AI and Ethics (2022) 2:167–175
https://doi.org/10.1007/s43681-021-00089-6

ORIGINAL RESEARCH

Moral exemplars for the virtuous machine: the clinician's role in ethical artificial intelligence for healthcare

Sumeet Hindocha[1,2] · Cosmin Badea[3]

Received: 5 May 2021 / Accepted: 29 August 2021 / Published online: 12 September 2021

Abstract
Artificial Intelligence (AI) continues to pervade several aspects of healthcare with pace and scale. The need for an ethical framework in AI to address this has long been recognized, but to date most efforts have delivered only high-level principles and value statements. Herein, we explain the need for an ethical framework in healthcare AI, the different moral theories that may serve as its basis, the rationale for why we believe this should be built around virtue ethics, and explore this in the context of five key ethical concerns for the introduction of AI in healthcare. Some existing work has suggested that AI may replace clinicians. We argue to the contrary, that the clinician will not be replaced, nor their role attenuated. Rather, they will be integral to the responsible design, deployment, and regulation of AI in healthcare, acting as the moral exemplar for the virtuous machine. We collate relevant points from the literature and formulate our own to present a coherent argument for the central role of clinicians in ethical AI and propose ideas to help advance efforts to employ ML-based solutions within healthcare. Finally, we highlight the responsibility of not only clinicians, but also data scientists, tech companies, ethicists, and regulators to act virtuously in realising the vision of ethical and accountable AI in healthcare.

Keywords Artificial intelligence · Machine learning · Healthcare · Ethics · Virtue ethics · Morality

1 Introduction

Artificial Intelligence (AI) continues to pervade several aspects of healthcare from diagnosis, epidemiology and drug-discovery to operational performance and value improvement [1–9]. For example, Ardilla et al. present a deep-learning system for predicting lung cancer with an area under the receiver operator characteristics curve value of 94.4% [8]. Some have even gone as far as to suggest that AI will replace clinicians [10, 11]. Whilst this may seem sensationalist, by facilitating data analysis at a level far beyond the limit of human capability, AI does have the potential to disrupt and revolutionize how we see healthcare from the system, clinician and patient perspectives.

Naturally, the significant disruption posed by the introduction of AI to any industry is not straightforward to effectively manage, nor without ethical challenges. The need for an ethical framework in AI to address this has long been recognized but to date, whilst specialised technical solutions have been proposed to parts of this problem [12–14], most efforts have delivered only high-level principles and value-statements, understandably restrained in putting forward more detailed recommendations [15–18]. Even less work focuses specifically on bioethics and AI in healthcare and where it does, the stance often appears speculative [19–21]. As a result, the medical community remains largely uninformed as to the ethical intricacies introduced by AI [22].

While speculative and sensational arguments may serve to raise public awareness and focus debate on the ethical challenges that encompass the deployment of AI in healthcare, further discourse is needed particularly within the medical community, to evolve the argument towards tangible results. In this paper, we seek to contribute to this discussion. We collate relevant points from the literature and formulate our own to present a coherent argument for the central role of

✉ Sumeet Hindocha
sh806@ic.ac.uk

Cosmin Badea
cosmin.badea10@imperial.ac.uk

1 Artificial Intelligence for Healthcare Centre for Doctoral Training, Imperial College London, London, UK

2 Lung Unit, Royal Marsden Hospital NHS Foundation Trust, London, UK

3 Department of Computing, Imperial College London, London, UK

clinicians in ethical AI. With this, we propose ideas to help advance efforts to employ Machine Learning (ML)-based solutions within healthcare.

We begin with an introduction to AI and ML and the difference between them. Henceforth we center mainly on ML due to its increasing popularity as an AI technology applied to healthcare. In "Section 2: the need for ethics in healthcare AI" we explain why ethics is needed in healthcare AI. We first propose the use of virtue ethics in the healthcare setting, focusing for illustration purposes on Aristotelian virtue ethics and discuss our arguments for its suitability, contrasting it with other main moral theories encountered in AI Ethics literature—Kantian deontology and (utilitarian) consequentialism. We go on to propose that virtue ethics also be used in further developments of ethical ML, drawing on existing arguments from the literature. These arguments help to further our main point, that the advancement of ethical healthcare AI will act to safeguard the role of the clinician.

Having made our case for virtue ethics in both healthcare and ML, we take this forward in "Section 3: practical ethical considerations of ML in healthcare" to respond to some popular objections to the use of ML in healthcare. Here we discuss practical considerations, suggest future work and touch upon the very important issue of reasoning under uncertainty. In doing so we argue that the clinician will not be replaced, nor their role attenuated. Rather, we illustrate how they are and will be integral to the responsible design, deployment, and regulation of AI in healthcare, and will keep their role as the moral exemplars for the virtuous machine.

2 Section 1: artificial intelligence and machine learning

The terms AI and ML have gained buzzword-status in the healthcare domain but are erroneously often assumed to mean the same thing. Generally, AI describes the ability of machines to demonstrate intelligence, by performing tasks, problem-solving, or acting in ways that maximise its chances of reaching its goals. Artificial Intelligence is an all-encompassing field for related methods and agents that includes, amongst others, expert systems, ML and newer technologies and specialisations such as deep learning.

Expert systems make decisions by combing an extensive knowledge base using mainly logic-based "if–then" rules derived from domain-expert proficiency. They have long been trialed in healthcare and unfortunately have often met with limited adoption [23–25]. Due to this, as well as the growing frequency seen in the use of machine learning for novel AI implementations in healthcare, we shall henceforth focus our attention on ML. It has recently often yielded superior functionality compared to expert systems through its ability to learn from and adapt to the data it is provided with, without having been explicitly programmed to do so in any particular way [23–25]. ML algorithms are able to train on vast datasets, implicitly finding patterns and relationships therein and updating their parameters in an iterative fashion to create a system capable of deriving the most accurate predictions [26].

A tentative example that could illustrate the comparison between the above two methods is the following: imagine the expert system as a keen and able student of medicine with limited clinical experience, applying externally generated rules and general principles to new patients seen on the ward, and ML as a senior, experienced clinician, who has through their training reached a method of practice tailored to their particular experience, which takes precedence over any prior rule-based approaches, able to learn from this past experience and even learn from new cases while continuously self-improving their behaviour and maintaining a decision-making process that has a probabilistic aspect to it. This is an analogy also seen in [27].

3 Section 2: the need for ethics in healthcare AI

A lot of work on applying ethical AI has revolved around autonomous machine agents such as robots and self-driving cars that, by design, will be capable of making independent decisions that could impact on humans [13, 14, 28]. The need for a robust ethical framework to guide this decision-making is obvious through the lenses of safety, trust or explainability. However, whilst robots are already used in surgery, current advances in medical AI are not largely autonomous and have been designed to support rather than replace clinicians [29]. Even without autonomy, how we design and use such tools for virtuous means remains of utmost importance. For these reasons, just as clinicians' decision-making is informed by their ethical character, machine agents built for healthcare should also be informed by ethical character and this should be decided as far in advance as possible, so as to ensure control and maximal representation of medical practice.

4 The case for virtue ethics in healthcare

The rationale for why we believe virtue ethics to be the most suitable ethical framework for medical machine learning is made clear when we contrast it with the other two main families of ethical theories used as bases for work in AI ethics: Deontology and Consequentialism (Fig. 1).

Both of these theories center on the action being performed by an agent. Deontology, or duty-based ethics,

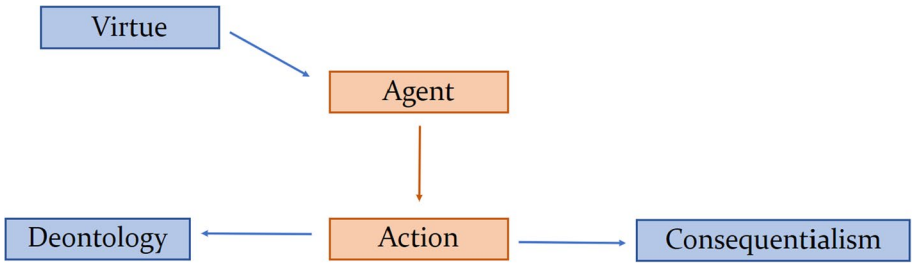

Figure. 1 The families of ethical theories considered. While deontology and consequentialism are based around discussing the ethical nature of the action being performed, as judged from the intentions/duties behind the course of action or the consequences stemming from it respectively, virtue ethics focuses more on the nature of the agent performing the action, and how one can evaluate the agent's character from his behavior

focuses on inviolable principles and duties driving the intent of one's action in a specific situation. We focus in particular on Kant's moral philosophy and our opinion follows the received view on reading Kant which sees one's rational will as the central element to be evaluated for goodness (as discussed in [30, 31]). In one formulation of his Categorical Imperative, Kant [32] argues that one should always 'act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end'. However, whilst this action may be beneficial for the object it does not always secure the same benefit for the wider population due, for example, to an uneven distribution of goods possible when acting without an explicit focus on wider consequences. Some formulations of Consequentialist ethics, for example maximising utilitarianism, take up the opposing view, that the moral weight of an action is based not on the benefit to one individual, but its maximal utility to the entire population—'the greatest good for the greatest number'[33]. However, our focus here is not on these two paradigms and so for brevity we shall take our eye away from the theory and instead focus on the practical aspects of their use.

In isolation, these contrasting ethical families could be difficult to apply and straightforwardly guide clinicians through the frequent ethical conundrums faced in healthcare, due for instance, to friction between conflicting principles. For example, whilst we have a duty to provide the patient in our consultation room with as much time as they need to understand their condition, the most suitable treatment plan and the best overall healthcare experience, we also owe this to all the patients in the waiting room. Thus, it is that for most healthcare systems, resources are finite and the responsibility of those who decide how to allocate them must be to employ a strategy that balances individual needs and those of society as a whole. This can lead to great difficulty in automating such decisions and also to some hybrid approaches such as using ML to inform a clinician's decision making.

In reality, there is no easily formalisable way to make such decisions to be readily found in the literature and, when forced to, clinicians may turn to their inherent moral reasoning and character, tempered by emotions, for guidance. Gardiner [34] states that 'the nature of our character is of fundamental importance' in moral decision making and that whilst care and consideration should be taken, that it is 'unwise to strip this process of affect or attitude and focus on reason alone'. This brings to mind the ideas found in virtue ethics and resonates well with Aristotle's theory that mirrors the above anecdotal evidence in reference to developing one's moral intuition, which is then to be relied on in making good decisions. Furthermore, as Aristotle also points out (and as has been proven true by the lack of homogeneity or agreement in moral theory in the past millennia), there may not be any ground-truth upon which everyone agrees in terms of ethics, and thus we turn to moral exemplars and the building of moral character through habit.

In contrast to deontology and consequentialism, virtue ethics focuses on the character of the moral agent performing the action, rather than the action itself. In Aristotle's formulation of this moral paradigm, a virtuous agent is one whose good character has been developed through learning and habitual practice of virtuous traits, such that they have become second-nature. Moving away from straightforward and rigid normative principles, it holds that the decisions an agent makes are rooted in its character and this makes it particularly relevant to ethical dilemmas in healthcare, where the decision reached (which stems from the clinician's *arete* (excellence)) may not always satisfy all parties. This failure to satisfy all is mitigated by the very fact that the agent is acting virtuously. As Aristotle [35] states, 'the virtue of man also will be the state of character which makes a man good and which makes him do his own work well'. Ideally, the clinician will make the best decision, because the clinician wants to make the best decision and he has built the virtuous character that makes him do his work well. Confidence in the clinicians' decision-making can then be developed through

repeated success and this focus on character in training may facilitate trust from patients through the transparency of its methods and visibility of its results.

5 The case for virtue ethics in ML and how it has been made before

So far, we have argued for the importance of virtue ethics in facilitating a deeper appreciation of the ethical dilemmas seen in healthcare, but what of virtue ethics in AI? Truly autonomous agents will require the ability to abide by ethical principles [36]. In identifying which of the three families would be most suitable to ML particularly, Berberich and Diepold [37] see virtue ethics as an obvious choice based on the way ML specifically is designed. We agree with their point and also hold that deontological and consequentialist frameworks can be seen as products of the top–down approach in building a moral expert system that can be obtained, for instance, by feeding it moral facts and either asking it to perform logical inference (deontological, rule-based) or by defining an explicit value function which is maximized by the consequences of the correct action (consequentialist). They argue that ML design on the other hand takes the bottom–up approach in learning moral actions and that this is more aligned with virtue ethics, as 'it allows non-cognitivistic and anti-universalistic positions, and thus needs philosophical justification in the form of an underlying moral theory'. They go on to argue that the central notions of learning and habituation in virtue ethics make it a more seamless and natural choice compared to deontology or consequentialism.

In addition to this, the use of virtue ethics would give us a way of learning good habits through imitating the behaviour of those we consider to be good agents (the *Phronimi),* without needing a thorough theoretical understanding of action, as required by consequentialism (in evaluating consequences and total utility) or deontology (in selecting and operating with rules and duties). This mirrors well the ethical training given to medical students, whereby one learns from prior cases to emulate the experienced approach of an existing practitioner, who can be seen as the Phronimos in that context. There are four pillars of medical ethics, the concepts of autonomy, beneficience, non-maleficence and justice [38]. The wide-spread reliance on them demonstrates how medical ethics already has virtues at its core, as striving to integrate these four ethical values into one's character is a markedly virtue-building, character-driven approach.

However, whilst we believe virtue ethics represents the best chance at making progress, we note that it is not the only paradigm that would safeguard the clinician's importance. The clinician's role is also maintained should the other moral theories be used. Data used to inform both consequentialist learning algorithms and deontological rule-based ones would arise from existing clinical experience (and by extension the clinician), and, therefore, both consequentialist and deontological theories would revolve around the clinician's activity.

6 Behind every virtuous machine…

In their paper, 'Toward the Engineering of Virtuous Machines', Govindarajulu et al. [39] draw on Aristotle's 'Nicomachean Ethics' to argue that the habitual practice of virtuous traits is learned from moral exemplars—Aristotle's *Phronimi* or 'wise ones'—'no one can become a master tennis player or pianist without, specifically, playing tennis/the piano with an eye to the mastery of great exemplars in these two domains'. Their argument is furthered by referencing the work of Vallor who states that these virtuous traits are cultivated states of character 'manifested by those exemplary persons who have come closest to achieving the highest human good'.

Given this, regardless of whether it refers to a truly autonomous machine agent or a ML algorithm designed to support clinicians, considering the lack of clear-cut rules for moral decision making in this context offered by other moral theories, it naturally follows that a machine may never be virtuous or used virtuously in healthcare without the clinician as its moral exemplar.

There may be a worry that once we have trained a virtuous machine using human moral exemplars, this machine could then act as the exemplar for others, rendering human exemplars redundant. However, we believe that this would only be possible if there were a virtuous machine that could act well in every conceivable clinical scenario, both now and in the future. Due to the continuously evolving nature of the clinical practice, new data and methods would need to be integrated into any model we have built. With every new treatment option, investigation modality or disease, come new ethical conundrums, and with them new agreed-upon solutions and beliefs about how to act best. To integrate all of this, as ML models are trained on specific data, a new training phase (or on-the-fly adaptation) would once again require our guidance as human moral exemplars.

7 Section 3: practical ethical considerations of ML in healthcare

Let us now explore some practical considerations and concerns around the use of AI in Healthcare, focusing in particular on ML. Through its very goal-oriented design, ML is expected to provide significant utility in healthcare by

performing tasks, more efficiently and faster than humans. The benefits of this will be crystallised into decreased workload, freeing up healthcare staff to handle more complicated tasks, helping optimise resource allocation and improving quality. Take for example, ML algorithms that can accurately distinguish a benign from a malignant lung nodule [8] or predict which patients are likely to miss a particular hospital appointment and send them targeted reminders or even prescribe them an appointment they are more likely to attend [2].

There have, however, been ethical concerns raised by the prospect of the introduction of ML in healthcare. Here, we explore and challenge some popular concerns in more detail, highlight the central role of the clinician and, where relevant, point out the benefits of a virtue-based approach:

8 "ML will erode the patient experience"

A patient's experience of healthcare is more than diagnosis and treatment. Good care requires psychological as well as physical wellbeing and is determined by the emotional support received along a patient's journey, shaped by their interactions with healthcare staff. Whilst ML applications may influence these interactions they are less well skilled at detecting non-verbal communication, tone of voice or other subtle cues and are not thought to be capable of emotionally driven decision-making, so would not be able to replace this human facet of care [40, 41]. This represents a key barrier to patients welcoming the introduction of AI in healthcare [42]. To allow the erosion of this important element of care would subvert the patient experience and would, therefore, be unethical [43]. Thus, as we see that traits of character such as trustworthiness or bedside manner are important, a virtue ethics or character-based approach may allow us to incorporate some of these traits, maintaining these aspects of the patient experience as much as possible. We look at this from another perspective in the next passage.

9 "ML will undermine the clinician"—decision support and reasoning under uncertainty

Framed in the notion that ML algorithms represent epistemic peers because they are trained and validated on the opinions of experts, Grote et al. raise the concern that such algorithms pose a source of uncertainty and challenge the authority of a clinician [26]. We believe that their argument, while valid, may rely on a yet unrealisable and perhaps even undesirable premise and thus may be unsound. The role of ML in healthcare does not necessarily need to be that of an autonomous peer setting off to manage cases whilst clinicians take a back seat. Taking the function of diagnosis as an example, it can instead serve as a decision support tool—the latest addition to the clinician's arsenal, alongside the stethoscope or the CT scan. Indeed, at their introduction skeptics may have argued that these tools eroded a clinician's diagnostic skill. In actuality, we believe it is obvious that they have enhanced the practice of medicine, allowing the diagnosis and treatment of more and more patients with more and more complicated diseases.

Furthermore, clinicians are well trained to deal with uncertainty. Taking a thorough history, examining the patient for physical signs, deriving further clues through blood tests and imaging and crucially, drawing on years of experience, they are experts in the multi-system integration and analysis of data and know that these clues don't always paint a clear picture. They are able to evaluate risk and make valued judgements. The maxim "don't treat the numbers, treat the patient" that senior clinicians use to teach junior colleagues exemplifies this. Consider a patient with suspected lung cancer recurrence on their CT scan. Where the clinician is uncertain as to whether this change is indeed lung cancer, they do not simply accept the CT findings and treat the patient. They can instead order a PET scan and/or request a biopsy for confirmation. Even the most sophisticated ML algorithm is unlikely to render uncertainty extinct and so in the same way, the clinician who has *arete*, or moral excellence and virtue, would know how to interpret and use the view of the algorithm, not simply letting it "run the show" in isolation or allowing it to undermine their expertise or confidence.

10 "ML threatens shared decision-making"

Whilst not always strictly adhered to in all healthcare cultures, ethical and shared-decision making is now commonplace in most settings and is recognised through "professional codes of practice and regulatory frameworks that establish fiduciary duties (of clinicians) towards their patients" [15]. These fiduciary duties are guaranteed through clinicians' professional values and self-governance. Ethical and shared decision-making is, therefore, the product of a clinician's commitments to promoting their patient's best interests. Importantly, this is not solely the prolongation of life and what may be in one patient's best interests may not be in another's. Where two patients have the same diagnosis of advanced cancer with a poor prognosis, one may choose the best supportive care and to focus on the quality of life whilst the other may choose further chemotherapy. Neither is necessarily the wrong choice and it is the *arete-imbued* clinician that works with both patients to arrive at that which is best for them.

McDougall [21] has suggested that ML algorithms pose a threat to shared decision-making as they (currently) do not facilitate consideration of patient values as parameters influencing their outputs. Although patient autonomy may be applied to the range of treatment options recommended based on a standardised value such as maximising lifespan, she argues that this is but a secondary consideration. It also fails to accurately replicate the clinician-patient decision-making process whereby a clinician would consider the patient's values from the outset when formulating management options. In reference to the above example, if an algorithm was built largely on data from patients electing for further chemotherapy, it may be more likely to offer this to the patient who preferred best supportive care, unless the patient's values were an input variable.

There are few cited examples of AI models that facilitate individual patient values as input variables. IBM's Watson for Oncology, which was trained on data from MD Anderson Cancer Centre in America has been criticised for its lack of concordance with clinicians in Asia [44, 45]. One argument for this is that cultural (patient-centered) as well as clinical factors are key in shared decision-making.

In response, McDougal calls for value-sensitive design whereby an individual patient's values can be used to weight the algorithm from the outset. We also espouse this value-based approach, and have upcoming work aiming to show how to use this to construct artificial moral agents. As it allows for the inclusion of such external values, this is another reason why a virtue-based approach is useful in answering such concerns.

Di Nucci [46] raises an important criticism to McDougall's argument with said fiduciary duties in mind. It is because of these that moral clinicians would not coerce or bias patient decision-making. Furthermore, he argues that the neutrality posed by algorithms would help to counter biased decision-making and help facilitate personalised, patient-centred choice (It should be noted that the alleged neutrality of AI algorithms is highly contested with numerous examples of biased decision-making due to skewed training data [40, 47]). In response, McDougall raises the point that the way in which ML tools are used in practice can be discordant to the intentions of their designers [48]. However, her point here relies on the assumption that ML designers had first considered patient-values, which is not necessarily the case. As such fiduciary duties as professional codes of practice, values, external regulatory structures and self-governance do not yet exist between technology developers and patients, there is no obligation or foundation for this to be built upon as of yet.

This, therefore, provides the virtuous clinician, together with the patients for whom they advocate, with a key objective: to engage with technology developers to ensure healthcare ML algorithms are developed with patient values at their core and to engage in research and debate on how such tools should be used ethically in practice.

11 "ML cannot be understood, and this is dangerous for patient care and informed consent"

ML has been labelled as a 'black box' system, and it requires complex technical skill to understand the mathematical theory underpinning it and the programming languages used to write it. Also, and of most consequence, because there exists an (at present) impassable gap between its high-dimensional mathematical ability and that of human reasoning and interpretation [49]. The former can be resolved through education and engagement of users (clinicians and patients) but although research into 'explainable AI' is widely ongoing, the latter may represent a challenge to informed consent. Due to the opacity in understanding how a ML algorithm has reached a particular outcome, the patient may lack sufficient information as to the accuracy of a diagnosis or the rationale for a particular treatment [26]. In addition to this, if the clinician is unable to explain the machine's decision, which is the crux of the 'black box' argument, this may undermine a patient's confidence and autonomy and presents another example of how ML may threaten shared decision-making.

12 From one black box to another—medicine and ML

On the contrary, there are examples of black boxes in medicine itself where our understanding of pathophysiology and mechanisms of interventions is limited. London [50] argues that the opacity seen with ML is not radically different from that seen in some aspects of clinical decision-making. Drugs with unknown mechanisms of action are routinely prescribed for numerous conditions. Bjerring et al. [51] argue that this pharmacological black-box is different to that of ML as the clinician can share with patients at least some information such as general characteristics of drug-trial participants, trial design and statistical measures that informed the result. However, is this really so different to a ML model in healthcare? If the clinician is able to explain to the patient how the ML algorithm was trained and validated, how it has performed in previous cases and

how its performance compares to the next best option, wouldn't that be sufficient to inform consent?

Furthermore, London [50] highlights Aristotle's belief that knowledge of particular facts is critical to success in action and, therefore, more important than knowledge of the principles that explain them. Patients appear to support this belief, valuing accuracy over explainability in at least some healthcare scenarios [43].

One could, therefore, argue that considered, clinical judgement, based on empirical validation of an intervention's benefits is more important. One may not be able to explain how the machine reached its conclusion, but if the conclusion is validated as being the best option for the patient, how much does this matter?

13 Who is to blame when ML gets it wrong?

It is true that ML decision-making is at risk of an underperformance that could potentially lead to harm and it is important to highlight the scale of the impact that such ML-mistakes could have on not just one, but thousands of patients [52]. Of course, such ML algorithms should not be deployed until they have passed robust audit and validation measures, however, this raises the question of who is to be held morally and legally responsible when ML makes an error. In particular, note that any moral paradigm chosen at the system's core would need to prove it is trustworthy as well as comply to relevant regulations and manage public accountability. Again, we can make a case for virtue ethics as the best central framework for ethical AI in healthcare. Every actor with *arete* playing their best part in the pathway from conceptualisation, construction and clinical use of an ML algorithm could reasonably feel responsible in the event of it failing. Schiff et al.[53] list a subset of actors who could principally be held responsible, detailing the roles they can play in mitigating medical ML errors.

All of these groups must fully understand the potential consequences for their intended course of action from the outset. Clinicians and patient-groups should engage with data scientists to produce responsible medical ML and regulatory bodies should work with these group to tackle issues around transparency, accountability, fairness and bias before ML can have a firm footing in healthcare. NHSX has already begun to do this through its code of conduct for data-drive technology which encourages innovation and technological development within the confines of ethical and regulatory boundaries [54].

We propose that future work examines these aspects, and in particular the role that virtue ethics could play in building responsibility into such systems compared to other approaches. We anticipate that being able to point to specific virtues or values will be of great use in understanding decisions and attributing responsibility, more so than that of consequences or specific rules having been followed. This is because legal and moral responsibility can be separated, as in certain forms of corrective justice where there remain reasons to assign legal responsibility to actors who may not hold moral responsibility (for example, tort law and more generally criminal law, as discussed here [55]).

14 Conclusion

Herein, we have sought to explain the need for an ethical framework in healthcare AI and the rationale for why this should be built around virtue ethics. We have gone on to argue that the introduction of AI in healthcare will not displace the clinician, but instead cement their role as the virtuous *phronimos*, or moral exemplar, for the moral machine. We have explored this further in the context of five key ethical concerns for the introduction of AI in healthcare and argued for virtue ethics' suitability in dealing with them.

We strongly believe that it would be clearly wrong to deprive patients of the significant benefits that healthcare AI brings, provided that we can ensure its safety and ethical behaviour, and, therefore, healthcare systems must evolve to accommodate it. Medical ML, for instance, can significantly improve clinical decision-making capability but for this to happen, clinicians must actualise their responsibility in engaging with and understanding it. This includes algorithm architectures, the integrity of the data they are trained on and their limitations—how their accuracy and performance vary and measure up against non-ML alternatives. Moreover, they must be able to explain the relevant parts of all these to patients to maintain patient autonomy and preserve informed consent.

Furthermore, other actors in the healthcare AI arena will also need to demonstrate virtue and play their part responsibly. For data-scientists and technology companies, this means a commitment to creating explainable AI as much as possible and to conforming to the same principalism and similar fiduciary duties to patients as their clinical counterparts. Finally, it is vividly clear that tech companies, patient-groups, clinicians, ethicists and regulators must work together to realise the vision of ethical and accountable AI in healthcare.

Author contributions All authors contributed to the paper conception and design. Material preparation and the first draft of the manuscript was written by Sumeet Hindocha and all authors commented on and revised versions of the manuscript. All authors read and approved the final manuscript.

Funding No funding was required for this work. Dr Sumeet Hindocha was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. EP/S023283/1).

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no conflicts of interest or competing interests to declare.

References

1. Ahuja, A.S.: The impact of artificial intelligence in medicine on the future role of the physician. PeerJ **2019**, e7702 (2019)
2. Nelson, A., Herron, D., Rees, G., Nachev, P.: Predicting scheduled hospital attendance with artificial intelligence. npj Digit. Med. **2**, 1–7 (2019)
3. Esteva, A., et al.: A guide to deep learning in healthcare. Nat. Med. **25**, 24–29 (2019)
4. Shah, P., et al.: Artificial intelligence and machine learning in clinical development: a translational perspective. npj Digit. Med. **2**, 1–5 (2019)
5. Doan, M., Carpenter, A.E.: Leveraging machine vision in cell-based diagnostics to do more with less. Nat. Mater. **18**, 414–418 (2019)
6. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**, 89–94 (2020)
7. Materials, N.: Ascent of machine learning in medicine. Nat. Mater. **18**, 407 (2019)
8. Ardila, D., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat. Med. **25**, 954–961 (2019)
9. Wu, J.T., Leung, K., Leung, G.M.: Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet **395**, 689–697 (2020)
10. Kocher, B., Emanuel, E.J.: Will robots replace doctors? *USC-Brookings Schaeffer on Health Policy*. https://www.brookings.edu/blog/usc-brookings-schaeffer-on-health-policy/2019/03/05/will-robots-replace-doctors/ (2019)
11. Lee, T.B.: Here's the best argument that computers could replace doctors, teachers, and even nannies | The new new economy. (2016)
12. Dietrich, F., List, C.: What matters and how it matters: a choice-theoretic representation of moral theories. Philos. Rev. **126**, 421–479 (2017)
13. Winfield, A.F., Michael, K., Pitt, J., Evers, V.: Machine ethics: the design and governance of ethical ai and autonomous systems. Proc. IEEE **107**, 509–517 (2019)
14. Winfield, A.F.: Ethical standards in robotics and AI. Nat. Electron. **2**, 46–48 (2019)
15. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**, 501–507 (2019)
16. Hagendorff, T.: The ethics of AI ethics—an evaluation of guidelines. Minds Mach. (2019). https://doi.org/10.1007/s11023-020-09517-8
17. Ben-Israel, D., et al.: The impact of machine learning on patient care: a systematic review. Artif Intell Med **103**, 101785 (2020)
18. The Economist. The EU wants to become the world's super-regulator in AI | The Economist. https://www.economist.com/europe/2021/04/24/the-eu-wants-to-become-the-worlds-super-regulator-in-ai (2021)
19. Char, D.S., Shah, N.H., Magnus, D.: Implementing machine learning in health care ' addressing ethical challenges. N. Engl. J. Med. **378**, 981–983 (2018)
20. Keskinbora, K.H.: Medical ethics considerations on artificial intelligence. J. Clin. Neurosci. **64**, 277–282 (2019)
21. McDougall, R.J.: Computer knows best? The need for value-flexibility in medical AI. J. Med. Ethics **45**, 156–160 (2019)
22. Rigby, M.J.: Ethical dimensions of using artificial intelligence in health care. AMA J. Ethics **21**, 121–124 (2019)
23. Alder, H., et al.: Computer-based diagnostic expert systems in rheumatology: where do we stand in 2014? Int J Rheumatol (2014). https://doi.org/10.1155/2014/672714
24. Vihinen, M., Samarghitean, C.: Medical expert systems. Curr. Bioinform. **3**, 56–65 (2008)
25. McCauley, N., Ala, M.: The use of expert systems in the healthcare industry. Inf. Manag. **22**, 227–235 (1992)
26. Grote, T., Berens, P.: On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics **46**, 205–211 (2019)
27. Obermeyer, Z., Emanuel, E.J.: Predicting the future-big data, machine learning, and clinical medicine. N. Engl. J. Med. **375**, 1216–1219 (2016)
28. Bird, E. et al.: The ethics of artificial intelligence: Issues and initiatives. STUDY Panel for the Future of Science and Technology vol. 27. http://www.europarl.europa.eu/thinktank (2020)
29. Abràmoff, M.D., Tobey, D., Char, D.S.: Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. Am. J. Ophthalmol. **214**, 134–142 (2020)
30. Kant, I., Gregor, M.J.: Groundwork of the metaphysics of morals. Cambridge University Press (2017)
31. Cohon, R.: Hume's Moral Philosophy (Stanford Encyclopedia of Philosophy). https://plato.stanford.edu/entries/kant-moral/ (2008)
32. Kant, I., Wood, A.W., Kant, I., Wood, A.W.: Groundwork of The metaphysics of morals (1785). In: Immanuel Kant: practical philosophy , pp. 37–108. Cambridge University Press (2012). https://doi.org/10.1017/cbo9780511813306.007
33. Mill, J.S., Bentham, J.: Utilitarianism and other essays. Penguin, UK (1987)
34. Gardiner, P.: A virtue ethics approach to moral dilemmas in medicine. J. Med. Ethics **29**, 297–302 (2003)
35. Ross, D.: Nicomachean Ethics Aristotle Translated by W (1999)
36. Wallach, W., Allen, C.: Moral machines: teaching robots right from wrong. Moral machines: teaching robots right from wrong. Oxford University Press (2009)
37. Berberich, N., Diepold, K.: The Virtuous Machine-Old Ethics for New Technology? *arXiv* (2018)
38. Gillon, R.: Medical ethics: four principles plus attention to scope. BMJ **309**, 184 (1994)

39. Govindarajulu, N. S., Bringsjord, S., Ghosh, R.: Toward the engineering of virtuous machines. *arXiv* (2018)
40. AOMRC. *Artificial Intelligence in Healthcare*. http://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf (2019)
41. Colvin, G.: Humans are underrated. Portfolio (2016)
42. Tran, V.-T., Riveros, C., Ravaud, P.: Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. npj Digit. Med. **2**, 53 (2019)
43. Morley, J., et al.: The ethics of AI in health care: a mapping review. Social Sci Med **260**, 113172 (2020)
44. Lee, W.-S., et al.: Assessing concordance with Watson for oncology, a cognitive computing decision support system for colon cancer treatment in Korea. JCO Clin. Cancer Informatics **2**, 1–8 (2018)
45. Somashekhar, S.P., et al.: Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann. Oncol. **29**, 418–423 (2018)
46. Di Nucci, E.: Should we be afraid of medical AI? J. Med. Ethics **45**, 556–558 (2019)
47. Asan, O., Bayrak, A.E., Choudhury, A.: Artificial intelligence and human trust in healthcare: focus on clinicians. J. Med. Internet Res. **22**, e15154–e15154 (2020)
48. McDougall, R.J.: No we shouldn't be afraid of medical AI; it involves risks and opportunities. J. Med. Ethics **45**, 559 (2019)
49. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. **3**, 205395171562251 (2016)
50. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. **49**, 15–21 (2019)
51. Bjerring, J.C., Busch, J.: Artificial intelligence and patient-centered decision-making. Philos. Technol. (2020). https://doi.org/10.1007/s13347-019-00391-6
52. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**, 44–56 (2019)
53. Schiff, D., Borenstein, J.: How should clinicians communicate with patients about the roles of artificially intelligent team members? AMA J. Ethics **21**, 138–145 (2019)
54. Joshi, I., Morley, J.: Ethics is our competitive advantage: how the NHS can lead the world in AI-based healthtech - Technology in the NHS. https://healthtech.blog.gov.uk/2019/02/20/ethics-is-our-competitive-advantage-how-the-nhs-can-lead-the-world-in-ai-based-healthtech/ (2019)
55. Duff, A.: Legal and moral responsibility. Philos Compass **4**, 978–986 (2009)
56. Badea, C., Gregory, A.: Morality, machines and the interpretation problem: a value-based, Wittgensteinian approach to building Moral Agents (2021). arXiv:2103.02728
57. Badea, C.: Have a break from making decisions, have a MARS: the multi-valued action reasoning system, arXiv e-prints (2021). https://ui.adsabs.harvard.edu/abs/2021arXiv210903283B. Accessed 9 Sept 2021