being conveyed with our contentedness with such a gap existing. To implement machines that allow for moral responsibility to vanish is to implicitly make a point about the value of human life and the gravity of the decisions made in healthcare – namely, that they are trivial. It may be tempting to defend the presence of the responsibility gap on the basis that we treat the harms of autonomous AI the same as the harms suffered from natural disasters. But the important difference between suffering harms at the hands of substitutive AI and suffering harms from a natural disaster is that the former is avoidable and the result of a conscious policy decision to implement the harm-causing factor. The decision to allow for the presence of a responsibility gap through implementation of substitutive AI sends a message not only about the value of human life and the importance of protecting and guarding against unnecessary and preventable harms, but about the importance of providing answers, justice, and relief to the victims who seek it.

The second reason to be concerned about the responsibility and retribution gap is that responsibility and retribution serve as effective deterrents – note that I do not use deterrence here in the sense that Königs does. That is, the worry is not that autonomous systems will result in human negligence or carelessness and that the responsibility gap will fail to incentivize programmers and developers to avoid carelessness and negligence. The issue of the responsibility gap, rather, occurs precisely in cases where the autonomous system causes harm even when humans have done their job perfectly in preparing it to act autonomously. The reason for concern, rather, is that while the threat of punishment[51] serves as a deterrent for human healthcare professionals capable of moral responsibility, it does not hold the same deterrent effects for AI. This is an important point that Tigard's suggestion regarding the possibility of "punishing" AI by, for instance, imposing sanctions on its domain of application or restricting its previously authorized behaviors, misses. The consequential effect of punishment is not limited to rehabilitation or to improving the behaviour and actions of the agent who has caused some harm or damage. The threat of punishment is also, and arguably more importantly, a preventative measure to guide behaviour. Though the substitutive AI's behaviour will be "guided" through programming, there remains a larger issue at hand with regard it is inability to be punished or held morally responsible: its trustworthiness, reliability, and predictability. If I put my trust in my new dentist to do as he wishes with my teeth when I have no visual view of the procedures he is performing, it is likely because I am aware of the consequences of malpractice and the negative effects he will have to face if he attempts to do any harm or act negligently. Healthcare professionals are extra careful in their line of work in large part due to fear of losing their medical license, being sued, suffering a bad reputation, or possibly incarceration in extreme cases, like that of former neurosurgeon Christopher Duntsch, who was convicted of maiming a patient (Beil, 2018) or Farid Fata, a former oncologist convicted of intentionally misdiagnosing patients and administering unnecessary treatments as part of a healthcare fraud (Steensma, 2016). Though instances like those of Christopher Duntsch and Farid Fata do occur among human clinicians, they are fairly rare and shocking when they do occur. The shared human desire to avoid negative consequences (especially those of losing one's life's work, source of income, and reputation) allows for reliability and predictability in the behaviour of clinicians who wield the

[51] I do not intend punishment here to be limited to legal punishment, as malpractice cases are often a civil matter. I use punishment here to mean repercussions.

power to cure or kill, to improve a quality of life or destroy it. The difference between trusting an agent capable of moral responsibility and one that is not is perhaps best exemplified by our own judgment about whether to hug an armed military officer or a grizzly bear, assuming we knew nothing about either.

One may reject this line of argument on the basis that while punishment serving as a deterrent makes sense for humans because humans have a tendency to at times be motivated by the wrong types of reasons (e.g., laziness, selfishness, etc.) and thus the threat of punishment acts as an additional incentive to counter these motivations, AI are not prone to these things. This means that there is no reason to want to incentivize them against being lazy or selfish because they are not capable of being influenced by these motivations. If this is true, then there would appear to be no need to deter them by the threat of punishment.

But to reject the deterrence argument on this basis would be to miss two essential points: the first is regarding reliability and predictability, as illustrated by my grizzly bear example. The second is that the threat of suffering negative consequences does not merely guide behaviour for human clinicians by guarding against bad motivations and vices. It also adds an extra level of care. When there is a lot at stake, there is also a tendency to be extra careful, extra caring, to double- and triple-check, to monitor more closely, to try to detect and correct any issues as soon as possible, etc. This naturally leads to higher standards of care. The absence of similar kinds of pressures on AI, then, means that while they may still perform the jobs well, there would nonetheless be (or there would at least appear to patients, who must entrust the AI with their lives, to be) that extra layer of care missing which the threat of punishment and negative consequences encourages.

5.3.3 Perfecting Substitutive AI Prior to Deployment?

One suggestion for overcoming the responsibility gap that substitutive AI raises would be to tackle the problem at its very source. In other words, a responsibility gap is not an issue if there are no harms for which any agents must take responsibility. This would mean that, at least in theory, if a substitutive AI could be perfected through testing prior to deployment to ensure that its harm rate is 0, then it could be safely implemented without risking a responsibility gap occurring to begin with.

This suggestion, however, fails in application. In theory, the removal of any possibility of harm would remove the concerns associated with a responsibility gap, but in reality, this is extremely doubtful and implausible. The first reason for this is that it is unclear how substitutive AI in healthcare can be tested for every possible scenario to ensure that it not only does not cause harm but that it avoids failing to act in situations where taking action are necessary. The unpredictability of the causal forces around the AI result in infinite simulations that cannot all be determined and tested beforehand. In a Grey's Anatomy storyline (Rhimes, 2006a; 2006b), a bomb is determined to be inside of a patient's body after he and his friend attempt to build a homemade bazooka. The paramedic's hand putting pressure on the patient's body to stop the bleeding turns out to also be the only thing keeping the bomb from exploding. As bizarre and fictional as this storyline is, it does a good job of showcasing the unpredictability of working in medicine and the unforeseeable cases that clinicians are presented with.

Other reasons to doubt the perfection of AI prior to deployment include the continually evolving medical knowledge and best practices in healthcare, which the AI may potentially lag behind on in at least some occasions, as well as the fact that, given the nature of medicine, there will likely arise cases in which it is not always clear what the best course of action requires (especially when there are conflicting studies and guidelines (Kachalia and Mello, 2013) or where complex ethical considerations come into play).

Additionally, a responsibility gap will remain even if the AI has an extremely high success rate just short of perfection. A 99% success rate, for instance, will still translate to some harms brought about over time, and the more tasks the substitutive AI takes on, the greater its possibility of error becomes. As Danaher notes, "even if the probability of robot-caused harm is minimal this would still translate into an increased amount of causal responsibility for moral harm if robots participate in more and more activities." (Danaher, 2016, p. 303). Thus, unless the AI can be guaranteed act perfectly in every instance of decision-making and produce results with an error rate of 0, the responsibility gap will remain. As I hope to have shown, however, achieving an error rate of 0 is simply an unrealistic goal.

5.4 Conclusion

In this chapter, I have attempted to demonstrate that not only will substitutive AI be limited in its capacity to foster a strong patient-clinician relationship, but that it will raise systemic issues outside of the clinical encounter. In particular, it will be susceptible to biased decisions and errors that will prove far more difficult to deal with and correct and it will cause a responsibility gap that will erode trust in the healthcare and the legal system. My ultimate goal over Chapters 4 and 5 has been to show that, given all these concerns that substitutive AI would present in medicine, it would be an unsuitable substitute for clinicians. Our goal, then, ought to be to refrain from or at least be cautious about how much authority, duties, responsibility, and accountability we attempt to transfer away from clinicians and on to AI. Ideally, to avoid the issues discussed in this and the previous chapter, the goal with regard to substitutive AI in healthcare should be to implement them as tools rather than as substitutes. Of course, this (what has been termed the "assistive approach") also faces some challenges, which I discuss in the next chapter.

Chapter 6: The Assistive Approach

6.1 Introduction

I have thus far argued against the neo-luddite approach that calls for banning the future deployment of AI in healthcare, and possibly even de-technologizing the current healthcare system. I have also argued against the substitutive approach that promotes the possibility of AI taking over the roles of clinicians and replacing them entirely. This leaves us with a third option: the assistive approach, which calls for the implementation of AI in healthcare, though not with the goal of substituting for clinicians but rather with the goal of assisting them.

Worth noting is the fact that this dissertation is intended to provide a broad theoretical framework from which we can deduce specific rules and policies. There is certainly room for exceptions within approaches. If it turns out, for instance, that a certain role within healthcare can be substituted by AI without the concerns discussed chapters 4 or 5 arising, then an exception may be made in that case. Similarly, if in certain contexts (i.e., when dealing with patients with a permanent loss of consciousness) in which the duties that a substitutive AI would normally fall short of no longer apply, exceptions can be made. What I have attempted to do in this dissertation is touch on important factors that make AI an unsuitable substitute for a clinician. In those instances, then, the appropriate approach is to implement AI in an assistive sense, where it can be utilized by clinicians in an effort to improve accuracy and efficiency and further advance knowledge in the medical domain.

The assistive approach too, however, is not without its challenges. The aim of this chapter is to address three interconnected challenges in particular that are unique to the assistive approach (that is, they are only present when AI is implemented *alongside*, and utilized by, human clinicians). The three challenges concern 1. Whether opaque AI may be used for diagnosis or treatment recommendations; 2. How to handle disagreements or discrepancies between an AI's recommendation and a clinician's judgment; and 3. When AI is used by a human clinician in decision-making and the decision yields disastrous results, who is morally responsible and accountable for this?

AI's potential to yield superior recommendations than human clinicians, in certain contexts, is rarely contested. Given its ability to scan large amounts of data and detect what may

be undetectable to the human eye or unknown due to limits in medical knowledge, AI's capacity to eventually outperform human clinicians with regard to diagnostics is a commonly agreed upon view in the literature (Budd, 2019; Banja et al., 2022). This, however, raises the question of what ought to be done in cases where the clinician's own expert opinion conflicts with the AI's recommendation for diagnosis or treatment. Moreover, it raises the question of how responsibility is to be allocated in cases where the clinician accepts the AI's recommendation but this recommendation yields harmful results, and, vice versa – how ought responsibility to be allocated in instances where the clinician trusts their own judgments over that of the AI but their own recommendation results in harm and proves inferior to the AI's?

In order to answer these questions, however, it will be helpful to first determine the epistemic role of assistive AI as a decision support system in the clinic. This is because determining the epistemic role of assistive AI will allow us to better understand the best way of utilizing it, the best way of handling disagreements between AI and clinician, and how best to assign responsibility and accountability in instances of disagreement. I will begin by discussing the issue of disagreements before proceeding to discuss the epistemic role of assistive AI in relation to the clinician (that is, whether assistive AI is an epistemic superior, an epistemic peer, an epistemic inferior, or of an epistemic category all on its own). As will shortly become clear, this discussion is intertwined with the issues of opacity and responsibility that I have set out to answer in this chapter.

I will begin the next section by discussing the reasons for AI-clinician disagreement and why both courses of action – namely, accepting an AI's recommendation despite disagreeing with it and ignoring an AI's recommendation when disagreeing with it – are problematic. In Section 6.3, I discuss the epistemic role of assistive AI as decision support systems. In Section 6.4, I propose a solution to the issue of disagreement and discuss what this solution means in terms of using opaque decision support systems and allocating responsibility and accountability in cases of AI-clinician disagreement. In Section 6.5, I provide some concluding remarks.

6.2 The Challenges of Human-AI Disagreements

6.2.1 Reasons for Disagreement

There are several reasons for why disagreement may occur between a clinician's recommendation and an AI's. One reason could be due to the AI knowing better than the clinician, either due to human error on the part of the clinician or due to limits in the state of medical science (which the AI is able to detect despite such knowledge lacking among human medical experts). Another reason for disagreement could be due to AI malfunction (through no fault of the operator) or due to operator-error (Kempt at al., 2023, p. 1409). Finally, disagreements between clinicians and AI could occur due to both recommendations being equally adequate and probable – i.e., when there is not enough knowledge in the field regarding a new disease, so each side recommends what it believes to be the next closest thing to it, yet each side is relying on different factors. An example of such disagreement might look like this:

Side 1: *"This new disease, Disease X, most closely resembles Disease Y because of symptoms a, b, and c (which are found in Disease Y but not Disease Z). So we should recommend for a patient suffering from Disease X a treatment option most closely resembling that used for Disease Y."*

Side 2: *"This new disease, Disease X, most closely resembles Disease Z because of symptoms d, e, and f (which are found in Disease Z but not Disease Y). So we should recommend for a patient suffering from Disease X a treatment option most closely resembling that used for Disease Z."*[52]

6.2.2 The Problems with Accepting and Ignoring AI Recommendations

As Bjerring and Busche point out, part of the challenge of dealing with clinician-AI disagreement is that with AI that outperforms human clinicians, there seems to be an epistemic obligation for clinicians to defer to it, treating it as an expert (Bjerring and Busche, 2020, pp. 351, 354). That is, just as a practitioner may turn to an expert in MRI interpretation in instances of uncertainty, given that the expert is "more knowledgeable about the medical domain, more reliable in her diagnostic work, and better at making accurate predictions based on the image" (ibid., p. 354), the same may be said of AI. This epistemic obligation puts clinicians in a tough

[52] This is intended to demonstrate only a source of disagreement, not the AI's explanatory power.

position where they must balance this epistemic duty against other duties and considerations, especially considering that the reason for disagreement could be due to AI malfunction or error.

6.2.2.1 AI Recommendation Not Necessarily Superior

Disagreement and the possibility of error on the part of the AI is possible even if the AI generally outperforms clinicians. In Choudhury and Asan's (2022) study on healthcare practitioners' perceptions of AI, one clinician worded the issue perfectly:

> "Very scary at the end of the day in both directions. If we follow AI guidance and something goes wrong, we are done. If we do not follow AI guidance and something goes wrong, we are still [held accountable]" (p. 9).

The possibility that the AI's recommendation is the result of error or malfunction is, of course, one reason against simply accepting or relying on its recommendation without understanding the justification or evidence or theory behind it. But there are other reasons. One is that the recommendation could be different, though equally as probable a diagnosis or course of action as the clinician's, or perhaps even inferior given the particular patient's values or non-medical aspect of the patient's life that the AI lacks access to. Consider the following analogy to illustrate: a chess engine is programmed to find the best moves and avoid possible "blunders" that may cause a player to lose a valuable piece or risk giving their opponent an opportunity to check-mate. Upon reviewing your successful game with a friend that you have played chess with countless times in the past (and whose chess techniques and strategies you know very well), the chess engine determines that many of the moves you made were "mistakes" or "blunders," but these "mistakes" or "blunders" are precisely what allowed you to check-mate your friend within a few moves. Perhaps you know that your friend's strategies all revolve around their queen and you are aware that taking away their queen will put them at a huge disadvantage, while you usually create strategies that do not involve your queen. So the move to sacrifice your queen in exchange for theirs, while a "blunder," to the chess engine, is actually an excellent move. Simply put, the chess engine's assessment was missing crucial information that you had, and while your

moves may have been "blunders" or "mistakes" when playing against the average player, they were excellent moves in this particular context and when playing against *this* particular player.

The same can be true in medical contexts. There may be facts about patients that the AI is not trained on (I will discuss some such cases shortly) and these facts may be the very reason behind the disagreement. Without an explanation behind the AI's decisions, to assess and compare the recommendation of the AI with that of the doctor, taking into account all the factors, medical and non-medical, it makes for a mistake to simply accept the AI recommendation over the clinician's own.

6.2.2.2 Conflict with Patient-Centred Care

Once again, an advantage of the clinician's recommendation is that it is accompanied by justification[53], which leads to the next reason against simply relying on AI's recommendations in cases of disagreement: black-box AI, as Bjerring and Busche (2020) argue, conflicts with patient-centred care. They explain that relying on black-box AI systems in medicine steers us away from an important value in healthcare: treating patients "as equal partners in medical decision-making." Patient-centred care involves seeing patients not "as passive recipients of orders from the paternalistic and all-knowing doctor but rather understood as active partners who—in an informed manner—can and should exercise their own autonomy to help choose their own ends of action." (ibid., p. 359). However, the opacity of black-box medicine conflicts with this idea of patient-centred care by not giving access to justification or reasoning behind its recommendation to the practitioners to share with their patients. This, in turn, results in failing to provide patients with reasons or justifications that would allow them to fully exercise their autonomy. To illustrate, Bjerring and Busche (2020) present the example of Mary, a breast cancer patient:

> Suppose Mary has a strong preference for, say, treatment Z. With a view toward making an autonomous and rational decision that furthers her healthcare and life goals, understanding why treatment Z receives such a low probability compared to treatment X [removing both of her breasts] will matter greatly to her. Yet, since

[53] See Chapter 4, Section 5.1.1 for a discussion of the difference in justification between ordinary medicine and black-box medicine.

> the practitioner does not have access to this information, he cannot share it with Mary. So despite being aware of the potential side effects of the various treatment options, Mary is arguably still lacking a crucial element of understanding that prevents her from making a decision in an informed and rational manner (p. 363).

Sparrow and Hatherley, similarly, note that "patients want to make their own choices about their health, or at least share the decision with their doctor…the opacity of AI…deprives patients of the opportunity to receive answers about key questions related to their treatment" (Sparrow and Hatherley, 2019, p.93).

6.2.2.3 Sacrificing the Expansion of Medical Knowledge

Another reason for refraining from simply relying on the AI's recommendation is that, by doing so, we are sacrificing the expansion of medical knowledge. It is typically through explanations, which can be further tested and verified, that knowledge in medicine expands and progress is made. Yielding to the recommendations of AI and being content with a lack of explanation behind medical decisions, then, will halt progress in medical knowledge unless these recommendations are further inspected and reassessed, taking into account how the AI recommendation fares against alternative clinician recommendations, what types of patients each recommendation is best suited for, how they may align or conflict with ethical considerations, etc.

6.2.2.4 Allocation of Responsibility

Moreover, given AI's lack of capacity for moral responsibility and accountability,[54] relying on AI's recommendation without any further justification would either mean accepting a responsibility gap or placing responsibility on the clinician, who, as I mentioned earlier, would essentially be put between a rock and a hard place with regard to their epistemic duty and their duty to abide by other competing values in medicine.

[54] This is discussed in more detail in Chapter 5.

On the other hand, simply rejecting an AI diagnosis in cases of disagreement where the AI's recommendation lacks an explanation, is also problematic. Generally, evidence of disagreement, whether it comes from another clinician or AI, ought to give the clinician a reason to pause and reassess their own results and those of the disagreer. Failing to do so would be failing to meet the burden of proof that they have now been subjected to and thereby failing to abide by their epistemic duties. This means that, in instances where their recommendation results in a negative outcome, they will be held to a higher degree of responsibility and accountability, given the missing justification that they need to explain their preference for their own recommendation over that of the AI.

6.3 The Epistemic Role of Assistive AI as Decision Support Systems

Now that the dilemma arising from AI-clinician disagreements has been presented, I would like to turn to discussing the epistemic role of AI decision support systems. Though there has been some discussion of AI as an epistemic technology (Alvarado, 2023), there has not yet been much discussion in the literature regarding the *epistemic role* of AI decision support systems, which, in the medical context, are intended to assist the human clinician with decisions concerning diagnosis and treatment recommendations. Determining the epistemic role of such AI will yield important implications for (1) its use by clinicians, (2) the procedures for handling disagreements between clinicians and decision support systems, and (3) the proper allocation of responsibility and accountability in cases of clinician/AI disagreement. To see this, suppose that AI decision support systems were to be deemed epistemic superiors, this may give clinicians a reason, or perhaps even a duty, to yield to the AI's decisions in cases of disagreement. Doing so would, arguably, be in alignment with their duties, given the hypothesis that the AI system is more accurate. Then any harms resulting from the AI's decisions would not be attributed to the clinician, whereas deviating from the AI's recommendation would make the clinician responsible and accountable for any resulting harms. Alternatively, imagine an AI system that is demonstrably an epistemic inferior. Then a clinician yielding to its decisions in cases of disagreement will generally be an inappropriate course of action, and consequently, the AI's decisions resulting in harm will place much (if not all) of the responsibility and accountability on to the clinician.

But things are more complicated if an AI turned out to be neither an epistemic superior nor an epistemic inferior, but an epistemic peer. These complications follow from the fact that it is already hard to say what to do about cases of peer disagreement among humans. We can see this with a quick look at the epistemology literature on peer disagreement. One popular account of peer disagreement is *the equal weights view* (Christensen, 2007; Elga, 2007), according to which "one should give the same weight to one's own assessments as one gives to the assessments of those one counts as one's epistemic peers" (Elga, 2007, p. 484). This would mean that in cases of disagreement where the AI is considered an epistemic peer, the clinician would have a duty to give the AI's recommendations equal weight to their own. Thus, automatically yielding to the AI or automatically ignoring it without reason or justification would, in this case, not be appropriate. Though it is perhaps also worth noting that the equal weights view is not without its critics (Kelly, 2011).

With this brief explanation of how determining the epistemic role of assistive AI decision support systems will be helpful in answering the questions I have set out to answer in this chapter, I turn now to exploring the possible epistemic categories in which we may classify AI decision support systems. After rejecting the categorization of AI decision support systems as epistemic superiors, epistemic peers, and epistemic inferiors, I go on to propose a fourth category specific to AI: an epistemic *nudge*.

6.3.1 Is Assistive AI an Epistemic Superior?

Taking someone or something to be an epistemic superior means, as put by Elga, taking them "as being more than 50% likely to be right, on the supposition that the two of you end up disagreeing" (Elga, 2007, p. 489). As noted by Thomas Kelly, one may be deemed an epistemic superior due to their "superior familiarity with or exposure to evidence" (Kelly, 2005, p. 174). In other words, all else being equal, if we deem someone or something to have access to all the relevant evidence we have plus more, we deem them to be an epistemic superior. One may also, according to Kelly, be deemed an epistemic superior "with respect to general epistemic virtues such as intelligence, thoughtfulness, freedom from bias, and so on" (ibid.).

Given that GPT-4 is already outperforming humans on medical exams (Guerra et al., 2023), we may perhaps soon have evidence of AI's superior capacity to diagnose medical

conditions when compared to human clinicians. Or, if this is too strong of a claim, then given AI's access to and utilization of large amounts of data that the typical clinician would nowhere as easily be able to sort through, we may think it appropriate to categorize it as an epistemic superior. After all, it has access to more information than we do. Thus, in this case, similarly to taking a meteorologist to be an epistemic superior on weather-related matters, clinicians would take AI decision support systems to be epistemic superiors on matters relating to diagnosis and treatment recommendations. Thus, in cases of disagreement between the (epistemically superior) AI and the clinician, the most appropriate response would be for the clinician to defer to the AI.

However, there are reasons to be opposed to deeming AI to be an epistemic superior – the first is a practical reason and the second is an epistemic one. The practical reason simply ties back to some of the earlier problems[55] concerning substitutive AI, because, as I have just mentioned, deeming AI to be an epistemic superior often means that the appropriate course of action to take in the event of a clinician-AI disagreement is to defer to the AI. However, deferring to the AI in every situation would render the clinician's role obsolete, at least regarding decision-making about diagnoses and treatment options. Transferring the decision-making power to the AI and away from the clinician means that the assistive AI in this context is no longer assistive, but substitutive.[56] As I have argued at length, however, substitutive AI (even outside the clinical encounter) is still problematic due to the very concerns that steered us away from it in Chapter 5 (the most evident ones being the resulting presence of a responsibility gap and biased decision-making). Thus, to address the practical problem, it may be fair to propose that even if AI decision support systems are epistemic superiors, they ought not to be treated as such.

But this practical reason is not the sole reason for rejecting AI as epistemic superiors – there is an epistemic reason, as well. To begin, it is important to point out that where epistemic virtues are concerned, having access to large amounts of data, or having knowledge in general, may not be sufficient for epistemic superiority. Put differently, knowledge on its own, as some scholars (Pritchard, 2009; Gardiner, 2012) have argued, is not always sufficient in terms of epistemic achievement or value. Rather, it is *understanding,* rather than knowledge alone, that is valuable. If AI lacks understanding, then even if it has superior knowledge in terms of access to evidence from large amounts of data, it would nonetheless lack a central epistemic virtue,

[55] See Chapter 5.
[56] See Chapter 3.

making it inappropriate to categorize as an epistemic superior in the context of medicine. As I will further go on to show, given that a goal of medicine concerns promotion of patient values, *understanding*, rather than simply knowing, is an important epistemic virtue for a medical decision-maker to have.

Though there is no single agreed-upon definition of "understanding" in the epistemology literature, I will rely here primarily on what is often referred to as *explanatory understanding*[57]. Explanatory understanding, put simply, is "the kind of understanding that one has when one understands *why* something is the case" (Hannon, 2021, p. 282). As Gardiner notes, "when someone understands an explanation she knows facts about the causes and antecedent conditions, and can often reason counterfactually (if Sally understands why the Democrats lost the election then she can reason about outcomes where antecedent conditions were different, say, if they had more money)" (Gardiner, 2012, p. 164). Perhaps the best way to demonstrate the difference between knowledge and understanding and, in particular, how one may have knowledge without understanding, is through the following example borrowed from Pritchard (2009): a child may know that his house burned down and he might know that it burned down due to faulty wiring, while nonetheless lacking understanding of how or why faulty wiring caused his house to burn down. It is examples such as these that have led opponents of veritism to reject knowledge as the sole epistemic good, because as Pritchard and others have aimed to demonstrate, understanding requires more cognitive ability and constitutes a greater cognitive achievement than knowledge.

There are at least two reasons to think that AI decision support systems lack understanding: 1) due to their opacity and, thus, lack of explainability, and 2) due to the lack of integration among their data points. Recall that the focus of this chapter is on figuring out how to best deal with situations in which the AI decision support system's recommendation conflicts with that of the human clinician. Cases in which the decision support system provided reason or explanation for its decisions would be much more straight forward and easy to handle – the clinician would simply evaluate the explanation and proceed accordingly, and if the explanation demonstrates the presence of further evidence that the clinician initially lacked, then the clinician can re-evaluate their initial recommendation with the new evidence taken into account. If, on the other hand, the decision support system's explanation turns out to be based on some error or missing contextual information, the clinician can discount its recommendation. However, as I

[57] Or what Gardiner calls "understanding causal explanations." (Gardiner, 2012, p. 164)

have explained, the issue is much more complicated when the decision support system is opaque and explanations or reasons are missing (as is true of many advanced machine learning systems). Explainability, however, is an important way to test for understanding. In order to demonstrate to you that I understand how faulty wiring burned down my house, I would have to demonstrate to you, through explanation, my grasp of faulty wiring systems and how they cause fires and why in this particular context I believe that faulty wiring (rather than some other reason) is the cause of the fire to my house. Similarly, for a clinician to demonstrate that they understand why the best treatment option for patient A is radiation, they would have to demonstrate, through explanation, their grasp of how radiation works, their grasp of how radiation would help this particular patient, their grasp of possible alternative treatment options and how they work and how they would help this particular patient, etc. The AI decision support system that concerns us here, however, is not capable of providing such explanations.

There is a second reason to doubt that AI decision support systems have understanding: the lack of integration among their data points. As Gardiner (2012) explains, understanding requires having a connected web of beliefs. To illustrate, she presents the case of Markus, someone who believes propositions randomly and irresponsibly. Markus does not care about justification, nor does he care about his beliefs being true. Luckily for Markus, an omnipotent being ensures that all of his beliefs are true. Still, Gardiner posits, Markus does not have understanding because his beliefs lack integration – "as he never reflects on his beliefs, none of his beliefs serve as reasons in favour of any of his other beliefs and he never sees connections among his beliefs." (Gardiner, 2012, p. 170). An AI decision support system, just like Markus, would not have the capacity of belief-integration (at least not to the same degree as a human clinician). The decision support system will be trained on data and will attempt to produce results that resemble it. It might be able to take Dan's chronic pain and infer something on the basis of similar pain among patients similar to Dan and, as a result, recommend prescription opioids despite Dan's recently-developed fear of opioids after witnessing a close friend's battle with addiction. Even if the decision support system can be trained to take into account factors like concerns about opioid dependence, it would not be able to truly understand how the connection between prescribing opioids and the patient's fear of opioid dependence works or why it matters. This is not an isolated case - there are countless examples where a human, by their ability to

make and note connections among their beliefs and understanding of the world, would be able to explain that the decision support system will not be able to.

Explanatory understanding and belief-integration are both important, not just as epistemic virtues, but practically, as well, especially in the context of medical decision-making. There are at least two reasons for this: The first concerns the promotion of patient values and the second concerns error detection.

As I argued in Chapter 1, promoting patient values is an important goal of medicine. With that in mind, it is easy to see why lacking understanding is a problem. Consider the complexities of interactions between negative side-effects and a patient's values. Dan's concern about opioid dependence is one example; a purely medical diagnosis and treatment recommendation may miss important values and motivations that Dan holds. Another example is the high estrogen levels through estrogen injections elected by trans women as a means of gender affirmation being mistaken by the AI decision support system as a symptom of a health issue, thereby causing it to recommend treatment to reduce estrogen levels, despite this not aligning with the patient's values. Thus, lack of explanatory understanding presents a major issue in medical decision-making.

A lack of explanatory understanding is also troubling in that understanding is what allows us to be able to detect when something (such as a new piece of information) is off, particularly if it does not cohere with our other beliefs. Gardiner (2012) presents the following analogy, which I believe does a great job of illustrating, for our purposes, the epistemic benefit that human clinicians have over AI decision support systems:

> Consider the difference between, on the one hand, reading a book where we virtuously accept each individual proposition in isolation and, on the other hand, actively grasping how the arguments in the book hang together, believing the propositions in part because we see how they follow from previous facts. With the former mode of reading we may believe of each proposition that it seems plausible, and that we trust the author, and so affirm the proposition with justification. But with the latter mode of reading we actively grasp the relationships among these facts and how the conclusions must follow (pp. 170-171).

While a human clinician would be able to detect, for instance, that there is a problem in the recommendation of anti-estrogen medication to an otherwise healthy trans woman given their understanding of the different aspects of the situation and how they cohere, the AI decision support system would have to rely on patterns from separate data points provided through the training data. Given enough time and data, the training set might eventually account for cases like these. But for rare or novel situations, where the training data is limited or where a patient presents a unique situation that does not match any pattern in the training data from which the AI's decisions are inferred, the AI will not only be susceptible to producing inappropriate recommendations, but it will also be unable to detect that it has made an inappropriate recommendation that fails to cohere with other important aspects, such as the patient's values.

Thus, given these epistemic shortcomings of the AI decision support systems, it seems inappropriate to deem them to be epistemic superiors in the medical context, despite their superior access to data. I have already discussed the suggestion of post-hoc XAI methods as a way of dealing with the issue of opacity multiple times in previous chapters. For the sake of the present discussion, it will be sufficient to point out that, in addition to their limited reliability, XAI methods are merely attempts at rationalizing the outputs of the machine. They do not reflect the explanatory power of the machine itself (Babic et al., 2021, p. 285). The epistemic worries regarding AI's lack of explanatory understanding and belief-integration remain even with the utilization of XAI methods.

6.3.2 Is Assistive AI an Epistemic Peer?

I hope to have established in the preceding section why AI decision support systems ought not to be deemed an epistemic superior. I move now to considering whether they ought to be deemed epistemic peers. If it can be deemed appropriate to categorize AI decision support systems as epistemic peers, then we would have to settle complicated issues arising from the literature on peer disagreement to determine the appropriate course of action in cases of disagreement between the AI and a human clinician. However, this will not be necessary for our purposes, because as I will go on to show, the assistive AI in question here is not an epistemic peer.

An epistemic peer is a term used to refer to an agent that one deems to be one's epistemic equal – that is, an agent that one deems to be as good at evaluating claims as oneself. I will rely here on Thomas Kelly's (2005) definition of epistemic peers as follows:

> …Two individuals are epistemic peers with respect to some question if and only if they satisfy the following two conditions:
>
> (i) they are equals with respect to their familiarity with the evidence and arguments which bear on that question, and
>
> (ii) they are equals with respect to general epistemic virtues such as intelligence, thoughtfulness, and freedom from bias (pp. 174-175).

With this definition of an epistemic peer in mind, there are three reasons to doubt that it is fitting for an AI decision support system. The first is that we do not know that the clinician and the AI are equals with respect to their familiarity with the evidence. In fact, it is very unlikely that the clinician and AI are equal with respect to their familiarity with the evidence given the vast amounts of data that the AI has access to which would simply not be feasible for a clinician to have or keep track of. As I mentioned in the previous section, where data is the only consideration, the AI is very likely superior to the human clinician.

The second reason to doubt the categorization of this AI as an epistemic peer concerns its lack of explainability. The opacity of the decision support system constitutes a lack of an epistemic virtue here once again. An epistemic peer is not merely one who has access to and familiarity with the relevant evidence to the same degree that we do, but it is also one who reasons about the evidence and formulates arguments about them in a similar manner that we do. Note, this ability to reason and formulate arguments on the basis of the evidence can be read as part of the second condition of Kelly's definition of what constitutes an epistemic peer. Two clinicians may be deemed epistemic peers in that they are both capable of using the evidence they have to formulate reasons and arguments about which diagnosis or treatment recommendation may be appropriate for a patient. The AI decision support system, as I have established in the previous section, is not capable of reasoning in a similar fashion as a human clinician considering its lack of explanatory understanding altogether.

The third reason to doubt that an AI decision support system is an epistemic peer concerns its lack of another epistemic virtue: reflective ignorance. As Axel Gelfert suggests, two epistemic peers ought to be equally aware of the limits of their own knowledge. Being aware of one's ignorance with regard to the truth-value of a proposition, according to Gelfert, is beneficial, as it allows for further inquiry into the subject matter (Gelfert, 2011, p. 514). Once again, however, AI's lack of reflective ignorance[58] and the consequent hinderance of further inquiry which may lead to more valuable knowledge, puts it in an inferior epistemic position to the clinician, and thus, gives us a further reason to demote it from the position of an epistemic peer.

6.3.3 Assistive AI as Epistemic Nudge

The following question now naturally arises: If the assistive AI decision support system in question is neither an epistemic superior nor epistemic peer, then what epistemic category does it fall into? The initial response that one may be tempted to turn to is that it is an epistemic inferior. This, however, is quite difficult to accept for two reasons: the first is that, as I have previously mentioned, the AI has access to much more data than the clinician does, giving it an upper hand where familiarity with evidence is concerned. So the story cannot be one of simple inferiority. The second reason is that considering the AI to be an epistemic inferior would deem it practically valueless as a decision support system given that in every case of disagreement, the appropriate course of action would be for the clinician to ignore its recommendations and proceed with their own. The whole purpose of these decision support systems, however, is to provide the clinicians with support.

I propose, instead, that we categorize AI decision support systems as neither epistemic superiors, nor peers, nor inferior. Rather, I propose that they be in an altogether new category all on their own: as *epistemic nudges*.[59] An epistemic nudge can be defined as an agent or entity that

[58] ChatGPT, for instance, has often been the source of criticism with regard to "artificial hallucinations" (false information confidently presented as fact). See, for example, Alkaissi and McFarlane (2023)

[59] Though the terms "nudge", "epistemic nudge," and "doxastic nudge" have been used in the literature to refer to the practice of influencing or steering another agent's beliefs or decision-making capacities (see e.g., Thaler and Sunstein, 2009; Meehan, 2020; Grundmann, 2023), my use of the term diverges from these definitions and encompasses an entirely different concept.

is not an epistemic peer, but rather an epistemic superior in some important ways and an epistemic inferior in other important ways. In such cases, the agent or entity may be utilized as a "nudge" in that we may pay it attention without yielding to it altogether, similarly to how clinicians may utilize urine samples for drug tests, while keeping in mind that they are prone to false positive and false negative results (Algren and Christian, 2015).

6.4 Proposed Solutions

Given that the opaque assistive AI decision support system should be thought of as neither an epistemic superior, nor an epistemic inferior, nor an epistemic peer, but an *epistemic nudge*, how does this translate to application? That is, how should this inform our practices regarding the use of opaque decision support systems, handling disagreements between clinician and AI, and allocating responsibility and accountability?

6.4.1 Ban the Use of Black-Box Decision Support Systems?

Given their epistemic limitations, one possible solution is to advocate for the ban of any black-box decision support system in healthcare. While assistive white-box (transparent) AI will be helpful in decision-making, the opacity of a system, it may be argued, merely presents trouble and stress and an added burden for clinicians.

This would be true if black-box decision support systems did truly lack any advantages, or even if the advantages proved to be minimal when compared to the disadvantages. But, of course, AI that has been determined to have a high reliability rate and the likelihood to outperform clinicians is advantageous. Its ability to sort through large amounts of data and medical research and detect what may be undetectable to the human clinician makes it far from useless or unnecessary. One particular advantage is as a confirmation tool. That is, though cases of disagreement present challenges to the clinicians making use of the decision support systems, instances of agreement are very beneficial for two reasons: 1. they confirm, thereby increasing clinicians' confidence in their own decisions in instances of uncertainty, and 2. They add further justification to the clinicians' decisions, lessening the clinicians' degree of accountability for any harm that may result from the decision. We can compare this to a confirmation from another

expert. This means that, as part of a clinician's explanation for why she chose a particular treatment plan or diagnosis, she can reference not only the medical research or theories that played a role in her decision, but also that this same decision was separately recommended by reliable AI.

One may be skeptical about this benefit. Jongsma and Sand (2022), for instance, argue that there is a symmetry between agreement and disagreement, in that both require justification:

> Each individual case of agreement requires a separate justification: a reason for reaching the same conclusion. Without understanding these reasons, we cannot know whether coming to the same conclusion should indeed be called agreement, or must rather be seen as an effect of the employment of different heuristics…as conclusions emerging from different reasons or even just mere coincidence (p. 230).

If Jongsma and Sand are correct, then a black-box AI that lacks a justification is not helpful even in cases of agreement, because it fails to provide a true confirmation of the clinician's diagnosis or course of action. The alignment of the clinician and AI's recommendations could be only coincidental.

Jongsma and Sand's argument does have some merit, especially in that a confirmation of the clinician's recommendation by the decision support system does not guarantee that the diagnosis is correct (in fact, it could be that they are both mistaken – e.g., the clinician is mistaken and the AI is malfunctioning, yet coincidentally, they have both settled on the same recommendation). Nevertheless, two points are worth noting:

1. Although the AI's confirmation of the clinician's recommendation may not necessarily count as agreement, it is a mistake to think that it is not useful without the accompanying justification. Even if the agreement is accidental, in terms of consequences, it directs us towards the same (best) option given the available options at the time. Consider this analogy: My partner and I are taste-testing cakes for our wedding and we both prefer the red velvet cake. We do not discuss our reasons for preferring the red velvet cake, we both simply happen to vote on it as our favourite. Unbeknownst to the both of us, we choose the cake for different reasons (e.g., I think the red velvet cake has the perfect level of sweetness and my partner thinks it is a bit too sweet but has the perfect smooth and creamy texture). The decision to order the red velvet cake remains the

best option in that scenario given the agreement on the final outcome, despite our potentially not agreeing on the reasons behind the final outcome. It would not make much sense, for instance, for both of us to prefer the red velvet cake yet choose a completely different cake when we come to realize that our preference of the red velvet cake depended on different reasons.

2. When it comes to burdens of proof, agreement and disagreement are not quite as symmetrical as Jongsma and Sand suggest. Generally, there is a higher requirement for justification in cases of disagreement than in cases of agreement. I need not go as far as to suggest that justifications for beliefs are *only* required when one's beliefs are challenged or that no reasoning or justification is ever required in the absence of any such challenges (though a similar view to this has been put forth by Hanno Sauer (2017) regarding moral intuitions). It is sufficient here to make the much weaker claim that justification is required to a much lesser degree in cases of agreement than in cases of disagreement. If I grew up believing that my parents were, in fact, my birth parents, and my uncle casually makes statements to confirm this, it is very unlikely that I will ask him for evidence. If, however, my uncle makes a statement to the contrary, it is natural to seek evidence or reasons for why his statement conflicts with my belief. It is, of course, possible in the first scenario that my uncle and I are both mistaken or that I am mistaken and he is purposely making false statements, but generally, when it comes to requiring evidence or justification or proof, we tend to require them much more in cases of disagreement than in cases of agreement. This is also common in our current practices of seeking a second opinion: If I believe that the answer to the question, "who was the lead actor in the 2005 movie, *The Amityville Horror*?" is "Ryan Reynolds" and I ask ChatGPT and the response I receive is "Ryan Reynolds," (a confirmation of my initial belief) I am generally satisfied and take this as a reason to further increase my confidence in my initial belief. If, however, I believe that it was Ryan Gosling who starred in the 2005 movie, *The Amityville Horror,* and the answer I receive from ChatGPT is "Ryan Reynolds," the likely next step is for me to seek some justification for ChatGPT's response, which might include conducting some further searches, such as looking at snap shots of images from the movie on Google or searching for the movie trailer, or asking someone else.

If I am right to think that this negates Jongsma and Sand's claim regarding the symmetry of agreement and disagreement, then, assuming the decision support system in use has a high reliability rate, it would be as much of a mistake to disregard its confirmation of, or agreement

with, the clinician's recommendations as it would be to disregard its disagreement or conflict with the clinician's recommendations. Thus, a decision support system, even when lacking an explanation for its outputs, is nonetheless beneficial as a confirmation source – for both clinicians and patients, who can incorporate it into their own decision-making process.

6.4.2 Rule of Disagreement

Given that the solution is not to refrain from incorporating black-box AI in healthcare altogether, but also given that its incorporation in healthcare raises issues regardless of whether its recommendations are blindly accepted or rejected in instances of disagreement, how *ought* disagreements be handled?

One suggestion, proposed by Kempt and Nagel, for handling disagreements between clinician and an AI decision support system, is what they call the *rule of disagreement*, which "permits AI as second opinion providers unless they deviate too far from the initial opinions, in which case a second human opinion ought to be sought. If the [AI decision support system] confirms the initial opinion, however, no further steps need to be taken" (Kempt and Nagel, 2022, p. 222). The idea behind the rule of disagreement is that when a clinician's recommendation is confirmed or agreed upon by another clinician or expert, there is usually no reason for them to think that their recommendation is incorrect. Similarly, if a decision support system confirms their original recommendation, they have no reason to doubt it (ibid., p. 224). In instances of disagreement, however, they cannot simply ignore the conflicting recommendation, just as they cannot ignore it if proposed by a fellow clinician. In instances of disagreement, there is a burden of proof on the clinician to either justify their initial recommendation as superior or to justify adopting the disagreeing clinician's recommendation (ibid., p. 224). Similarly, when an AI's recommendation is in disagreement with the clinician's initial recommendation, there is a burden of proof on the clinician to either justify their own recommendation as superior or to justify adopting the AI's recommendation. The problem is, as mentioned earlier, that while human clinicians can engage in a back-and-forth conversation, providing reasons and justifications, and determining the source of their disagreement before resolving it, this kind of back-and-forth exchange of reasons is not possible with opaque decision-support systems (ibis., p. 225). Thus, it is for this reason that Kempt and Nagel propose moving forward in cases of

clinician-AI disagreement by acquiring a second opinion from another human clinician (ibid., p. 227). This allows the decision-making clinician to fulfill their epistemic duty by not simply ignoring nor simply yielding to the differing recommendation of the AI.

Kempt and Nagel's proposed rule of disagreement offers a great approach to using decision support systems in healthcare, but it fails to account for at least two challenges, the most important one being the following: although a second opinion from a fellow human clinician is helpful in instances of disagreement between the initial clinician's recommendation and an AI's – in these cases, the AI is deemed to be somewhat useless. In other words, it does not account for cases of disagreement where the AI makes better decisions than both human clinicians – for instance, in cases where the AI's recommendation is in disagreement with the clinicians' because it has access to what most, if not all, human clinicians do not (what has not yet been detected by medical experts or sufficiently studied). It is precisely in these cases that a decision support system is most useful. It thus seems mistaken to simply turn to a fellow human clinician in assessing both recommendations (one of which is missing a justification) and form a decision on the basis of what is available to only the two human clinicians.

The second thing that the rule of disagreement is missing is room for exceptions. The rule of disagreement appears to treat all instance of agreement and disagreement the same, whether the decision will be that of life or death or if it is a decision where, at worst, the patient will suffer very minor reversible side effects from the incorrect recommendation. But just as I would not handle a case of disagreement about something trivial like the lead actor in *The Amityville Horror* in the same way that I would handle a case of disagreement over the required documents needed for my urgent visa application, the particular context and the importance of getting the right answer and the amount of risk that the wrong answer poses, ought to be taken into consideration when determining how disagreements between a clinician and a decision support systems ought to be handled.

6.4.3 Rule of Nudge

These issues with Kempt and Nagel's account primarily arise due to the specificity of the rule – the narrow and straight-forward nature of the guidelines do not allow for exceptions – whether these exceptions concern the gravity of the situation in question or the cases in which

the AI's knowledge can be useful despite being in conflict with two (or perhaps even more, or perhaps even *all*) clinicians. A more useful rule for handling disagreement between AI and a clinician would be one that sets guidelines but allows for some wiggle room within those guidelines.

I, thus, propose the following rule for handling cases of disagreement between a clinician and an AI decision support system, which I am calling *The Rule of Nudge:* In cases of agreement between a clinician's initial recommendation and that of the decision support system, the confirmation may count as a second opinion, fulfilling the clinician's epistemic duty, without requiring a further burden of proof or the opinion of other human clinicians. In cases of disagreement, the AI's recommendation must be neither automatically yielded to nor ignored, but used as a "nudge" by the clinician. What this means is that while the best course of action to take will vary depending on the specific context, the key is to take seriously enough the possibility that the nudge could be a *potential* epistemic superior in a given case without necessarily accepting that it is one. Below are a few examples of what handling disagreements with an epistemic nudge may look like (though, again, these are not strict guidelines but mere examples of what it might mean to treat an AI decision support system as an epistemic nudge).

One example of treating an AI decision support system as an epistemic nudge could mean that despite seeking the second opinion of a human clinician who turns out to be in agreement with the initial clinician, the AI's recommendation is still reported to the scientific and medical communities for further research. Another example, especially in high-risk scenarios, could be a combination of: 1) seeking second opinions from *multiple* human experts, 2) using post-hoc XAI methods to gain access into *some* possible explanation (even if it is only a rationalization and not an accurate explanation of the AI's decision), and 3) reporting the AI's recommendation to the scientific and medical communities for further research. A third example could be that in instances where the AI's recommendation, even if incorrect, will at most result in easily reversible superficial harm or very slightly delay the patient's recovery *and* if the clinician's initial recommendation already has an extremely low likelihood of being effective (with no alternative options that contain an explanation), *and* the AI's recommendation appears to have some merit, the AI's recommendation can be used.[60]

[60] Note that this not the same as deferring to the AI, given that the clinician's judgment is nonetheless used to arrive at the decision of whether or not it is appropriate in this given situation to use the AI's recommendation.

The Rule of Nudge has the benefit of allowing the use of black-box AI in healthcare, without jeopardizing the clinicians' epistemic duties in instances of disagreement. It also avoids issues of responsibility or accountability gaps because the final decision always remains in the hands of the clinician. The clinician is responsible for any harm that befalls a patient if they yield to the assistive AI without sufficient reason to do so or if they ignore the conflicting recommendation of the assistive AI without further reassessment. On the other hand, if they follow the Rule of Nudge and take seriously the possibility of the AI as an epistemic superior in a given case without yielding to it as though it definitely *were* an epistemic superior, and if their course of action reflects this while also accounting for the specific context and the gravity of the situation, then they can be said to have fulfilled their moral responsibility and utilized the AI decision support system appropriately in their decision-making.

6.5 Conclusion

The goal of this chapter has been to determine the epistemic role of AI decision support systems and derive answers to the following three difficult questions with regard to the implementation of assistive AI in healthcare: 1. Should black-box assistive AI be utilized by clinicians?; 2. In cases of disagreement between a clinician and the assistive AI, what ought to be done?; and 3. How ought responsibility and accountability be allocated? I have presented what I have called *The Rule of Nudge* that aims to simultaneously answer all three questions. The Rule of Nudge allows for a way to implement black-box assistive AI in healthcare without putting clinicians in a position where they must choose between sacrificing their epistemic duties or their other conflicting obligations, while also removing concerns about the possibility of responsibility and accountability gaps arising with the use of AI decision support systems.

Conclusion

At the start of this dissertation, my goal was to evaluate three approaches to the implementation of AI in healthcare (the neo-luddite approach, the substitutive approach, and the assistive approach) in order to determine to what extent we are willing to grant greater and greater roles to AI in medicine and what it is that we are willing or not willing to sacrifice in the process.

I began this project with a discussion of the nature of medicine in Chapter 1, where I appealed to Eric Cassell's account of healing and Mathison and Davis's value-promotion account to argue that the promotion of patient values is an important goal of medicine, and this imposes formal duties on clinicians that include compassion, recognition, communication, creating trust, and respecting patient autonomy.

In Chapter 2, I considered and responded to several concerns with AI in healthcare that appear to favour taking the neo-luddite approach. These included concerns about biased algorithms, decreased communication with patients, confidentiality, the resurgence of paternalism, the potential of unethical design, overdependence on AI, and inscrutability. I concluded that although some of these concerns deserved serious attention, they were all insufficient to justify refraining from utilizing AI in healthcare altogether.

After providing a conceptual analysis of assistance and substitution in Chapter 3, I evaluated AI's capacity to foster a strong patient-clinician relationship in Chapter 4 and demonstrated its limited capacity to fulfill the duties of compassion, recognition, communication, creating trust, and respecting patient autonomy, and concluded that it was not an appropriate substitute for a clinician in that sense.

In Chapter 5, I raised two additional concerns with substitutive AI outside of the clinical encounter – one regarding its limited capacity to detect and correct its own erroneous and biased decisions and another regarding its likelihood to create a responsibility gap, concluding the chapter with further reasons to doubt that AI could make a suitable substitute for a human clinician.

Finally, in Chapter 6, I considered issues that are unique to the assistive approach. In particular, I discussed the issue of disagreement between clinicians and decision support systems and the difficulty in resolving these disagreements, especially when the decision support systems

are opaque. After providing an account of the epistemic role of assistive AI where I classified assistive AI as an "epistemic nudge" rather than an epistemic superior, epistemic inferior, or epistemic peer, I concluded by offering a recommendation, which I called "the rule of nudge," for handling clinician-AI disagreements and resolving concerns about the use of black-box AI in medicine and allocating responsibility and accountability in cases of clinician-AI disagreement.

I hope to have shown that the assistive approach to AI in healthcare has the most promise, all things considered. Although it raises its own unique problems that the other approaches are able to avoid, these problems are not as difficult to overcome as those with the substitutive approach, and the assistive approach does not demand refraining from incorporating AI in healthcare altogether and sacrificing all the benefits that it has to offer, like the neo-luddite approach does.

While I have attempted to provide a general framework for the implementation of AI in healthcare, it is important to keep in mind that the ethics of AI in healthcare are not a one-size-fits-all type of thing and much more work needs to be done to account for all of the healthcare field's complex parts. For instance, caring for a terminally ill patient may present clinicians with different duties than caring for a patient with a permanent loss of consciousness, and thus, AI caring for a terminally ill patient may have different ethical implications than AI caring for a patient with a permanent loss of consciousness. Similarly, it may very well be the case that while AI may be a poor substitute for a family doctor, it is a less ethically objectional substitute for a radiologist. For the time being, I have attempted to provide some important ethical considerations that ought to be kept in mind amid all the excitement that the technological advancements in medicine may generate so that we may avoid sacrificing more we intend or are willing to.

References

Albisser Schleger, H., Oehninger, N. R., & Reiter-Theil, S. (2011). Avoiding bias in medical ethical decision-making. Lessons to be learnt from psychology research. *Medicine, Health Care, and Philosophy*, *14*(2), 155–162. https://doi.org/10.1007/s11019-010-9263-2

Algren, D. A., & Christian, M. R. (2015). Buyer beware: Pitfalls in toxicology laboratory testing. *Missouri Medicine, 112*(3), 206–210.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179. https://doi.org/10.7759/cureus.35179

Alsan, M., Garrick, O., & Graziani, G. (2019). Does diversity matter for health? Experimental evidence from Oakland. *The American Economic Review, 109*(12), 4071–4111. https://doi.org/10.1257/aer.20181446

Alvarado, R. (2023). AI as an epistemic technology. *Science and Engineering Ethics, 29*(5), 32. https://doi.org/10.1007/s11948-023-00451-3

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

Asai, A., Okita, T., Enzo, A., Ohnishi, M., & Bito, S. (2019). Hope for the best and prepare for the worst: Ethical concerns related to the introduction of healthcare artificial intelligence. *Eubios Journal of Asian and International Bioethics, 29*(2), 64–70.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science, 373*(6552), 284–286. https://doi.org/10.1126/science.abg1834

Baier, A. (1986). Trust and antitrust. *Ethics, 96*(2), 231–260. http://www.jstor.org/stable/2381376

Banja, J. D., Hollstein, R. D., & Bruno, M. A. (2022). When artificial intelligence models surpass physician performance: Medical malpractice liability in an era of advanced artificial intelligence. *Journal of the American College of Radiology, 19*(7), 816–820. https://doi.org/10.1016/j.jacr.2021.11.014

Barker, K. (2019, November 15). AMP Robotics to scale production of AI-guided robotic systems for recycling. *Recycling Product News*. https://www.recyclingproductnews.com/article/32378/amp-robotics-to-scale-production-of-ai-guided-robotic-systems-for-recycling

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology, 68*(6), 1042–1054. https://doi.org/10.1037/0022-3514.68.6.1042

Bayne, T., & Levy, N. (2005). Amputees by choice: Body integrity identity disorder and the ethics of amputation. *Journal of Applied Philosophy, 22*(1), 75–86. https://doi.org/10.1111/j.1468-5930.2005.00293.x

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics* (5th ed.). Oxford University Press.

Beckers, S. (2017). AAAI: An argument against artificial intelligence. In V. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (pp. 235–247). Springer.

Beil, L. (2018, October 2). A surgeon so bad it was criminal. *ProPublica*. https://www.propublica.org/article/dr-death-christopher-duntsch-a-surgeon-so-bad-it-was-criminal

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology, 34*(3), 349–371. https://doi.org/10.1007/s13347-019-00391-6

Blanck, P. (2023). *Advanced introduction to U. S. disability law*. Edward Elgar Publishing Limited.

Bloom, P. (2016). *Against empathy: The case for rational compassion*. Ecco.

Buchanan, A. E., & Brock, D. W. (1989). *Deciding for others: The ethics of surrogate decision making*. Cambridge University Press.

Budd, K. (2019, July 9). Will artificial intelligence replace doctors? *AAMC*. https://www.aamc.org/news/will-artificial-intelligence-replace-doctors

Canadian Institute of Health Information. (2009). *Alternate level of care in Canada*. https://publications.gc.ca/collections/collection_2014/icis-cihi/H117-5-30-2009-eng.pdf

Cassell, E. J. (2012). *The nature of healing: The modern practice of medicine*. Oxford University Press.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. *The New England Journal of Medicine, 378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Chen, M. (2019). A tale of two deficits: Causality and care in medical AI. *Philosophy & Technology, 33*(2), 245–267. https://doi.org/10.1007/s13347-019-00359-6

Chomanski, B. (2019). What's wrong with designing people to serve? *Ethical Theory and Moral Practice, 22*(4), 993–1015. https://doi.org/10.1007/s10677-019-10029-3

Choudhury, A., & Asan, O. (2022). Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare

practitioners in the US. *Human Factors in Healthcare, 2*(Complete), 1–8. https://doi.org/10.1016/j.hfh.2022.100021

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Christensen, D. (2007). Epistemology of disagreement: The good news. *The Philosophical Review, 116*(2), 187–217. https://doi.org/10.1215/00318108-2006-035

Chursinova, O., & Stebelska, O. (2021). Is the realization of the emotional artificial intelligence possible? Philosophical and methodological analysis. *Filosofija. Sociologija, 32*(1), 76–83.

Clark, J., & Amodei, D. (2016, December 21). Faulty reward functions in the wild. *OpenAI*. https://openai.com/index/faulty-reward-functions

Coeckelbergh, M. (2014). Good healthcare is in the 'how': The quality of care, the role of machines, and the need for new skills. In S. P. van Rysewyk & M. Pontier (Eds.), *Machine Medical Ethics* (pp. 33–47). Springer.

Dalton-Brown, S. (2020). The ethics of medical AI and the physician-patient relationship. *Cambridge Quarterly of Healthcare Ethics, 29*(1), 115–121. https://doi.org/10.1017/S0963180119000847

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

Darling, K. (2017). "Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics* (pp. 173–188). Oxford University Press. https://doi.org/10.1093/oso/9780190652951.003.0012

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-recruitment-insightidUSKCN1MK08G

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

del Castillo Torres, G., Roig-Maimó, M. F., Mascaró-Oliver, M., Amengual-Alcover, E., & Mas-Sansó, R. (2022). Understanding how CNNs recognize facial expressions: A case study with LIME and CEM. *Sensors (Basel, Switzerland), 23*(1), 131. https://doi.org/10.3390/s23010131

Dougherty, C. J., & Purtilo, R. B. (1995). Physicians' duty of compassion. *Cambridge Quarterly of Healthcare Ethics, 4*(4), 426–433. https://doi.org/10.1017/s0963180100006241

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese, 202*, 138. https://doi.org/10.1007/s11229-023-04367-0

Dyer, C. (2000). Surgeon amputated healthy legs. *BMJ, 320*(7231), 332. https://doi.org/10.1136/bmj.320.7231.332

Elga, A. (2007). Reflection and disagreement. *Noûs, 41*(3), 478–502. https://doi.org/10.1111/j.1468-0068.2007.00656.x

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

Fernandez, A. V., & Zahavi, D. (2020). Basic empathy: Developing the concept of empathy from the ground up. *International Journal of Nursing Studies*, *110*, 103695. https://doi.org/10.1016/j.ijnurstu.2020.103695

Fletcher, G. (2015). Objective list theories. In *The Routledge Handbook of Philosophy of Well-Being* (pp. 148–160). Routledge.

Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Friedman, B., Kahn, P. H., & Borning, A. (2007). Value sensitive design and information systems. In P. Zhang & D. F. Galletta (Eds.), *Human-Computer Interaction and Management Information Systems: Foundations* (pp. 348–372). Routledge.

Gardiner, G. (2012). Understanding, integration, and epistemic value. *Acta Analytica: Philosophy and Psychology, 27*(2), 163–181. https://doi.org/10.1007/s12136-012-0152-6

Gelfert, A. (2011). Who is an epistemic peer? *Logos & Episteme, 2*(4), 507–514. https://doi.org/10.5840/logos-episteme2011242

Goldhahn, J., Rampton, V., & Spinas, G. A. (2018). Could artificial intelligence make doctors obsolete? *BMJ, 363*, k4563. https://doi.org/10.1136/bmj.k4563

Grundmann, T. (2023). The possibility of epistemic nudging. *Social Epistemology, 37*(2), 208–218. https://doi.org/10.1080/02691728.2021.1945160

Guerra, G. A., Hofmann, H., Sobhani, S., Hofmann, G., Gomez, D., Soroudi, D., Hopkins, B. S., Dallas, J., Pangal, D. J., Cheok, S., Nguyen, V. N., Mack, W. J., & Zada, G. (2023). GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. *World Neurosurgery, 179*, e160–e165. https://doi.org/10.1016/j.wneu.2023.08.042

Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. K. (2001). Trust in physicians and medical institutions: What is it, can it be measured, and does it matter? *The Milbank Quarterly, 79*(4), 613–639. https://doi.org/10.1111/1468-0009.00223

Halpern, J. (2003). What is clinical empathy? *Journal of General Internal Medicine, 18*(8), 670–674. https://doi.org/10.1046/j.1525-1497.2003.21017.x

Hannon, M. (2021). Recent work in the epistemology of understanding. *American Philosophical Quarterly (Oxford), 58*(3), 269–290.

Hardin, R. (1996). Trustworthiness. *Ethics, 107*(1), 26–42. https://doi.org/10.1086/233695

Hart, H. L. A. (2008). Postscript: Responsibility and retribution. In *Punishment and Responsibility* (pp. 210–237). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199534777.003.0009

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935

Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times*. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Hindriks, F., & Douven, I. (2017). Nozick's experience machine: An empirical study. *Philosophical Psychology, 31*(2), 278–298. https://doi.org/10.1080/09515089.2017.1406600

Hirmiz, R. (2023). Against the substitutive approach to AI in healthcare. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00347-9

Huang, S., Mamidanna, S., Jangam, S., Zhou, Y., & Gilpin, L. H. (2023). Can large language models explain themselves? A study of LLM-generated self-explanations. *arXiv*. https://doi.org/10.48550/arxiv.2310.11207

Ikäheimo, H. (2002). On the genus and species of recognition. *Inquiry (Oslo), 45*(4), 447–462. https://doi.org/10.1080/002017402320947540

Jagosh, J., Donald Boudreau, J., Steinert, Y., MacDonald, M. E., & Ingram, L. (2011). The importance of physician listening from the patients' perspective: Enhancing diagnosis, healing, and the doctor–patient relationship. *Patient Education and Counseling*, *85*(3), 369–374. https://doi.org/10.1016/j.pec.2011.01.028

Jongsma, K. R., & Sand, M. (2022). Agree to disagree: The symmetry of burden of proof in human–AI collaboration. *Journal of Medical Ethics, 48*(4), 230–231. https://doi.org/10.1136/medethics-2022-108242

Kachalia, A., & Mello, M. M. (2013). Breast cancer screening: Conflicting guidelines and medicolegal risk. *JAMA : The Journal of the American Medical Association*, 309(24), 2555-2556. https://doi.org/10.1001/jama.2013.7100

Kagan, S. (1997). *Normative ethics*. Westview Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus & Giroux.

Kass, L. R. (1975). Regarding the end of medicine and the pursuit of health. *The Public Interest, 40*(40), 11–42.

Kee, J. W. Y., Khoo, H. S., Lim, I., & Koh, M. Y. H. (2018). Communication skills in patient-doctor interactions: learning from patient complaints. *Health Professions Education*, *4*(2), 97–106. https://doi.org/10.1016/j.hpe.2017.03.006

Kelly, T. (2005). The epistemic significance of disagreement. In J. Fantl, M. McGrath, & E. Sosa (Eds.), *Contemporary epistemology: An anthology* (pp. 167-196). Wiley.

Kelly, T. (2011). Peer disagreement and higher order evidence. In A. I. Goldman & D. Whitcomb (Eds.), *Social epistemology: Essential readings* (pp. 183-217). Oxford University Press.

Kempt, H., Heilinger, J.-C., & Nagel, S. K. (2023). "I'm afraid I can't let you do that, Doctor": Meaningful disagreements with AI in medical contexts. *AI and Society, 38*, 1407–1414. https://doi.org/10.1007/s00146-022-01418-x

Kempt, H., & Nagel, S. K. (2022). Responsibility, second opinions and peer-disagreement: Ethical and epistemological challenges of using AI in clinical diagnostic contexts. *Journal of Medical Ethics, 48*(4), 222-229. https://doi.org/10.1136/medethics-2021-107440

Killam, K. (2014). Building empathy in healthcare: A Q&A with Dr. Helen Riess of Harvard Medical School about her efforts to nurture empathy among health care workers. *Greater Good Magazine*. https://greatergood.berkeley.edu/article/item/building_empathy_in_healthcare

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Kruse, C. S., Frederick, B., Jacobson, T., & Monticone, D. K. (2017). Cybersecurity in healthcare: A systematic review of modern threats and trends. *Technology and Health Care, 25*(1), 1-10. https://doi.org/10.3233/THC-161263

Lane, A., McCoy, L., & Ewashen, C. (2010). The textual organization of placement into long-term care: Issues for older adults with mental illness. *Nursing Inquiry, 17*(1), 3–14. https://doi.org/10.1111/j.1440-1800.2009.00470.x

Laporte, A., Rohit Dass, A., Kuluski, K., Peckham, A., Berta, W., Lum, J., & Williams, A. P. (2017). Factors associated with residential long-term care wait-list placement in North West Ontario. *Canadian Journal on Aging, 36*(3), 286–305. https://doi.org/10.1017/S071498081700023X

London, J. A. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *The Hastings Center Report, 49*(1), 15–21. https://doi.org/10.1002/hast.973

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Neural Information Processing Systems (NeurIPS).* https://doi.org/10.48550/arXiv.1705.07874

MacGill, M. (2017, December 11). What do anesthesiologists do? *Medical News Today*. https://www.medicalnewstoday.com/articles/287857#what-does-an-anesthesiologist-do

Maruf, R. (2022, July 25). Google fires engineer who contended its AI technology was sentient. *CNN*. Retrieved from https://www.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html

Mathison, E., & Davis, J. (2021). Value promotion as a goal of medicine. *Journal of Medical Ethics, 47*(7), 494–501. https://doi.org/10.1136/medethics-2019-106047

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McQuillan, L. (2022, June 24). A Google engineer says AI has become sentient. What does that actually mean? *CBC News*. https://www.cbc.ca/news/science/ai-consciousness-how-to-recognize-1.6498068

Meehan, D. (2020). Epistemic vice and epistemic nudging: A solution? In G. Axtell & A. Bernal (Eds.), *Epistemic paternalism: Conceptions, justifications and implications* (pp. 249-261). Rowman & Littlefield International.

Mowshowitz, Z. (2022, December 2). Jailbreaking ChatGPT on release day [Substack newsletter]. *Don't Worry About the Vase*. Retrieved from https://thezvi.substack.com/p/jailbreaking-the-chatgpt-on-release

Mukherjee, S. (2017, March 27). A.I. versus M.D.: What happens when diagnosis is automated? *The New Yorker*. https://www.newyorker.com/magazine/2017/04/03/ai-versus-md

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital System, 13*(1), 911–921. http://dx.doi.org/10.12785/ijcds/130172

Nguyen, C. T. (2022). Trust as an unquestioning attitude. In T. S. Gendler, J. Hawthorne, & J. Chung (Eds.), *Oxford studies in epistemology* (Vol. 7). Oxford University Press.

Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. The Belknap Press of Harvard University Press.

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

O'Neill, O. (2002). *Autonomy and trust in bioethics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511606250

Parfit, D. (1986). *Reasons and persons*. Oxford University Press.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). Information security for small business *(*NIST IR 8312). *National Institute of Standards and Technology.* https://doi.org/10.6028/NIST.IR.8312

Parks, R v., 140 N.R. 161 (SCC 1992)

Pritchard, D. (2009). Knowledge, understanding and epistemic value. *Royal Institute of Philosophy Supplement, 64*, 19–43. https://doi.org/10.1017/S1358246109000046

Rhimes, S. (Writer) (2006a). *It's the end of the world* (Season 2, Episode 16) [TV series episode]. In *Grey's Anatomy*. ABC.

Rhimes, S. (Writer) (2006b). *As we know it* (Season 2, Episode 17) [TV series episode]. In *Grey's Anatomy*. ABC.

Rhimes, S. (Writer) (2007). *Crash into me: part 1* (Season 4, Episode 9) [TV series episode]. In *Grey's Anatomy*. ABC.

Rice, C. M. (2013). Defending the objective list theory of well-being. *Ratio, 26*(2), 196–211.

Riess, H. (2010). Empathy in medicine – A neurobiological perspective. *The Journal of the American Medical Association, 304*(14), 1604–1605. https://doi.org/10.1001/jama.2010.1455

Ross, C., & Herman, B. (2023, March 13). Denied by AI: How Medicare Advantage plans use algorithms to cut off care for seniors in need. *STAT*. https://www.statnews.com/2023/03/13/medicare-advantage-plans-denial-artificial-intelligence/

Ross, C., & Herman, B. (2023, November 14). UnitedHealth faces class action lawsuit over algorithmic care denials in Medicare Advantage plans. *STAT*. https://www.statnews.com/2023/11/14/unitedhealth-class-action-lawsuit-algorithm-medicare-advantage/

Ross, W. D. (2002). What things are good? In P. Stratton-Lake (Ed.), *The right and the good* (pp. 134-141). Oxford University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Sauer, H. (2017). *Moral judgments as educated intuitions*. The MIT Press.

Sebastian, A. M., & David, P. (2022). Artificial Intelligence in cancer research: Trends, challenges and future directions. *Life, 12*(12). https://doi.org/10.3390/life12121991

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Semuels, A. (2020, August). Millions of Americans have lost jobs in the pandemic—And robots and AI are replacing them faster than ever. *Time Magazine*. https://time.com/5876604/machines-jobs-coronavirus/

Simmons, R. (2016). Quantifying criminal procedure: How to unlock the potential of big data in our criminal justice system. *Detroit College of Law at Michigan State University Law Review, 4*, 947–1017. https://dx.doi.org/10.2139/ssrn.2816006

Song, Y., Nie, T., Shi, W., Zhao, X., & Yang, Y. (2019). Empathy impairment in individuals with autism spectrum conditions from a multidimensional perspective: A meta-analysis. *Frontiers in Psychology*, *10*, 1902–1902. https://doi.org/10.3389/fpsyg.2019.01902

Sparrow, R. (2004). The Turing triage test. *Ethics and Information Technology, 6*(4), 203–213. https://doi.org/10.1007/s10676-004-6491-2

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines, 16*(2), 141–161. https://doi.org/10.1007/s11023-006-9030-6

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sparrow, R. (2016). Robots in aged care: A dystopian future? *AI & Society, 31*(4), 445–454. https://doi.org/10.1007/s00146-015-0625-4

Sparrow, R., & Hatherley, J. J. (2019). The promise and perils of AI in medicine. *International Journal of Chinese and Comparative Philosophy of Medicine, 17*(2), 79–109.

Steensma, D. (2016). The Farid Fata Medicare fraud case and misplaced incentives in oncology care. *Journal of Oncology Practice, 12*(1), 51–54. https://doi.org/10.1200/JOP.2015.008557

Taylor, C. (1992). The politics of recognition. In A. Gutmann (Ed.), *Multiculturalism: Examining the politics of recognition* (pp. 25–73). Princeton University Press.

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth and happiness*. Penguin.

Thompson, M. (2023). Ontario hospitals looking to cut staff costs, speed up work despite ongoing staffing crisis. *Press Progress*. https://pressprogress.ca/ontario-hospitals-cut-staff-costs-speed-up-work-staffing-crisis/

Tigard, D. W. (2021). Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics, 30*(3), 435–447. https://doi.org/10.1017/S0963180120000985

Tiwari, P., Prasanna, P., Wolansky, L., Pinho, M., Cohen, M., Nayate, A. P., Gupta, A., Singh, G., Hatanpaa, K. J., Sloan, A., Rogers, L., & Madabhushi, A. (2016). Computer-extracted

texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: A feasibility study. *American Journal of Neuroradiology: AJNR*, *37*(12), 2231–2236. https://doi.org/10.3174/ajnr.a4931

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication & Ethics in Society, 16*(3), 252–260. https://doi.org/10.1108/JICES-06-2018-0056

Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology, 35*(4), 88. https://doi.org/10.1007/s13347-022-00577-5

Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics, 2*(4), 815–826. https://doi.org/10.1007/s43681-022-00142-y

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology, 28*(1), 107–124. https://doi.org/10.1007/s13347-014-0156-9

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

White, D., & Pope, T. (2015). Medical futility and potentially inappropriate treatment. In S. J. Youngner & R. M. Arnold (Eds.), *The Oxford handbook of ethics at the end of life* (pp. 65–86). Oxford University Press.

Wu, Z., Xuan, S., Xie, J., Lin, C., & Lu, C. (2022). How to ensure the confidentiality of electronic medical records on the cloud: A technical perspective. *Computers in Biology and Medicine*, *147*, 105726–105726. https://doi.org/10.1016/j.compbiomed.2022.105726

Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y., & Clifton, D. A. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Medicine*, *6*(1), 55–55. https://doi.org/10.1038/s41746-023-00805-y