AIM, Philosophy and Ethics

Stephen Rainey, Yasemin J. Erden, and Anais Resseguier

Contents

1 Introduction 2

2 Promises of AI in Medicine 2

3 AI and Medical Epistemology: A Changing Paradigm 3
3.1 Data 3
3.2 Data-Utopianism 4
3.3 Data Curation and Use 5

4 AI and Medical Epistemology: Limits, Risks, and Biases 6
4.1 Human Biases and Prejudices: Language and Interpretation 7
4.2 Computational Biases: Programming and Algorithms 8

5 AI and Medical Ethics 9
5.1 The Patient-Doctor Relationship 10
5.2 The Medical Profession in the Era of Digital Capitalism 11

6 Conclusion and Recommendations 12

References 12

Abstract

This chapter explores AI through a philosophical and ethical lens. This includes an examination of how AI impacts on medicine in terms of uses and promises, limitations, and risks, as well as key questions to consider. While AI offers scope for complex and large-scale data processing, with the promise of an increase in efficiency and precision, some central limitations need to be highlighted. The use of AI also brings some pertinent and predictable, as well as unpredictable risks, such as those due to biases. Also considered is what may be lost where AI replaces established processes, not least those relational and interpersonal aspects that are central to healthcare. By covering these and related issues, this chapter offers ways to evaluate, and also balance, key benefits and

S. Rainey (✉)
University of Oxford, Oxford, UK
e-mail: stephen.rainey@philosophy.ox.ac.uk

Y. J. Erden
University of Twente, Enschede, The Netherlands
e-mail: y.j.erden@utwente.nl

A. Resseguier
Trilateral Research, Waterford, Ireland
e-mail: anais.resseguier@trilateralresearch.com

N. Lidströmer, H. Ashrafian (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-58080-3_243-1

risks arising from the application of AI to the medical sector.

Keywords

Philosophy · Artificial intelligence · Medicine · Care · Bias · Data · Medical epistemology · Medical ethics · Algorithms

1 Introduction

This chapter examines AI through a philosophical and ethical lens; it explores some fundamental impacts on medicine in terms of promises, limitations, risks and key questions. On the one hand, AI has the capacity to process huge amounts of data on medical issues. This may increase efficiency and precision for diagnosis and treatment and offer progress in medical research. On the other hand, the possibility emerges of a drift toward over optimistic techno-solutionism in medicine and healthcare. This includes the application of problem-solving approaches which are apt for data-centered practice, but which are not obviously improvements when it comes to other dimensions of medicine, especially its relational aspect. Whereas medicine in general has a clear anchoring in scientific accounting for the biological body, it also has responsibilities in recognizing subjective, interpersonal, sociopolitical, and historical realities about health and illness. The inclusion of AI in medical practice raises interesting and challenging philosophical, ethical, as well as practical questions. In exploring these topics related questions about asymmetric interpersonal relations in medicine, and on topics of personal identity as they relate to AI and medicine will be considered.

This chapter begins by considering the promises of AI (Sect. 2), but gives only limited time to this endeavor since many of the chapters in this volume offer more detailed examples of how AI can fruitfully be used in medicine. The primary aims in this chapter therefore concern on the one hand, AI in relation to philosophy, and more specifically in terms of medical epistemology (Sect. 3), and on the other, in relation to medical ethics (Sect. 5). The argument is that the epistemological issues are at the core of many of the ethical issues that follow, and so this earlier section is necessarily longer and more detailed. The suggestion is that understanding the limitations of AI in terms of issues arising from data and from bias will enable the reader to more easily anticipate and assess the other ethical issues that are outlined in Sect. 5. Many of these also overlap with, or build on, those earlier issues and so can be explained more concisely by that stage.

Before beginning, it is important to note that AI and ethics is a vast area, and there is plenty of literature already written on this topic generally [1], as well as in terms of AI as it pertains to medicine, healthcare and clinical settings. See for instance recent work on whether AI and medical principles can align [2], on the application of AI to specific areas of medicine, including intensive care [3] and on the use of AI in surgery [4]. Meanwhile there are also texts that offer high level mapping, thereby outlining general ethical issues as they relate to AI in healthcare [5]. Mapping offers a useful general picture, while detailed accounts on single applications offer scope for focused analysis of specific issues. However, the aim in this chapter is to offer a practical, middle ground that introduces the reader to a few central topics that are core to thinking philosophically and ethically about AI in medicine. By restricting the focus in this way there can be some detailed discussion of each, while not precluding the importance of other ethical issues relevant to these topics. The reader can therefore use this chapter as an introduction to some foundational issues, and a way to begin the important work of ethical and philosophical analysis of AI as it pertains to medicine. It should thereby be considered a stepping stone to further reading and analysis.

2 Promises of AI in Medicine

AI applications in medicine may be claimed to be more objective than standard, human-based approaches. Some of the methods employing AI might appear able, to some extent, to bypass

clinician bias. Where, for instance, a human may see a person and make snap judgements or harbor implicit biases about them, AI "sees" only data (this apparent potentiality is explored more critically later in this chapter). This hope for objectivity would be a boost for justice in healthcare. Certainly, it may be true in cases whereby AI is employed to assist in the managing or assessment of data and/or in ensuring that approaches to clinical procedures and processes are consistent and rigorous. While there may be these improvements brought by AI in medicine in terms of justice – through more equitable access or distribution of resources – they may alternatively be seen simply as efficiency boosts for the practical implementation of medical decision-making. Where AI can take over more laborious tasks that otherwise tie up clinicians' time, this can free those clinicians to do more valuable work.

A large amount of what AI can do in medicine centers on imaging, given the aptitude for pattern recognition in AI applications. This can be seen in radiology, pathology, ophthalmology, and dermatology [6]. In these areas, AI can be used to detect fractures from scanning x-ray images, or potential eye or skin problems from processing photographs of retinas and skin. In this way, the system can assist clinicians to prioritize their work. Besides imaging, an emerging and potentially very powerful application comes from classifying text. Clinicians make detailed notes as they work, including their observations and practices. AI can be set loose on amassed clinical note data, and spot common findings, approaches, mistakes, and inefficiencies [6]. This processing can then lead to recommendations that would allow protocols to be fine-tuned based on analysis of many instances of clinical notes. AI could do this classification in an unsupervised way, freeing up time for clinicians.

Even where the promises of objectivity and efficiency are assumed, a further query arises. *How sure can we be that AI can deliver on those promises?* It isn't automatically clear that more and more data applied to individual cases is necessarily better than careful attention to the individual specifics. Imaging and clinical note consolidation are valuable applications, but as explored below it would be prudent not to expect an AI replacement for clinical expertise in general, given much can be lost as well as gained. One risk to be avoided is that of imbuing AI with an "enchanted" status [7], which would serve to close down scrutiny of those applications. Advances in AI technology don't inevitably lead to improvements in applications. They may represent changes to applications that require evaluation on their own terms.

3 AI and Medical Epistemology: A Changing Paradigm

3.1 Data

Doctors have always collected data (observations, case notes, research findings) in order to come to diagnosis. Typically, these data have been handled so as to inform medical theories, bolstering clinical insights through broad observational regularities. One hope for the future of medicine includes the idea that AI can be used on data in a much more directly instrumental way in order to boost the efficiency, predictive power, and ultimately the effectiveness of medicine. Such a data-centric approach – where research, diagnosis, and treatment primarily derive from the processing of digitized health data, rather than patient observation – is thought by some to permit a computational approach to medicine. This would diminish the role of medical theories, and expertise, replaced by exploratory data science [8].

The promise of data in general, and big health data specifically, is that it can represent vast arrays of knowledge based on samples, processed in various ways, without guiding theoretical knowledge required. The data-centric approach surpasses limitations present in case study or cohort analyses by aggregating wide ranges of quantitative and qualitative observations. This provides scales not available by other means. These mass aggregates of data can be transformed into new knowledge, through applying statistical transformations and pattern analyses. At its simplest, the idea is that where there are patterns in data, there can be reasons to explain those patterns [9]. These

patterns would be taken to signify relationships among arrays of observations. Doubtless, owing to the way in which humans are primed to spot patterns anyway – like pareidolia and seeing faces in clouds – such things will be seen frequently. But there will be a set of patterns among datasets that are not mere coincidence. By exploring data closely, these can be shifted from the burgeoning whole.

Patterns found within data that aren't mere coincidence will have some other causal explanation. This means that the patterns in the data, representing structures implicit among the observations of some phenomena, will reveal casual structures among those phenomena. The relationships among data such that a pattern emerges are thought to reveal fundamental information about whatever the researcher is investigating in this way. And in finding new causal structures among phenomena by examining patterns in data, the formulation of new insights into those phenomena are enabled by examining the data. This in turn can be used as a basis for new practical approaches to the field. This has application in medicine where medical and other data can be combined and from the whole predictions made to explain patterns. The patterns in data are then expected to relate to causal factors in diagnosis, prognosis, health outcomes, treatment, and so on.

For example, demographic information might be aggregated into a large dataset, along with clinical data, genetic information, observations from specimen biological samples, and other meta-data [10]. By turning to data analysis, relationships among this trove of heterogeneous material might serve to yield patterns that suggest underlying structures among lifestyle, medical history, illness, and health outcomes [11]. This in turn can offer novel prevention, diagnostic, and treatment strategies for an illness, or even suggestions for public health policy. The insights would be gained from predictions based on patterns among data. This means whole populations needn't be interviewed, nor even specifically sampled for a purpose. The data processing appears to do it alone. Nevertheless, in order to curate, populate, maintain, and operationalize such datasets with sufficient quality, expert knowledge is required across a number of disciplines.

Data scientists, software and hardware engineers, and developers are needed to ensure quality systems and structures. Medical expertise is also required, able to identify health-relevant data and connections between that and other data. Likewise, to point out what is irrelevant, meaningless, obviously wrong, etc. Clinical expertise aside, there are also technical decisions that must be made about data which can bear upon their medical relevance. Indeed, how and in what respects data are accurate to the medical phenomena they represent is a question in need of careful scrutiny. As other contexts have shown, data isn't neutral with respect to its collection, storage, or mediation [12, 13].

3.2 Data-Utopianism

In an idealized data-utopian scenario, the AI approach to medicine would constitute what has been in another context termed a "screen and intervene" paradigm [14]. This is not "diagnosis" as it is known today, where interview, examination, and observation are essential elements. Biomarkers are of central importance in a more datafied approach. AI will look for these as patterns in data, on the basis that over time they have been established as indicators of illness. But biomarkers are not without problems in themselves, owing at least in part to questions over consistency and standardization [15]. The reduction to biomarkers represented as patterns in data, apt for automated detection, bypasses pertinent questions including some concerning what is being detected, and why that is being looked for in the first place (see Sect. 3.3 below).

It might be that the promises of AI prompt the datafication of medical investigation too quickly. If the data is approached in a spirit of exploration, and so too is the diagnostic field, then there appear to be overlapping explorations without specific guiding strategies. In human behavior generally, Tversky and Kahneman provide examples of decision-making under uncertainty wherein it is not necessarily helpful to gain more and more

information [16]. That is, when the future is uncertain, those facing a decision can be hampered in making optimal choices if they are presented with more to think about. Yet in this context of multiple uncertainties and explorations, the solution is taken to include amassing data. This is not without risk, especially in medicine where human health and wellbeing is at stake.

Unlike examples of human reasoning examined by Tversky and Kahneman, the AI approach doesn't proceed by gathering information and then coming up with decisions. In certain respects, AI centering on data turns traditional medical-scientific investigation on its head. Scientific research is often thought of in terms of heavily disciplined normal human reasoning: hypothesis-formation, investigation, then confirmation or refutation in the light of evidence gathered specifically for a carefully defined purpose. A typical way to characterize a clinical encounter might be along these lines: a patient presents with a complaint. The clinician carries out an examination, and makes observations. Based on these, and expert knowledge, a clinical diagnosis is made, following which, treatment or further examination is recommended. This stresses interpersonal skills, like empathy and including patients in the process overall [17]. The data-centered paradigm emphasizes a more thoroughly empirical drive, in which facts, in the shape of digitized data, are prioritized over testimony.

This approach instead amasses data from a variety of sources, and seeks hypotheses based on what emerges from the data. This is an exploratory approach that claims to need no guiding theory, instead seeking correlative information among data to prompt explanatory work. The models to explain patterns are "born from the data" [18]. Such approaches have seen some success in areas like predictive data analytics to forecast likely Parkinson's Disease development [19]. The data can also be harnessed to make narrow predictions about disease development for specific patients. Again, using Parkinson's as an example, whether a patient is likely to suffer falls or not can be taken from analysis of swathes of data far beyond the scale of comprehension of an individual clinician [20]. This kind of predictive power concerning likely disease course is made available by data, and seems a dimension not open to the interpersonal clinical encounter. How to conceptualize the potential pros and any emerging cons of this approach, and how to weigh them against one another, remains a difficult endeavor. The following sections explore this difficulty in order to throw some light on what otherwise remains obscure.

3.3 Data Curation and Use

The aggregation of heterogeneous, large datasets, taken from myriad sources, makes the data in these applications complicated to deal with. This raises ethical questions about data ownership, privacy, consent, purpose, re-use, anonymity, and others, as well as the nature of dataset-making as a scientific, sociopolitical, and technological endeavor [21]. Once created, the analysis of these huge and varied datasets cannot be carried out by humans. Owing to the scale and complexity of data, algorithms and machine processing techniques more widely are required. Using techniques like Compressive Big Data Analytics (CBDA), algorithmic processing of data sets aims to be "model free" or "model agnostic" [22]. This kind of analysis can be taken to imply the objective nature of the data being collected, and of its subsequent analysis. But this objectivity is anything but secure.

One of the largest medical AI datasets at the moment is known as "ChestX-ray14." As Kulkarni et al. discuss [6] this is an interesting case that illustrates some of the issues raised here. This dataset was used in a study to train an AI-detection model called "CheXNet." In order to know what CheXNet should look for, a "ground truth" had to be established such that positive cases might be defined. This would represent what CheXNet was looking for. Ground truthing was carried out by text mining clinicians' radiology reports. This is the inclusion of medical expert opinion in the course of training the AI, mentioned above as necessary. But, "Intriguingly, CheXNet's performance mirrored human weaknesses in many respects; the algorithm had much

greater accuracy in detecting hiatal hernias, a radiographically distinctive diagnosis, compared to pulmonary infiltration, which is frequently ill-defined" [6].

This is taken to be the case because the notes from which the dataset is comprised, and on which the AI diagnostic model was trained, themselves are replete with uncertainty. Diagnoses aren't always binary cases. One can speculate that the kinds of uncertainty present throughout clinical notes are exactly the sort of thing expertise is well honed to draw conclusions from, or otherwise use in informed medical decision-making. If the data is all there in the dataset, and machine learning techniques discover patterns in that set, then there is a ring of objectivity to it. The pattern is really there, somehow. But there are uncertainties in the data that are themselves hallmarks of medical expertise. Radiologists rightly record their uncertainties in their notes. What's more, besides these kinds of responsible uncertainties, there are other potentially unconscious or historically derived biases in data. This is especially so where women and people of color are poorly represented in health data. Successful, data-driven skin cancer lesion detection algorithms, for instance, are effective mainly for light-skinned people [23]. The dataset here is incomplete following specific sociopolitical non-inclusivity. Issues such as these may be compounded if they become hidden behind the supposed objectivity of data processing.

Just as there are questions about dataset creation and its completeness, illustrated here with ChestX-ray14, others arise around algorithmic processing (see next section). How these questions and how the requirements of data in general relate to medical reasoning is in need of careful scrutiny. It is not clear that by deploying data in their investigations, clinicians will necessarily get closer to better answers simply owing to datafication.

There remain those who emphasize that the use of data ought to be handled judiciously, and used in tandem with expert clinical decision making [24]. Here, tools are deployed as decision support rather than considered as silver bullets. As with the scope for predicting falls in Parkinson's patients, data here can provide value. A clinician, faced with specific observations about their patient can draw upon arrays of information relating to similar observations. The singular case can thereby be related to a population-level sample. This could aid clinicians and patients alike in providing a rich backdrop to the otherwise one-to-one clinical encounter. It could further serve to minimize healthcare disparities by boosting clinician confidence and performance, not least through eliminating more burdensome dimensions of clinical work [25].

This would ideally translate into better patient outcomes, with world-class healthcare data available in even the most under resourced locations. While this promise sounds worthwhile, it does entail some changes to expected medical practice. The biostatistical reduction of the patient to correlations among data points not only methodologically detaches the person from their body and his/her sociopolitical context, but also the functioning of that specific body from clinical observation. This is replaced by an aggregating view of bodily function and deficit. The clinical encounter then takes on the function of harmonizing the deficient with the aggregate, mediated in depersonalized data processes. This represents a challenge to the traditionally interpersonal doctor-patient relationship. Whether and how this represents a problem, or a medical advance is an open question, though one addressed below.

4 AI and Medical Epistemology: Limits, Risks, and Biases

There are certain inevitabilities with regard to the limitations of a technology, including that it will serve more or less narrow purposes, and that it will be defined by those who imagine, determine, fund, and build it. In those senses then, "limits" and "biases" can be understood in terms of the parameters within which a technology is designed and developed. This includes the aims, ambitions, and outcomes, as well as the structure within which it comes to be, e.g., financial and political. A preference to develop one method and not another would count as a nontrivial bias on that

account, and the development of a technology that can do one thing but not another would count as one of its not unreasonable limitations (cf. [26]). Such limits and biases need not be either inherently negative or positive, nor are they necessarily problematic. Whether a limit or a bias matters would likely be linked to whether the technology "works" in fulfilling its aims and objectives, considered in context, and whether its benefits outweigh any harms.

This section turns to those limitations and biases that are necessarily negative, and in those respects the concern is with problematic limitations and biases that need not occur. These weaken the overall quality of the technology and its impact, and may cause a variety of harms, including physical, psychological, social, and environmental. The argument is that these kinds of limitations and biases are tied to a number of predictable, sometimes inevitable, negative consequences and therefore also risks. There are many negative biases that impact on AI, some of which are discussed here under two categories. The first concerns human biases and prejudices, which includes the kinds of biases contained and expressed in language as well as in interpretative endeavors including perception. The second concerns computational biases, and these are biases that are contained in programming and algorithms. This latter variety emerges to some extent from human biases, whether implicit or explicit, such as in the selection or categorization of data.

4.1 Human Biases and Prejudices: Language and Interpretation

Biases are unavoidable to some extent, and they can be either implicit, e.g., unacknowledged or even to some extent unknown, or explicit, e.g., stated, acknowledged, and to some extent known. Whether a bias is implicit or explicit, it can be demonstrated in both language and interpretation, i.e., words that are chosen, used, and understood, as well as in judgement and action, i.e., in decisions made and resulting behaviors. These biases may be intentional or unintentional, and they can result in a variety of consequences. For instance, they can be seen in a tendency to connect certain illnesses to gender or culture, ethnicity, or socio-economic context. There are many examples of these biases, especially negative, in medical practice.

For instance, it has been suggested that there is a link between diagnosis of borderline personality disorder (BPD) and gender, including the possibility that it is overdiagnosed in young women [27], and elsewhere there is evidence to suggest that schizophrenia is overdiagnosed in young black men [28]. Meanwhile the expectation that someone who is overweight will necessarily suffer as a result of their weight has led to non-weight-related illnesses and diseases being misdiagnosed, missed, and even ignored [29]. Bias can also be tied to judgements made about a person based on their perceived class and socio-economic circumstances, including as these intersect with ethnicity [30]. Such problematic judgements are sometimes equated with understanding and knowledge. For instance, biases tied to perceptions of ability and disability intersect with gender with the outcome that disabled people describe being ignored or dismissed [31]. For example, people with visible disabilities report a lack of recognition regarding their complex situations. This includes where the focus centres only the disability, to the neglect of other topics, such as basic checks for blood pressure and cholesterol, and tests related specifically to prevention of disease. Patients report a lack of effective communication, particularly acute for those patients with severe disabilities, which sometimes results in excessive communication between physicians and caregivers to the neglect of the patient's own perspectives [32].

How much credibility is given to a speaker, and how much credence to their testimony, has clear, direct, and serious consequences in medical contexts. A lack of care to the specificity of a patient's concrete situation can lead to epistemic injustice, whereby negative biases and prejudices impact on the scope within which the speaker's credibility is assessed and what then follows in terms of time given to their account, as well as outcomes and decisions (cf. [33, 34]). The kinds of biases in the examples noted above might be explicitly or

verbally expressed as prejudices, or they can be seen in the terminology that is used and the actions that follow. For instance, where the term "hysterical" subsumes and thereby also neglects a whole set of (typically women's) physical and psychological health conditions [35]. Biases can also be enacted without being expressed, for instance, in the framing of a physician's expectations and actions, which may nevertheless be recognized by patients, and which can frame their experience of the medical experience, as per the examples described above.

Recognition of some key principles can help to avoid such problematic biases in medical contexts. First, that empirical observation is never neutral, and as such empirical positions need to be viewed as at least partly normative [29]. Second, that the clinical gaze of individual physicians is itself normatively structured. As Foucault [36] argues, "The clinical gaze is not that of an intellectual eye that is able to perceive the unalterable purity of essences beneath phenomena. It is a gaze of the concrete sensibility, a gaze that travels from body to body, and whose trajectory is situated in the space of sensible manifestation." This concrete sensibility is tied to contexts with specificity, cultural meaning, and, inevitably, more or less prejudice. Third, embodied humans have tacit bodily knowledge, which can be considered in terms of intercorporeal ways of knowing, i.e., as beings-in-the-world ([29] ibid.). It will suffice for now to highlight that the necessarily concrete nature of subjective experience both feeds into and is informed by biases. Even if are ways in which to recognize and mitigate these. The suggestion here is that these principles can also inform how AI is developed and used in medical contexts.

4.2 Computational Biases: Programming and Algorithms

Without careful attention, biases and prejudices can feed into both the data (see previous section) and the production of the algorithms on which AI relies. In such instances, a negative bias captured or reified in a technology like AI is always morally problematic regardless of whether the technology is otherwise "successful" in its practical aims. An ethical approach to AI requires that people and ethics should not be sacrificed on the altar of "faster" or more efficient technologies, for instance benefiting some while harming others. It is worth recognizing that negative biases can, and often do, affect the overall success of a technology, but that conflating practical solutions with ethical solutions is not sufficient, and the ethical ought not to be subsumed within the practical. For instance, Google's immediate response to the racism caused by their facial recognition algorithms – that identified some black people as gorillas – was to amend the recognition categories, i.e., by removing the identification of gorillas [37]. This technical workaround might solve a short-term practical problem, but it is very far from an ethical solution. The latter requires instead a broader recognition of the problems and their causes, as well as a wider range of solutions, some of which include greater representation both in terms of data and in the teams developing the AI [38].

To consider ethics first includes the recognition that human biases impact on computational biases, and that to some extent this may be unavoidable. It has been suggested that linguistic biases inevitably find their way into programming, and thereby into AI [39]. The argument suggests that biases are inherent to language, and since language is the framework within which an AI is developed and structured, it is also inevitable that bias will find its way into the programming. If this is accepted to be the case, then it becomes clear that vigilance for bias may not be enough. In medicine, this is especially problematic, and there are already many examples of where harm can occur if medical practitioners bring prejudices and biases to their practice, as discussed above. It is therefore essential to take this into account when planning, designing or using AI, if the risk of simply replicating and reifying those same biases is to be avoided in the development and use of these new tools. Where AI is used for automated decision-making and judgement, rather than as a tool for data management, the risk is even greater given the possibility that such systems will become embedded in medical

structures and future changes may be difficult or even impossible.

Once it is accepted that bias may be unavoidable, mitigation of negative biases in particular needs to be considered. For instance, it is clear that data scientists and engineers will not necessarily have expertise about the medical fields for which an AI is being developed. Evidence for this can be found in the language about medical conditions in papers that primarily describe the development of AI systems for use in those fields, e.g., the development of AI systems for the assessment of autism (cf. [40]). AI developers working on technologies for medical applications may demonstrate cursory understanding at best, and limited or flawed interpretations at worst. It's clear that the task of developing AI for fields outside of computing and engineering ought to be an interdisciplinary endeavor. Yet when expert knowledge is sought, this can lead to a replication of the biases of those whose theories are then privileged. As with the biases noted above, this may be inevitable to some extent. An expertly-informed AI system can serve to amplify the practice of the experts chosen to inform it. The views and preferences in theory and practice that are evident in an expert's perspective are implicit or explicit choices, judgements, and decisions. In selecting an expert to provide input to an AI system's development, elements of that specific perspective are being tacitly endorsed. Selecting expert advice is thereby, to some extent, omitting the theories and ideas that might animate the practice of an equally expert, but contrasting practitioner.

The above situation is not itself unusual, in so far as any clinician would also bring their preferences and biases to their individual practice. A key difference, however, concerns the scope and the reach that can be achieved by an embedded AI. In other words, what happens once a technology becomes ubiquitous? In the case of the individual physician, their prioritizing of certain theories, to the neglect of others, in their individual practice may only impact on the number of people who are patients in that practice, as well as whatever influence they may have on those they train or mentor. An AI that is developed with their input would extend the reach of those biases and preferences, while also lending credence and authority to the selected theories. Autism is again a useful example here, especially as dominant theories about traits and characteristics have already led to exclusionary diagnostic practices [41]. Yet many of those problematically exclusionary practices are already being replicated in AI [40].

Mitigation requires a number of factors. First, that expertise is sought in the development of an AI, but also that such expertise is handled critically by an interdisciplinary team with a breadth of knowledge. This range ought to be sufficient for both the identification of problems of negative bias early in the design and implementation stages, as well as to ensure that sufficient attention is given to whether theories on which AI programming rely avoid exclusionary practices. Transparency in such processes is also essential so that medical professionals who will be end users of an AI enabled technology can themselves identify biases in the programming. Without these elements, the quality of an AI will remain vulnerable to the replication of unchecked biases, including those that are dominant but not unproblematic and those that have form for historical inequalities, whether in terms of overrepresentation or exclusion, among others. Opportunities to change the programming of potentially expensive AI may be few once it has already been embedded, so understanding of these necessary limitations and risks are essential if the tools are to be used critically and applied cautiously.

5 AI and Medical Ethics

The previous two sections explored how AI is transforming medical epistemology and bringing particular value to the field but also challenges and risks. This changing epistemological foundation with potential threats is also having a significant impact on medical ethics. The last section of this chapter is therefore dedicated to exploring how AI is affecting medical ethics.

Medicine is fundamentally a relational practice founded on the relationship between a patient and

a doctor. A key aspect of this relation is its constitutive asymmetry – the vulnerability of a patient who asks for the help of a health professional equipped with the competence to respond to the identified vulnerability. However, this asymmetry of abilities, or power, may lead to abuses on either side: (a) the patient might abuse the professional's service or (b) the professional might abuse his or her power over the patient. As the philosopher Worms puts it: "Care cannot exist without a relationship in which the weakness of one person requires assistance, which can turn into submission, and the capability of another allows for devotion, which can turn into power, or even abuse of power" ([42] ii). Because of this constitutive asymmetry between the patient and the doctor, medical practice has been, from very early on, regulated by ethical codes and guidelines. The earliest of these is the Hippocratic Oath and many others have followed, notably the principles of biomedical ethics by Beauchamp and Childress [43].

How is the deployment of AI in medicine modifying this fundamental aspect of this practice as a relation, given that it is one that is essentially asymmetric? This is a key question for medical ethics in the era of AI, one that philosophers, ethicists, healthcare professionals, patients, and society at large need to explore to identify potential challenges and risks and propose mitigating strategies. Section 5 of this chapter explores this question through two different aspects: (a) the patient-doctor relationship and (b) the medical profession in the era of digital capitalism.

5.1 The Patient-Doctor Relationship

The fundamental relationship between the patient and the doctor is affected in various ways by the introduction of AI to medicine. There is the risk that the clinical encounter – this fundamental moment of the therapeutic relationship – is set aside, replaced by a mass amount of data automatically collected and analyzed. The claimed "'superhuman' accuracy and insight" [7] produced by AI risks replacing the value of this encounter. This is a potential threat for clinical diagnosis as an outcome from a doctor-patient encounter. Indeed, what causes a person to seek medical help may not be the primary issue for which they need most help. In other words, the narrow approaches to diagnostic processes as conducted by AI systems may lead to missing potential for incidental findings. For instance, referred pain, physical manifestations of mental health issues, hoarding and other self-harm behaviors can be indicators of a brain disorder, a mental health condition, or of abuse, such as domestic violence. These kinds of potentialities could be straightforwardly considered in an interpersonal clinical encounter, but missed by an AI.

Additionally, with the entrance of new, highly complex technical tools in medical practice, the doctor might not fully understand the results of the analysis produced by the AI, and therefore be unable to properly explain to the patient and their family the full rationale behind the diagnosis and prescription. As Campolo and Crawford put it, "claims about 'superhuman' accuracy and insight" are "paired with the inability to fully explain how these results are produced" ([7] ibid). This is a key challenge of explainability brought about by AI, also presented as the issue of an AI system as a black box [44]. This issue is particularly critical in the context of medicine since human life and wellbeing are at stake.

The clinical encounter between the patient and doctor is also affected by the introduction of new actors, i.e., data scientists and engineers, and their technical systems, especially where these gain a central role. This inclusion affects the intimate relationship that characterizes the clinical encounter and brings challenges in terms of confidentiality and privacy [45, 46]. For instance, a study has shown that "a well-trained deep learning system is able to recover the patient identity from chest X-ray data" [47]. Even if CheXNet, discussed above, were to become an effective prediction model, such revelations could undermine its usefulness in terms of patient willingness to undergo scanning. This also leads to challenges related to trust and informed consent: two essential aspects of the medical relationship. In turn, this threat to the patient-doctor dialogue and relation of trust

might make it difficult for the patient to make sense of their own illness or issue. Indeed, the interpersonal dialogue with the doctor plays a major role in this self-understanding.

5.2 The Medical Profession in the Era of Digital Capitalism

The entrance of AI in medicine is also impacting the medical profession in various ways. Data science is increasingly gaining a central role in the field, at the expense of more traditional medical expertise, including its intuitive dimensions. As Matuchansky puts it: "One quickly learns from clinical practice that medicine is as much an art as a science or a technique" [48]. With the entrance of AI in medicine, this should not be forgotten or obscured behind claims of precision, scientism, and objectivity of AI systems.

Connected to this challenge to the profession, there is also the risk of deskilling of medical professionals [49]. If AI systems are seen as better placed to conduct various activities that were initially conducted by a doctor, chances are high that they will progressively replace humans (hence, also redirecting money from clinician's wages to technology companies). For instance, use of deep learning techniques for chest radiography shows "the potential to exceed human performance" [47]. In turn, this leads to the risk of increased dependency on technology for key aspects of human existence (cf [49, 50]). Against this, many experts have called to ensure that AI remains an assistant to the healthcare professional, acting as a "support tool," and not as a replacement to the human [51].

Additionally, this trend further pushes medicine in the direction of a technical discipline, taking attention away from the relational aspects of this practice. As the ethics of care has shown, the relational aspects of care already tends to be undervalued in the medical sector. These non-technical tasks are primarily undertaken by people from marginalized groups, primarily women and migrants, and those who are poorly paid. This can be contrasted with those highly specialized and technical areas of medicine such as surgery, which are highly respected and valued, and rewarded accordingly [52]. The introduction of AI to the world of medical care further contributes to this over-valorization of the technical at the expense of relational aspects.

Exploring relations in medicine is not only restricted to interpersonal relationships (between patients, doctors, and other professionals). It also requires looking at relations within the broader sociopolitical landscape and the power asymmetries at this level. Here as well, the growing dependency on technology, and in particular on big technology companies needs to be investigated. The use of AI in the medical sector has brought about new actors, including Google, Apple, Facebook, and Amazon [53]. Because these companies are in possession of massive amounts of data and the ability to process them, they have found themselves well placed to enter the medical sector. For instance, Tamar Sharon talks about the "googlisation of healthcare" and its promises to "advance health research by providing the technological means for collecting, managing, and analysing the vast and heterogeneous types of data required for data-intensive personalised and precision medicine" [53]. This entrance of healthcare in "digital capitalism" is posing key questions in terms of privacy and confidentiality. Indeed, these big tech companies have a rather poor track record when it comes to the protection of personal data [50]. Medical data is enormously valuable and a particularly sensitive type of data that requires special protection. Access to such data only further increases the power of technology companies. Meanwhile, there is a general consensus that already powerful actors from insurance to recruitment companies should not have access to this data as it could lead to significant discrimination on the basis of health.

As the above has shown, the introduction of AI in medicine changes the power dynamics in the sector. With the digitalization of the field, data scientists, as well as big technology companies are gaining a central role in medicine. It is also pushing the sector toward "technicization" at the expense of relational aspects that are nonetheless central. Although AI in medicine brings great promises for the field in terms of efficiency and

precision, it is nonetheless essential to pay attention to these changing power dynamics as well as the technicization of the field to ensure that the interests and wellbeing of the patient remain the central consideration of medical practice. Additionally, it is essential to ensure that healthcare professionals, whether doctors or nurses, are well equipped for the digital transformation of their field and protected against abuses by industry actors. Finally, policy makers also have a role to play to ensure digital capitalism does not interfere with the interests of patients.

6 Conclusion and Recommendations

This chapter has investigated some of the promises for medicine in the algorithmic age and what these claims mean philosophically and ethically. After briefly exploring what AI promises to achieve in terms of medical advances, this chapter looked at how the introduction of AI in medicine deeply impacts medical epistemology. In particular, it explored the implications of a technology that claims to generate knowledge and diagnosis from data alone. Numerous questions raised by this epistemology, based on a form of "data-utopianism," and how this interrogates the nature of objectivity in the medical field have been pointed out. This chapter has then highlighted the limitations, risks, and biases of AI and how these impact medicine. Finally, it looked into the ethical implications of the introduction of AI to medicine, especially in relation to the patient-doctor relationship and how the medical profession is evolving as it enters the era of digital capitalism.

This chapter has highlighted some key challenges and risks brought about by AI in medicine, but also some ways to mitigate these. To begin with, it is essential to refer to relevant ethical guidelines and frameworks. A number of these have been developed for AI over the last few years, such as the 2019 "Ethics guidelines for trustworthy AI" of the High-Level Expert Group on Artificial Intelligence set up by the European Commission [54]. The SIENNA project has also developed a set of ethical instruments to promote an ethical development, deployment and use of AI, including an ethics by design framework and an approach for research ethics [55]. It is also important to carefully assess where AI can truly add value and where it does not (and might even be harmful). To these ends, ethical impact assessments as well as social science studies on the use of AI in medicine can help to better understand the impacts and consequences of AI for the medical sector, including in the short, medium, and long term [56].

Finally, it is essential for the healthcare community to develop their understanding of AI in technical terms. This would help to ensure appropriate and proportionate levels of trust, which is not too little, such that the AI system would become useless or poorly applied and adopted, nor too much, such that it is trusted and allowed to function independently of oversight and appropriate human intervention. It is also essential to raise awareness of the kinds of unique challenges that are brought by AI, including for doctors, for the healthcare community, and for society at large. Effective AI in medical contexts requires transparency so that people are aware of these tools in their work and their everyday lives, and so as to ensure a certain degree of oversight in these contexts. Finally, and as already mentioned above, AI should remain a tool that human beings can use if they deem it useful to achieve specific and identifiable objectives, but not as a replacement for human expertise and intervention, especially in medical domains where a person's needs and their vulnerabilities may be greatest.

References

1. Coeckelbergh M. AI ethics. The MIT press essential knowledge series. Cambridge, MA: The MIT Press; 2020.
2. Mittelstadt B. AI ethics – too principled to fail? SSRN Journal [Internet]. 2019 [cited 2021 Mar 22]. https://www.ssrn.com/abstract=3391293
3. Shaw JA, Sethi N, Block BL. Five things every clinician should know about AI ethics in intensive care. Intensive Care Med. 2021;47(2):157–9.
4. Schiff D, Borenstein J. How should clinicians communicate with patients about the roles of artificially

intelligent team members? AMA J Ethics. 2019;21(2): E138–45.
5. Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. Soc Sci Med. 2020;260:113172.
6. Kulkarni S, Seneviratne N, Baig MS, Khan AHA. Artificial intelligence in medicine: where are we now? Acad Radiol. 2020;27(1):62–70.
7. Campolo A, Crawford K. Enchanted determinism: power without responsibility in artificial intelligence. Engag Sci Technol Soc. 2020;6:1–19.
8. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired. 2008;16(07).
9. Good IJ. The philosophy of exploratory data analysis. Philos Sci. 1983;50(2):283–95.
10. Dinov ID. Volume and value of big healthcare data. J Med Stat Inf. 2016;4(1):3.
11. Butte AJ, Kohane IS. Unsupervised knowledge discovery in medical databases using relevance networks. In: Proc AMIA Symp. 1999;711–5.
12. van Dijck J. Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. Surveill Soc. 2014;12(2):197–208.
13. danah b, Crawford K. Critical questions for big data. Inf Commun Soc. 2012;15(5):662–79.
14. Rose N. 'Screen and intervene': governing risky brains. Hist Hum Sci. 2010;23(1):79–105.
15. Poste G. Bring on the biomarkers. Nature. 2011;469(7329):156–7.
16. Tversky A, Kahneman D. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol Rev. 1983;90(4):23.
17. Frankel RM, Stein T. Getting the most out of the clinical encounter: the four habits model. Perm J. 1999;3(3):79–88.
18. Kitchin R. Big data, new epistemologies and paradigm shifts. Big Data Soc. 2014;1(1):2053951714528481.
19. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. PLoS One. 2016;11(8):e0157077.
20. Gao C, Sun H, Wang T, Tang M, Bohnen NI, Müller MLTM, et al. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. Sci Rep. 2018;8(1):7129.
21. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics. 2016;22(2):303–41.
22. Marino S, Xu J, Zhao Y, Zhou N, Zhou Y, Dinov ID. Controlled feature selection and compressive big data analytics: applications to biomedical and health studies. PLoS One. 2018;13(8):e0202674.
23. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154(11):1247–8.
24. Bezemer T, de Groot MCH, Blasse E, ten Berg MJ, Kappen TH, Bredenoord AL, et al. A human(e) factor in clinical decision support systems. J Med Internet Res. 2019;21(3):e11732.
25. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care. AMA J Ethics. 2019;21(2):167–79.
26. Kuhn TS, Hacking I. The structure of scientific revolutions. 4th ed. Chicago/London: The University of Chicago Press; 2012. 217 p.
27. Becker D. Through the looking glass: women and borderline personality disorder [Internet]. 1st ed. Routledge; 2019 [cited 2021 Mar 23]. https://www.taylorfrancis.com/books/9780429964206
28. Metzl JM. The protest psychosis: how schizophrenia became a black disease [Internet]. Boston: Beacon Press; 2014 [cited 2021 Mar 23]. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=715745
29. Murray S. Corporeal knowledges and deviant bodies: perceiving the fat body. Soc Semiot. 2007;17(3):361–73.
30. van Ryn M, Burke J. The effect of patient race and socio-economic status on physicians' perceptions of patients. Soc Sci Med. 2000;50(6):813–28.
31. Olkin R, Hayward H, Abbene MS, VanHeel G. The experiences of microaggressions against women with visible and invisible disabilities. J Soc Issues. 2019;75(3):757–85.
32. Hamilton N, Olumolade O, Aittama M, Samoray O, Khan M, Wasserman JA, et al. Access barriers to healthcare for people living with disabilities. J Public Health (Berl) [Internet]. 2020 Oct 10 [cited 2021 Mar 23]. http://link.springer.com/10.1007/s10389-020-01383-z
33. Fricker M. Epistemic injustice: power and the ethics of knowing. Oxford/New York: Oxford University Press; 2007. 188 p.
34. Peled Y. Language barriers and epistemic injustice in healthcare settings. Bioethics. 2018;32(6):360–7.
35. Tasca C, Rapetti M, Carta MG, Fadda B. Women and hysteria in the history of mental health. CPEMH. 2012;8(1):110–9.
36. Foucault M. The birth of the clinic: an archaeology of medical perception. 1. publ., reprinted. London: Routledge; 2010. 266 p. (Routledge classics).
37. Vincent J. Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech [Internet]. The Verge. 2018 [cited 2021 Mar 24]. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai
38. Garcia M. Racist in the machine: the disturbing implications of algorithmic bias. World Policy J. 2016;33(4):111–7.
39. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356(6334):183–6.
40. Erden YJ, Hummerstone H, Rainey S. Automating autism assessment: what AI can bring to the diagnostic process. J Eval Clin Pract. 2020;27:485. https://doi.org/10.1111/jep.13527.

41. Bargiela S, Steward R, Mandy W. The experiences of late-diagnosed women with autism spectrum conditions: an investigation of the female autism phenotype. J Autism Dev Disord. 2016;46(10):3281–94.
42. Worms F. The two concepts of care. Life, medicine, and moral relations. Esprit. 2006;1:141.
43. Beauchamp TL, Childress JF. Principles of biomedical ethics. Oxford: Oxford University Press; 2012.
44. Castelvecchi D. Can we open the black box of AI? Nat News. 2016;538(7623):20.
45. Char DS, Shah NH, Magnus D. Implementing machine learning in health care – addressing ethical challenges. N Engl J Med. 2018;378(11):981–3.
46. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866–72.
47. Packhäuser K, Gündel S, Münster N, Syben C, Christlein V, Maier A. Is medical chest X-ray data anonymous? arXiv:210308562 [CS, EESS] [Internet]. 2021 Mar 15 [cited 2021 Mar 24]. http://arxiv.org/abs/2103.08562
48. Matuchansky C. Intelligence clinique et intelligence artificielle. Une question nuance med/sci. 2019;35:797–803.
49. Susskind RE, Susskind D. The future of the professions: how technology will transform the work of human experts. Oxford University Press; 2015.
50. Powles J, Hodson H. Google DeepMind and healthcare in an age of algorithms. Heal Technol. 2017;7(4):351–67.
51. Coeckelbergh M. Health care, capabilities, and AI assistive technologies. Ethical Theory Moral Pract. 2010;13:181–90.
52. Molinier P. De la civilisation du travail à la société du care. Vie sociale. 2016;14(2):127–40.
53. Sharon T. When digital health meets digital capitalism, how many common goods are at stake? Big Data & Society; 2018.
54. Hleg A. High-level expert group on artificial intelligence: ethics guidelines for trustworthy AI. European Commission, 0904; 2019.
55. Resseguier A, Brey P, Dainow B, Drozdzewska A, Santiago N, Wright D. D5.4: multi-stakeholder strategy and practical tools for ethical AI and robotics. SIENNA; 2021.
56. Resseguier A, Rodrigues R. Ethics as attention to context: recommendations for the ethics of artificial intelligence. Open Research Europe; 2021.