OPEN ACCESS

On the ethics of algorithmic decision-making in healthcare

Thomas Grote [1,2] Philipp Berens [3]

[1] Ethics and Philosophy Lab; Cluster of Excellence: "Machine Learning: New Perspectives for Science", University of Tübingen, Tübingen, Germany
[2] International Center for Ethics in the Sciences and Humanities (IZEW), University of Tübingen, Tübingen, Germany
[3] Institute for Ophthalmic Research, University of Tübingen, Tubingen, Germany

Correspondence to
Thomas Grote, Ethics and Philosophy Lab, Cluster of Excellence: "Machine Learning: New Perspectives for Science", University of Tübingen, Tübingen 72076, Germany; thomas.grote@uni-tuebingen.de

Received 30 June 2019
Revised 27 August 2019
Accepted 9 September 2019

Check for updates

© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Grote T, Berens P. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/medethics-2019-105586

ABSTRACT

In recent years, a plethora of high-profile scientific publications has been reporting about machine learning algorithms outperforming clinicians in medical diagnosis or treatment recommendations. This has spiked interest in deploying relevant algorithms with the aim of enhancing decision-making in healthcare. In this paper, we argue that instead of straightforwardly enhancing the decision-making capabilities of clinicians and healthcare institutions, deploying machines learning algorithms entails trade-offs at the epistemic and the normative level. Whereas involving machine learning might improve the accuracy of medical diagnosis, it comes at the expense of opacity when trying to assess the reliability of given diagnosis. Drawing on literature in social epistemology and moral responsibility, we argue that the uncertainty in question potentially undermines the epistemic authority of clinicians. Furthermore, we elucidate potential pitfalls of involving machine learning in healthcare with respect to paternalism, moral responsibility and fairness. At last, we discuss how the deployment of machine learning algorithms might shift the evidentiary norms of medical diagnosis. In this regard, we hope to lay the grounds for further ethical reflection of the opportunities and pitfalls of machine learning for enhancing decision-making in healthcare.

INTRODUCTION

Machine learning is increasingly being conceived as a technology with the potential to transform professional healthcare. Recently, there has been a surge of interest in machine learning for medical decision-making (reviewed by Esteva and Topol[1 2]), fuelled by a series of studies demonstrating 'expert-level' accuracy of machine learning algorithms, for example, in diagnosing eye diseases from fundus images,[3] and different types of skin cancer from images of skin lesions.[4] Moreover, a study made by Walsh and colleagues found that machine learning algorithms managed to predict the risk of imminent suicide attempts at high accuracy based on a large repository of clinical electronic health data (Walsh, p. 460).[5] In contrast, for clinicians, the ability to predict suicide attempts has been near chance for decades. Hence, machine learning algorithms promise to enhance the diagnostic as well as the predictive abilities of clinicians by assessing health risks of individual patients based on complex diagnostic data sets. Furthermore, the predictive abilities of machine learning algorithms might amplify an ongoing shift in healthcare, from curing disease towards prevention.[6]

Thus, high hopes are being put into machine learning making healthcare *smarter*. By examining the literature on machine learning in healthcare,[2 7] one typically encounters a type of narrative, brought up in favour of deploying said algorithms in healthcare. It applies to both the level of individual clinicians as well as the institutional level of healthcare. Regarding the individual level, it is being argued that deploying machine learning algorithms will improve medical decision-making—meaning it will make medical diagnosis and treatment decisions quicker and more reliable. Besides, proponents of machine learning in medicine are usually not shy of pointing out flaws of clinicians, such as their susceptibility to cognitive biases and to committing diagnostic errors (Topol, p. 52).[2] Therefore, machine learning algorithms might compensate for the weaknesses or even enhance the decision-making capabilities of individual clinicians. With respect to the institutional level of healthcare, it is inefficiencies in the workflow, a potential waste of resources, inequities and exploding costs which are being referred to here. Again, machine learning is assumed to mitigate these deplorable circumstances.[2]

As will be shown in this paper, this narrative relies on shaky assumptions. More importantly, one of our central claims is that even if we accept the narrative's individual premises, the enhancement of clinicians and healthcare institutions by means of machine learning is less straightforward than it might appear. As we aim to demonstrate, the deployment of machine learning algorithms in medicine goes hand in hand with trade-offs on the epistemic and the normative level. Moreover, these trade-offs might bring about a plethora of ethically non-beneficial effects. Drawing on work from social epistemology, we argue that the involvement of current machine learning algorithms challenges the epistemic authority of clinicians. This promotes patterns of defensive decision-making which might come at the harm of patients. Additionally, we argue that, on a structural level, machine learning algorithms can exert normative force regarding the evidential standards and the management of risks within healthcare institutions. Again, this raises new questions with respect to fairness in healthcare.

The paper's structure will be as follows: first, we give a brief outline regarding the current state of machine learning in healthcare. This is required to underpin our epistemological and ethical arguments. In the subsequent section, we examine some epistemic problems that arise, once clinicians try to make well-informed decisions based on a machine learning algorithm's output. Building on this, we will discuss how these epistemic problems are intimately linked to a broader range of ethical issues, such as problems of defensive medicine and patient autonomy. In the final section, we will focus on ethical issues stemming from the deployment of

J Med Ethics: first published as 10.1136/medethics-2019-105586 on 20 November 2019. Downloaded from http://jme.bmj.com/ on August 13, 2024 by guest. Protected by copyright.

machine learning algorithms at the structural level of healthcare. Here, we will mostly discuss how the involvement of machine learning in healthcare challenges the norms of accountability in addition to reshaping epistemic norms of healthcare. Furthermore, we will highlight some issues regarding fairness and the management of risks in healthcare which again emerge from the aforementioned epistemic problems.

PRIMER ON MACHINE LEARNING IN HEALTHCARE

Recent years have seen a rapid surge of interest in the applications of machine learning algorithms in medicine.[1 2] In classical artificial intelligence, expert systems contain a database of deductive rules by which—given a set of known facts—certain consequences can be inferred. In medicine, for example, the presence of certain symptoms may trigger the expert system to propose diseases commensurate with these symptoms and suggest diagnostic tests to reach unambiguous conclusions. Such systems are typically based on the explicitly encoded knowledge of experts. For example, the recently discussed IBM Watson Oncology system could be classified as an advanced expert system combining automatic text mining of clinical papers with a vast number of logical rules.[8] In imaging-based medical diagnostics, an expert system may look for expert-defined features and explicitly encode the decision rules as stated by clinicians.

In contrast to these rule-based algorithms explicitly aiming at encoding expert knowledge,[i] machine learning algorithms aim at extracting patterns or structure present implicitly in labelled examples. In medical applications, such examples often consist of images (eg, X-rays or fundus photographs). Further, labels may consist of a disease rating according to a diagnostic scale. The algorithm then automatically infers which setting for its internal parameters and which features lead to the most accurate predictions of the original labels. The art and science of machine learning consists in finding classes of algorithms that show a high generalisation performance to new data sets not seen during training. This brief section does not aim at covering the whole spectrum of applications of machine learning in healthcare but will focus on research in machine learning with respect to medical diagnosis. The recent interest in machine learning is fuelled by two ingredients: On the one hand, clinical data are increasingly collected digitally, making them amenable to analysis by machine learning algorithms. In addition, a special kind of machine learning algorithm, called deep neural networks, has rapidly gained popularity. These deep neural networks excel in recognising objects in images, a task that has been the cornerstone of computer vision research for decades.[9] Today, deep neural networks can label objects in images even more accurately than humans. As an algorithm, deep neural networks consist of layers of nodes that each use simple mathematical operations to perform a specific operation on the activation of the layer before, leading to the emergence of increasingly abstract representations of the input image. As deep neural networks have a very large number of parameters, very large data sets are typically required to achieve good performance. Alternatively, techniques that constrain the complexity of the mathematical function at each node can be used (so called 'regularisation'). Naturally, deep neural networks lend themselves to straightforward application in imaging-based medical specialties such as radiology, ophthalmology, dermatology or pathology. Several studies have shown by now that deep neural networks can match, or even surpass, medical experts in diagnostic accuracy, at least when tasked with classifying individual diseases from a given diagnostic modality.[2 3]

Importantly, algorithmic decisions can often be made more quickly than human ones, arguing that machine learning–based approaches may be particularly useful in emergency settings. Notably, progress has been made recently towards diagnosing any of a whole set of diseases from diagnostic images.[4 10] For example, the algorithm by de Fauw *et al* can assess diagnostic images of the retina for a variety of 50 different retinal diseases and suggest which patient needs urgent attention by trained medical personnel at high accuracy.[10] Despite these successes, transfer to clinical practice has been slow, partially due to difficult regulatory questions and the lack of rigorous prospective studies showing improved clinical outcomes.[2] Major hurdles for adopting machine learning–based diagnostic systems in clinical workflows are the lack of transparency of such systems due to their complex architecture (despite some progress, see de Fauw *et al* and Norgeot *et al*[7 10]) and the fact that machine learning–based algorithms tend to lack an accurate report of confidence in their decision (typically being overconfident). While machine learning algorithms will not cure any disease by themselves any time soon, there is clear potential to improve diagnostic decision-making based on the progress we are seeing today.

PITFALLS OF ALGORITHMIC DECISION-MAKING AT THE INDIVIDUAL LEVEL OF HEALTHCARE

Epistemic pitfalls

In medicine, making 'good decisions' constitutes much of the daily work of clinicians. They need to accurately diagnose diseases based on limited evidence, and in limited time, and determine the best treatment strategy among different possibilities for the patient at hand. In these tasks, clinicians are highly skilled experts. They have undergone multiple years of training and, throughout their careers, many have diagnosed and treated a five-digit number of patients. The gold standard for medical diagnosis is a complex process which involves multiple steps. According to a report from the National Academy of Science from 2015 on improving medical diagnosis, its basic structure might be described as follows:

> Once a patient seeks health care, there is an iterative process of information gathering, information integration and interpretation, and determining a working diagnosis. Performing a clinical history and interview, conducting a physical exam, performing diagnostic testing, and referring or consulting with other clinicians are all ways of accumulating information that may be relevant to understanding a patient's health problem. […] The continuous process of information gathering, integration, and interpretation involves hypothesis generation and updating prior probabilities as more information is learned.[11] (p. 32)

Moreover, clinicians deploy different diagnostic tools, such as medical imaging devices, allowing to assess physical conditions in high anatomical detail. At last, medical diagnosis is often a collaborative endeavour, as the patient herself and other clinical colleagues are being consulted.

This outline of the process of medical diagnosis also highlights some of the shortcomings of many of the studies comparing the performance of machine learning algorithms with clinicians. For instance, in the study from Esteva and colleagues,[4] which compared the ability of machine learning algorithms and dermatologists at classifying skin cancer, the dermatologists had to judge based on a slide of clinical images in a relatively short time

[i] Of course, a machine learning-based component could be part of an expert system, such that the two are not as such ruling each other out.

J Med Ethics: first published as 10.1136/medethics-2019-105586 on 20 November 2019. Downloaded from http://jme.bmj.com/ on August 13, 2024 by guest. Protected by copyright.

span. In more realistic settings, it might be fair to assume that the dermatologist would consider different sources of evidence and her diagnosis would not be a one-off decision.

Having said this, diagnostic errors are still an all too common phenomenon in healthcare. An estimate of the National Academy of Sciences states that around 5% of US adults seeking healthcare advice are subject to diagnostic error. Furthermore, relevant diagnostic errors are assumed to contribute to approximately 10% of all patient deaths[11] (p. 1). Diagnostic errors can be traced back to different causes. For instance, medical diagnosis necessarily involves different degrees of uncertainty. For a clinician, making medical diagnosis involves testing different hypothesis and therefore full certitude is unattainable. Moreover, different information gathering activities and medical treatments induce risks of their own for the patient's well-being, which need to be considered. In addition, clinicians often face time constraints. Some diseases require immediate treatment and thus, there might be limited time to properly assess all the evidence available[11] (p. 48).

Given these epistemic and structural constraints, it is easy to see how machine learning algorithms might enhance the decision-making capabilities of clinicians. When making a diagnosis, the algorithm can process complex sets of data in shorter time and—at least in theory—it might be less susceptible to cognitive biases than its human counterpart. Hence, the algorithm might provide an additional source of evidence for the clinician, allowing her to make a well-informed decision. Nevertheless, clinicians face obstacles when trying to infer information from a machine learning algorithm's output. Here, the underlying problem can be described as follows: both the clinician and the machine learning algorithm might be conceived as experts of sorts. Yet, they have been trained differently and they reason in very distinct ways. For the clinician, this poses a problem once we consider cases of peer-disagreement.[12 13] Here, 'peer-disagreement' describes cases of two (equally) competent peers with respect to a certain domain-related activity, whereby both parties disagree with respect to a certain proposition.[14]

Allow us to illustrate this problem by discussing a potential case of peer-disagreement between a clinician and a machine learning algorithm. Let us assume that she aims at making a medical diagnosis with respect to a skin disease. After assessing the evidence, she concludes that the patient has disease *x*, where she has a confidence of 0.8 in her proposition. However, when the machine learning algorithm screens the evidence, it states that the patient has disease *y*, with a similar degree of confidence. Now, when trying to make a well-informed decision, how much weight should the clinician assign to the algorithm's diagnosis? Bluntly put, should she be required to call her superior out of bed for an additional opinion? Or, would the superior be rightfully mad, given that the algorithm provided a clear diagnosis?

There is very little that the clinician might do on epistemic grounds to resolve the disagreement in question. For once, the algorithm might represent its output in form of a CI or a risk score. However, current algorithms often do not supply the clinician with an explanation of why it decided that way. This problem is intimately linked to the 'opacity' of machine learning algorithms—as opposed to transparent or interpretable machine learning algorithms. Based on a conceptual framework introduced by Burrell,[15] we can distinguish between three kinds of opacity: (1) opacity as corporate secrecy, (2) opacity as technical illiteracy and (3) opacity arising from the algorithm's complex form of mathematical representation which is not intelligible for humans. In principle, the first two kinds of opacity might be overcome by establishing regulatory standards or by better educating clinicians. In contrast, the third kind of opacity is an *intrinsic* problem of machine learning. It arises from architectural features of machine learning algorithms, especially deep neural networks, which are among the leading strands of research in machine learning. Here, some form of visualisation indicating where in the input image important features for the diagnosis can be found is currently the best one can achieve.[16]

In addition, the confidence score reported by the algorithm and that of a clinician may be given on the same scale (0: no confidence; 1: high confidence), but mean very different things: Many complex machine learning algorithms such as deep neural networks have been reported to be overconfident, and human self-reported confidence can be quite substantially differ from mathematical notions of confidence.[17] If the confidence reports of algorithms and humans were both accurate and well calibrated, the optimal decision rule for the clinician would be clear—as it stands, when making her diagnosis she basically has two options. She may stick to her initial proposition or she may defer to the algorithm. Fusing the two decisions is almost impossible, as the confidence judgments provide no path forward, and the algorithmic evidence cannot easily be accessed to change her mind.

Now, in the relevant philosophical debate, there are different theories about what would be reasonable for the clinician to do. According to the 'Equal Weight View',[12] learning that an epistemic peer's proposition differs from your own should diminish the confidence in one's judgment. Hence, deferring to the algorithm is the most reasonable choice. By contrast, the 'Steadfast View' emphasises the epistemically privileged status of one's own beliefs, which is why it is reasonable for the clinician to stick to her proposition.[18] Therefore, we end up with a stalemate.

We do not intend to settle a complex debate in social epistemology. It might be argued, however, that—given that the algorithm is likely trained and validated on the opinions of several expert clinicians—deferring would seem like a reasonable choice, especially for a novice. Nevertheless, no algorithm will every come completely foolproof. What should have become clear is that for a clinician trying to make a well-informed decision, the deployment of machine learning algorithms poses another source of uncertainty, which needs to be considered. In effect, this challenges the 'epistemic authority' of the clinician.

Some brief remarks concerning the notion of 'epistemic authority'. Drawing on ideas by Fricker, we reject the idea that a clinician should rely only on her own cognitive faculties.[19] Our modern society has been shaped by an extensive division of cognitive labour. Further, as a matter of course, we deploy technology in order to access information in virtually every domain of life. Therefore, deferring to the testimony of others can be reasonable, if certain epistemic conditions are being met. In this regard, Fricker developed a general principle specifying when deferring to the testimony of others should be accepted on the basis of trust. Here, the necessary and sufficient conditions comprise two factors. First, the testimonial source needs to be epistemically in a good enough position with respect to *p*, ensuring that *p* almost certainly qualifies as knowledge. Furthermore, the testifier's epistemic position needs to be better than the expert's herself. Second, the expert needs to recognise the testifier's superior epistemic position in addition to not being aware of any contrary evidence regarding (Fricker, p. 232).[19]

Building on this, we might be able to formulate the challenge of machine learning algorithms to the epistemic authority of the clinician more clearly. Given the opacity and the overconfidence of machine learning algorithms, assessing their epistemic position is currently not feasible. Hence, the decision to

J Med Ethics: first published as 10.1136/medethics-2019-105586 on 20 November 2019. Downloaded from http://jme.bmj.com/ on August 13, 2024 by guest. Protected by copyright.

defer a medical diagnosis to said algorithms lacks proper epistemic support. However, for reasons stated above (eg, machine learning algorithms being trained on many expert's data), it is still compelling to defer to it. By contrast, there are more reliable indicators for assessing the testimony of fellow clinicians. Let us assume that the clinician is a novice, whereas the testimonial source happens to be an acclaimed specialist in the relevant field. Here, she might be justified in concluding that the testimonial source is in an epistemic privileged situation.

Moreover, according to the Argumentative Theory of Reasoning by Mercier and Sperber,[16] dialogical engagement with other colleagues has a high chance of producing better epistemic outcomes. This is especially the case, when their views differ. As solitary reasoners, people are bad at evaluating their own arguments. Further, they are prone to cognitive biases such as overconfidence. That being said, they excel at spotting reasoning errors in the arguments of their interlocutors. Thus, by being confronted with the arguments of a fellow colleague, the clinician is likely to end up with more reliable propositions (cf. Mercier and Sperber[16]: p. 222). Going back to the interplay of clinicians and machine learning algorithms, it might be fair to say that while such an interaction is currently not feasible, it could perhaps be a conceivable model in the future.

Ethical pitfalls

What might seem to some as being a mostly an epistemological problem has, in fact, severe ethical implications. For a start, clinicians are being held accountable for their decisions. They are obliged to provide justification for their actions. Furthermore, in case a clinician causes harm to a patient by committing a severe diagnostic error, she might be blamed for acting irresponsibly. To mitigate that risks, one should decide according to the best evidence available. Now, let us return to the case of peer disagreement between the clinician and a machine learning algorithm. She knows that her and the algorithm's diagnosis diverge. Yet, she is unable to extract an explanation from the algorithm why it decided accordingly. At best, she might have some higher-order evidence about the algorithm's general degree of accuracy or a map of important regions in the original image. If we assume that the relevant general degree of accuracy is reasonably high, it is easy to see why it is tempting for her to defer to the algorithm.

For one, deferring to the algorithm provides her with a normative justification for her medical decision. Then again, if she sticks to her initial proposition—and her diagnosis turns out to be wrong—she might be conceived as acting irresponsibly as she ignored the evidence provided by the algorithm. Things might aggravate once we consider that many medical decisions are being made under imperfect conditions, such as time constraints obstructing a careful re-assessment of the evidence available. A further side effect could be that the clinician might be biased towards interpreting the evidence in a way confirming the algorithm's diagnosis. Thus, the interplay of machine learning algorithms and clinicians potentially risks the fostering of epistemic vices such as dogmatism or gullibility.[20] To sum things up, instead of enhancing their decision-making capabilities, the deployment of machine learning algorithms may impose mechanisms of 'defensive medicine' among clinicians.[21 22]

However, there is one caveat. It needs to be pointed out that the opacity of machine learning algorithms is a well-established problem within the machine learning community. Currently, research groups from the tech giants and the academic sector alike are working on solutions to make machine learning algorithms explainable.[23] Thus, it might be the case that some of the problems discussed above will need to be revised in the foreseeable future. Then again, some more fundamental questions with respect to explainable machine learning are likely to remain. First, explainable to whom? To a data scientist, the clinician or the patient? In each case, relevant explanation requires some trade-offs. Either the bar might be set too high for some stakeholders or the explanation might become too simplified, omitting meaningful information.

This also leads us to problems of patient autonomy. In a recent article, McDougall has argued that involving artificial intelligence in treatment decisions risks reintroducing a paternalistic model of medical decision-making—in the guise of a 'computer knows best'—attitude. According to her line of reasoning, algorithms enforce such a paternalistic model by dictating the values, by virtue different treatment options are being ranked. For instance, most algorithms will rank treatment decisions based on which treatment maximises the lifespan of a patient the most. However, a patient might prefer a treatment which minimises her suffering instead. Thus, the involvement of artificial intelligence potentially undermines a shared decision-making between the clinician and the patient, posing a threat to the autonomy and dignity of the patient.[24]

We believe this analysis to be basically correct (for a more critical view, cf. Di Nucci[25]). Nevertheless, based on our discussion of the epistemic trade-offs arising from the deployment of machine learning algorithms, its challenges to shared decision-making and patient autonomy might be even more severe. Deploying machine learning algorithms for medical diagnosis or treatment decisions might work for the good of the patient, as they allow for more accurate medical diagnosis. Yet, due to the opacity of relevant algorithms, sensitive information will be withheld from the patient. As the patient is not provided with sufficient information concerning the confidence of a given diagnosis or the rationale of a treatment prediction, she might not be well equipped to give her consent to treatment decisions. Again, while machine learning algorithms might become explainable in the future, relevant standards that appeal to the epistemic norms of informed consent still need to be established. For instance, which information from the algorithm's statistical model shall be regarded as being essential for the patient to make an informed decision? Conversely, which information can be withheld? These epistemic issues are particularly pressing, if informed consent's normative core (ie, protecting a patient's autonomy and dignity) is to be kept intact. Instead, without providing adequate information, it runs at risk of eroding into a merely formalistic and legal form of protection.[26 27]

PITFALLS OF ALGORITHMIC DECISION-MAKING AT THE STRUCTURAL LEVEL OF HEALTHCARE

Many of the ethical issues arising at the institutional level of healthcare are intimately linked to the problems of uncertainty and accountability discussed in the last section. Nevertheless, it is the institutional level where the ramifications of machine learning algorithms in decision-making processes become most apparent. At the individual level, the deployment of machine learning algorithms potentially fosters epistemic vices among clinicians as they are inclined to minimise their risks of being blamed for medical maltreatment. However, who is to blame in case an algorithm turns out to be flawed and systematically prescribing erroneous treatments? For instance, many of IBM Watson Health's cancer algorithm treatment recommendations have turned out to be erroneous and thereby induced iatrogenic risk to patients. In case of Watson, this problem can be traced back to relevant algorithms mostly being trained with a small

J Med Ethics: first published as 10.1136/medethics-2019-105586 on 20 November 2019. Downloaded from http://jme.bmj.com/ on August 13, 2024 by guest. Protected by copyright.

number of synthetic non-real cases, compared with a large set of real data stemming from oncologists (Topol,[2] p. 51).

From an ethical perspective, this example highlights three pitfalls posed by involving machine learning in medicine. For a start, it complicates the attribution of accountability. If we assume that Watson gave a wrong treatment recommendation, yet, it was the clinician who made the final decision, can responsibilities be delegated to the tech firm engineering the algorithm? The picture becomes even more complex once we take other stakeholders into account, such as the healthcare institution pressuring the clinician to base her medical decisions on the algorithm's output. Hence, we are in a situation where each of the stakeholders involved have contributed to medical maltreatment, with neither of them being fully to blame. The obscuration of accountability just described has its counterparts in many other domains where artificial intelligence is being deployed—most notably self-driving vehicles. By looking at the ethical literature in relevant debates, one take-home message might be that we might be required to implement less individualistic notions of responsibility, such as distributed or collective responsibility, to close potential 'responsibility gaps'.[28–30] Having said this, it remains unclear how these less individualistic notions of responsibility might translate into the legal system.

Irrespective of the attribution of accountability, the example of Watson making wrong treatment recommendations also illustrates that by deploying machine learning algorithms in medicine, we necessarily expose patients to risks. A certain degree of uncertainty about the algorithm's reliability is inevitable. Here, we do employ a rather non-technical notion of 'risk', meaning situations where there is a possibility of an undesirable event affecting one's functionings of health.[31 32] The problem, however, is that these risks are unlikely to be distributed fairly among society. Instead, they are likely to disproportionately affect people belonging to certain subpopulations, such as racial minorities. This claim requires some explanation. For a start, while an algorithm might have a high degree of accuracy overall, it still might fare worse with respect to certain subpopulations. For example, if the training and test data for the algorithm during development predominantly involves medical cases of middle-aged Westerners, its medical diagnosis might be less accurate regarding people with East Asian ancestry. Due to different dietary habits or genetic predispositions, the latter might be prone to developing other kinds of diseases compared with the former. Thus, the machine learning algorithm might face difficulties at providing a reliable medical diagnosis for a patient coming from Eastern Asia.

Moreover, studies indicate that image recognition software is prone to develop biases which put people from certain racial minorities at a disadvantage. For instance, an algorithm might perform worse at accurately detecting a skin disease in a darker-skinned person, as compared with persons having a lighter skin colour. Again, this fact can be attributed to machine learning algorithms often being trained with data sets not sufficiently diverse. This problem might be accommodated by training relevant algorithms with more diverse sets of data. Nevertheless, as there are very strict legal regulations in many countries regarding the protection of health data, collecting a well-balanced data set may not always be feasible. Furthermore, given the risks that, for example, a person's genetic data becoming public poses for her privacy, there are also good reasons why health data should not be shared lightly.[33]

Hence, measures need to be taken to ensure a fair distribution of health risks among populations stemming from the involvement of machine learning in medicine. From a technical perspective, this might be achieved by validating the algorithm for different subpopulations. Nevertheless, on egalitarian grounds, further compensation for disadvantaged subpopulations might be necessary. For instance, a second opinion from clinical colleagues could become mandatory to mitigate their risks of becoming subject to medical malpractice. Apart from issues of fairness, the deployment of machine learning raises some fundamental questions related to the role of evidence and the management of risks in healthcare. Even in the absence of machine learning, uncertainty seems to be an integral part of decision-making in medicine.[34] When prescribing medical treatment, clinicians often do not know *why* a certain drug works, as they lack an explanation of its biological mechanisms. Thus, if we have good enough reasons to believe that involving algorithms in medicine promotes more reliable decision-making, does that justify their deployment on consequentialist grounds? In this vein, how should we balance the values of transparency and evidence on the one hand, and reliability and efficacy on the other[30] (for a more sceptical view cf. Stegenga[35])?

The third ethical pitfall might be defined as the problem of 'normative alignment'. The basic idea is that by deploying machine learning algorithms to make well-informed decisions, a relevant algorithm reshapes the epistemic norms within a healthcare institution. Our conjecture is that the relevant alignment of norms is being constituted by means of three different mechanisms:

1. *Deference:* By deferring to a machine learning algorithm, its diagnostic or treatment decisions become causally effective.
2. *Shared epistemic background:* In order to draw meaningful inferences from the machine learning algorithm's output, a clinician needs to align her conceptual standards to the algorithm.
3. *Standardisation of medical data:* As a prerequisite for getting robust data sets, given data need to be labelled according to a common standard. By incorporating machine learning algorithms in healthcare, normative standards about how to define a disease will be established.

Allow us to illustrate the latter claim. In the philosophy of medicine, the status of concepts such as 'health' and 'disease' is heavily contested. Here, we can mostly distinguish between two camps, 'naturalists' and 'normativists'. On the one hand, naturalists assume that these concepts are value-free representations of the world. For instance, according to Christopher Boorse's influential account,[36] disease might be conceived as a biological dysfunction. On the other hand, normativists assume that disease is a value-laden concept, mostly employed by practical purposes, such as deciding who should get medical treatment.[37 38] As treatment decisions take place in a social context and are often made under uncertainty, it might be fair to assume that they hinge to some degree on value judgments made by clinicians. In particular, these value judgments refer to the definition of disease and the rationale of the treatment. Hence, in applied contexts at least, normativists should have the upper hand with respect to their definition of disease.

Now, in case medical diagnosis or treatment recommendations are being deferred to machine learning algorithms, it is the algorithm who sets the bar about how a disease is being defined. Here, its judgments of value will be implicit in the training data, which shapes the algorithm's conception of given disease. The training data might stem from other health institutions where different values prevail. For example, the conceptual norms of certain mental diseases might differ vastly between hospitals, countries or even world regions. Therefore, by deploying machine learning algorithms, the values of health institutions

J Med Ethics: first published as 10.1136/medethics-2019-105586 on 20 November 2019. Downloaded from http://jme.bmj.com/ on August 13, 2024 by guest. Protected by copyright.

might shift, as conceptual norms stemming from other health institutions are being imposed.

What makes this shift so intricate is that the algorithm will not make its conceptual norms explicit, which makes them difficult to detect for clinicians. Finally, the deployment of machine learning in medicine might resurge the debate between naturalists and normativists. By means of machine learning, it might become feasible to analyse sufficiently large fractions of data from certain populations. Now, if the algorithm provides us with a comprehensive model of the statistical reference classes of a given population, to what degree does that allow us to naturalise certain types of health and disease? If so, would this ensure a more reliable decision-making in medicine?

Driven by the aim to mitigate inefficiency and uncertainty in healthcare, there are some intriguing parallels between advocates of machine learning and the evidence-based medicine (EBM). In both cases, the preponderant attitude might be conceived as the attempt to improve medical decision-making by standardising the norms of evidence (see Timmermans and Angell[39]). In EBM, the standards are set by a hierarchy of different forms of scientific research. By contrast, in machine learning it is the algorithm who provides guidance for clinicians by conducting diagnostic tasks. However, in both cases, clinicians are confronted with puzzles of normative epistemology. Which inferences are warranted given the data or the algorithm's output, respectively to what standard should the evidence in question be held? And what are its limits? However, there are also some differences between the two camps. Most notably, EBM is a movement within healthcare, whereas in machine learning the tech industry and leading computer science departments are among the major drivers. Suffice it to say at this point that the engagement of the industry entails ethical problems of its own.

CONCLUSION

In this paper, we aimed at examining which opportunities and pitfalls machine learning potentially provides to enhance of medical decision-making on epistemic and ethical grounds. As should have become clear, enhancing medical decision-making by deferring to machine learning algorithms requires trade-offs at different levels. Clinicians, or their respective healthcare institutions, are facing a dilemma: while there is plenty of evidence of machine learning algorithms outsmarting their human counterparts, their deployment comes at the costs of high degrees of uncertainty. On epistemic grounds, relevant uncertainty promotes risk-averse decision-making among clinicians, which then might lead to impoverished medical diagnosis. From an ethical perspective, deferring to machine learning algorithms blurs the attribution of accountability and imposes health risks to patients. Furthermore, the deployment of machine learning might also foster a shift of norms within healthcare. It needs to be pointed out, however, that none of the issues we discussed presents a knockout argument against deploying machine learning in medicine, and our article is not intended this way at all. On the contrary, we are convinced that machine learning provides plenty of opportunities to enhance decision-making in medicine. Medical decision-making involves high degrees of uncertainty and clinicians are prone to reasoning errors. In this respect, the involvement of machine learning algorithms in medical decision-making might yield better outcomes. However, it needs to be accompanied by ethical reflection. In this regard, we hope that the article lays the ground for further debate.

Contributors Thomas Grote is the main author of the article. Philipp Berens did write the second section and did comment on the paper's other parts.

Funding This research was supported by German Research Foundation (DFG, Excellence Cluster 'Machine Learning—New Perspectives for Sciences', EXC 2064, project no. 390727645; BE5601/4-1). Philipp Berens was additionally supported by the German Ministry for Education and Research (BMBF, 01GQ1601; 01IS18039A).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD
Thomas Grote http://orcid.org/0000-0002-9832-6046

REFERENCES

1 Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med* 2019;25(1):24–9.
2 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44–56.
3 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316(22):2402–10.
4 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115–8.
5 Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017;5(3):457–69.
6 Broadbent A. *Philosophy of epidemiology*. Palgrave McMillan, 2013.
7 Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. *Nat Med* 2019;25(1):14–15.
8 Ross C, Swelitz I. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. Available: https://www.statnews.com/2017/09/05/watson-ibm-cancer [Accessed 20 Aug 2019].
9 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–44.
10 De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24(9):1342–50.
11 National Academies of Sciences, Engineering, and Medicine. *Improving diagnosis in health care*. Washington DC: The National Academies Press, 2015.
12 Christensen D. Epistemology of disagreement: the good news. *Philos Rev* 2007;116(2):187–217.
13 Kelly T. Peer disagreement and higher order evidence. In: Goldman AI, Whitcomb D, eds. *Social epistemology: essential readings*. Oxford University Press, 2010: 183–2017.
14 Frances B, Matheson J. Disagreement. In: Zalta EN, ed. *The Stanford encyclopedia of philosophy*. Spring, 2018. https://plato.stanford.edu/archives/spr2018/entries/disagreement/
15 Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data Soc* 2016;3(1).
16 Mercier H, Sperber D. *Enigma of reason*. Harvard University Press, 2017.
17 Gigerenzer G, Hoffrage U, Kleinbölting H. Probabilistic mental models: a Brunswikian theory of confidence. *Psychol Rev* 1991;98(4):506–28.
18 Enoch D. Not just a truthometer: taking oneself seriously (but not too seriously) in cases of peer disagreement. *Mind* 2010;119(476):953–97.
19 Fricker E. Testimony and epistemic autonomy. In: Lackey J, Goldmann A, eds. *The epistemology of testimony*. Oxford University Press, 2006: 225–53.
20 Cassam Q. *Vices of the mind: from the intellectual to the political*. Oxford University Press: Oxford, 2019.
21 Gigerenzer G, Gaissmaier W, Kurz-Milcke E, *et al*. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest* 2007;8(2):53–96.
22 Hawley K. Trust and distrust between patient and doctor. *J Eval Clin Pract* 2015;21(5):798–801.
23 Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. *Digit Signal Process* 2018;73:1–15.
24 McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. *J Med Ethics* 2019;45(3):156–60.
25 Di Nucci E. Who's afraid of Dr. Watson? On the supposed threat of medical AI. *J Med Ethics*.
26 Eyal N. Informed consent. In: Zalta EN, ed. *The Stanford encyclopedia of philosophy*. Spring, 2019. https://plato.stanford.edu/archives/spr2019/entries/informed-consent/
27 Gould CC. How democracy can inform consent: cases of the Internet and bioethics. *J Appl Philos* 2019;36(2):173–91.
28 Floridi L. Distributed morality in an information society. *Sci Eng Ethics* 2013;19(3):727–43.

J Med Ethics: first published as 10.1136/medethics-2019-105586 on 20 November 2019. Downloaded from http://jme.bmj.com/ on August 13, 2024 by guest. Protected by copyright.

29 Matthias A. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf Technol* 2004;6(3):175–83.

30 Nyholm S. The ethics of crashes with self-driving cars: a roadmap, II. *Philos Compass* 2018;13(7).

31 Wolff J. Disadvantage, risk and the social determinants of health. *Public Health Ethics* 2009;2(3):214–23.

32 Wolff J, Avner D-S. *Disadvantage*. Oxford: Oxford University Press, 2007.

33 Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci Eng Ethics* 2016;22(2):303–41.

34 London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep* 2019;49(1):15–21.

35 Stegenga J. *Medical nihilism*. Oxford: Oxford University Press, 2018.

36 Boorse C. Health as a theoretical concept. *Philos Sci* 1977;44(4):542–73.

37 Clouser KD, Culver CM, Gert B. Malady: a new treatment of disease. *Hastings Cent Rep* 1981;11(3):29–37.

38 Kingma E. Naturalism about health and disease: adding nuance for progress. *J Med Philos* 2014;39(6):590–608.

39 Timmermans S, Angell A. Evidence-based medicine, clinical uncertainty, and learning to doctor. *J Health Soc Behav* 2001;42(4):342–59.

J Med Ethics: first published as 10.1136/medethics-2019-105586 on 20 November 2019. Downloaded from http://jme.bmj.com/ on August 13, 2024 by guest. Protected by copyright.