Received: 31 October 2020 | Revised: 2 March 2021 | Accepted: 20 March 2021

DOI: 10.1111/bioe.12887

SPECIAL ISSUE: PROMISES AND CHALLENGES OF MEDICAL AI

bioethics WILEY

Responsibility beyond design: Physicians' requirements for ethical medical AI

Martin Sand[1] | Juan Manuel Durán[1] | Karin Rolanda Jongsma[2]

[1]TU Delft, Faculty of Technology, Delft, The Netherlands

[2]University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Correspondence
Martin Sand, TU Delft, Department of Values, Technology and Innovation, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628 BX Delft, The Netherlands.
Email: m.sand@tudelft.nl

Funding information
Netherlands Institute for Advance Study in the Humanities and Social Sciences; Dutch Science Organization, Grant/Award Number: 406.Di.19.089

Abstract

Medical AI is increasingly being developed and tested to improve medical diagnosis, prediction and treatment of a wide array of medical conditions. Despite worries about the explainability and accuracy of such medical AI systems, it is reasonable to assume that they will be increasingly implemented in medical practice. Current ethical debates focus mainly on design requirements and suggest embedding certain values such as transparency, fairness, and explainability in the design of medical AI systems. Aside from concerns about their design, medical AI systems also raise questions with regard to physicians' responsibilities once these technologies are being implemented and used. How do physicians' responsibilities change with the implementation of medical AI? Which set of competencies do physicians have to learn to responsibly interact with medical AI? In the present article, we will introduce the notion of forward-looking responsibility and enumerate through this conceptual lens a number of competencies and duties that physicians ought to employ to responsibly utilize medical AI in practice. Those include amongst others understanding the range of reasonable outputs, being aware of own experience and skill decline, and monitoring potential accuracy decline of the AI systems.

KEYWORDS

competencies, entrustable professional activities, forward-looking responsibility, medical AI, medical ethics, radiology, responsibility

1 | INTRODUCTION—BEYOND DESIGN REQUIREMENTS FOR MEDICAL AI

The rapid development of artificial intelligence (AI) is considered one of the most transformative forces of our time. In medicine, the development of AI, including machine learning and deep-learning, has spawned optimism regarding the enablement of personalized care, better prevention, detection, diagnosis, and treatment of disease.[1] Some medical AI systems have already been approved by the FDA—including IDx-DR, which can be used to speed diagnose diabetic retinopathy.[2] Many machine learning approaches, especially artificial neural networks for deep learning, have proven to be particularly useful for image processing. In image-based medicine such as radiology and pathology, image screening is a time-consuming task

[1]Fogel, A. L., & Kvedar, J. C. (2018). Artificial intelligence powers digital medicine. *NPJ Digital Medicine*, *1*, 5–5. https://doi.org/10.1038/s41746-017-0012-2

[2]Keane, P. A., & Topol, E. J. (2018). With an eye to AI and autonomous diagnosis. *NPJ Digital Medicine*, *1*, 1–3, article 40. https://doi.org/10.1038/s41746-018-0048-y; Nabi, J. (2018). How bioethics can shape artificial intelligence and machine learning. *Hastings Center Report*, *48*(5), 10–13. https://doi.org/10.1002/hast.895

and screening accuracy varies amongst different physicians, institutions, and countries.[3] Furthermore, the prevalence of human-error in image screening has caused additional concerns and has motivated attempts to use computational systems to assist with image-based diagnosis.[4] Studies have indicated that medical AI can perform equally well to, or even outperform expert radiologists and pathologists in terms of accuracy when detecting, classifying, and segmenting tumors in ultrasonography, X-ray imaging, MRI scans, and digitalized microscopy slides.[5] It is therefore no wonder that most of the currently developed and proposed AI applications in medicine aim at image-based diagnostics in fields like radiology and pathology.[6] In the present paper, we will focus on the context of AI for image-driven diagnostics.

The technological possibilities of medical AI have spawned an important ethical debate that primarily focuses on technical features of medical AI and design requirements.[7] Central concerns in this debate are: How can these technologies be designed to protect privacy,[8] to prevent bias and ensure fairness,[9] to ensure explainability and to ensure accuracy of results?[10] Solutions to these problems are often sought in the design and functioning of the AI system itself. In this manner, Thilo Hagendorff concludes based on his review of AI ethics guidelines that in

> AI ethics, technical artefacts are primarily seen as isolated entities that can be optimized by experts so as to find technical solutions for technical problems. What is often lacking is a consideration of the wider contexts and the comprehensive relationship networks in which technical systems are embedded.[11]

While current ethical debates regarding the design and technical features of medical AI are important,[12] we contend that for ethical medical AI, we have to move beyond the mere technical and design aspects of these systems. In particular, we argue for a more pronounced focus on forward-looking responsibilities of physicians using such systems: The design of these technologies is only one factor influencing their eventual alignment with societal values and their ethical acceptability.[13] Humans who operate those systems are another major factor influencing the moral acceptability of those systems and their effects. Their knowledge and (technical) competencies can foster or undermine their acceptability: even the best-designed technologies fail to perform reliably in the hands of someone unskilled or someone who does not use them properly. Indeed, even if these systems are technically reliable and concerns regarding their design are overcome, the potential of these technologies can only be realized when they are used correctly and implemented in clinical practice provided certain conditions. As AI systems are increasingly being implemented, concerns about human requirements become pertinent. The conditions for interacting with and using such systems will affect whether and to what extent it will be morally acceptable to use medical AI.

We will first sketch the current ethical debate on accuracy and accountability that forms the backdrop of our plea for forward-looking responsibility. We will then discuss clinicians' responsibilities by outlining *entrustable professional activities* (EPAs), which elaborate specific competencies and skills that resident radiologists ought to be taught and ought to acquire. We will show that the current list of EPAs does not address specific technological competencies that

[3]Topol, E. J. (2019). *Deep medicine - How artificial intelligence can make healthcare human again*. Basic Books.

[4]Castellino, R. A. (2005). Computer aided detection (CAD): An overview. *Cancer Imaging: The Official Publication of the International Cancer Imaging Society*, *5*(1), 17–19. https://doi.org/10.1102/1470-7330.2005.0018; Krupinski, E. A. (2003). The future of image perception in radiology: Synergy between humans and computers. *Academic Radiology*, *10*(1), 1–3. https://doi.org/10.1016/s1076-6332(03)80781-x; Tsai, T. L., Fridsma, D. B., & Gatti, G. (2003). Computer decision support as a source of interpretation error: The case of electrocardiograms. *Journal of the American Medical Informatics Association: JAMIA*, *10*(5), 478–483. https://doi.org/10.1197/jamia.M1279

[5]Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., & Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study. *The Lancet Oncology*, *21*(2), 233–241. https://doi.org/10.1016/S1470-2045(19)30739-9; Chen, J. H., & Asch, S. M. (2017). Machine learning and prediction in medicine - Beyond the peak of inflated expectations. *The New England Journal of Medicine*, *376*(26), 2507–2509. https://doi.org/10.1056/NEJMp1702071; Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., & the CAMELYON16 Consortium. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, *318*(22), 2199–2210. https://doi.org/10.1001/jama.2017.14585; Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, *37*(2), 505–515. https://doi.org/10.1148/rg.2017160130

[6]Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, *29*(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002; Pesapane, F., Codari, M., & Sardanelli, F. (2018). Artificial intelligence in medical imaging: Threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, *2*(1), 1–10, article 35. https://doi.org/10.1186/s41747-018-0061-6

[7]Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8; Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2; Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

[8]Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

[9]Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care - Addressing ethical challenges. *The New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229; Nabi, op cit. note 2.

[10]Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz*, *34*(2), 193–198. https://doi.org/10.1007/s13218-020-00636-z; Jobin et al., op cit. note 7; London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

[11]Hagendorff, op cit. note 7, p. 103.

[12]Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*. 26, 1771–1796. https://doi.org/10.1007/s11948-020-00213-5; Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5; van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, 30, 385–409. https://doi.org/10.1007/s11023-020-09537-4

[13]Stilgoe, J. (2020). *Who's driving innovation?* Palgrave Macmillan.

are important in the context of medical AI. We will show that existing accounts of physicians' responsibility in the medical AI literature incomprehensively address forward-looking responsibilities. Furthermore, we will outline, specify, and justify what physicians will have to learn and how they should interact with medical AI in the future. This includes, amongst others, recognizing the range of reasonable output values (understanding normal functioning, abnormal deviation), understanding which type of data is processed, monitoring possible accuracy decline and variation, and awareness of AI's task specificity to be able to responsibly utilize those devices in practice.

2 | THE CURRENT DEBATE ON RESPONSIBLE AI—ACCURACY, ACCOUNTABILITY AND THE NEED FOR FORWARD-LOOKING RESPONSIBILITIES

The demand for systems that outperform humans in terms of accuracy has motivated the development of more opaque models like artificial neural networks. This has caused a tension between accuracy and explainability of these systems, as opaque models seem to be more accurate, yet difficult to interpret.[14] It has been proposed that diagnostic AI systems' principled opacity might be acceptable, if their accuracy in detecting cancer, for instance, is much higher than that of physicians.[15] Yet, it is often argued that the condition of accountability—which is linked to the explainability of such medical AI systems—has to be met too: If the outputs of medical AI cannot be explained due to the opacity of the system, trust in their decisions might be undermined and, thus, care cannot effectively and morally be provided.[16] Opacity of these systems is also legally problematic, given that under the European General Data Protection Regulation (GDPR) a right to explanation of automated decisions is legally required (General Data Protection Regulation, 2016, p. Articles 21 & 22), even though it remains disputed what such explanation should entail.[17] The requirement of accountability is often understood as a feature that has to be designed into the system: In order to identify parties responsible for damage, we should to be able to reconstruct the internal workings of the algorithm to trace the origins of a failure.

Accountability, liability, and blame are part of an understanding of responsibility that is backward-looking and that naturally arises *after* damage has occurred.[18] Forward-looking responsibility, in contrast, aims at justifying and imposing moral requirements in order to prevent damage from happening in the first place.[19] Institutionally, this perspective resonates with training and education of physicians in both professional and moral ways; such structures aim at *preventing* harm and moral misconduct *beforehand*. Forward-looking responsibilities can be understood as a safeguard to decrease the risk of harm in cases of cognitive misalignment between the physicians and the AI system—when an AI output cannot be confirmed (verified or falsified). If the promise of increased accuracy can be fulfilled and such safeguarding mechanisms are in place, this might override remaining concerns about opacity and cognitive misalignment. Forward-looking responsibilities might entail, for instance, attentiveness regarding one's own shortcomings and awareness of the technological limitations (of an AI system), which can forestall harm to patients and should therefore be fostered in clinical settings. Such forward-looking responsibilities are often expressed in less specific formulations, meaning that health care professionals can translate these responsibilities into practice suitable to the respective context in which they actualize them. Terminology from virtue ethics has proven particularly beneficial to guide educatory practices and is therefore suitable to formulate the competencies of physicians in terms of forward-looking responsibilities.[20] Unfortunately, very little has been said in the literature on medical AI about such forward-looking responsibility related to the human operators of these newly emerging socio-technical systems.[21] This is clearly a blind spot: As argued before, even if problems of algorithmic opacity were technically resolved, that does not necessarily mean that these technologies are utilized in an ethical manner. Here, more insight and guidance are needed as to how these technologies can be responsibly used.

[14]Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA*, *318*(6), 517–518. https://doi.org/10.1001/jama.2017.7797

[15]London, op cit. note 10.

[16]Grote, T., & Di Nucci, E. (2020). Algorithmic decision-making and the problem of control. In B. Beck & M. Kühler (Eds.), *Technology, anthropology, and dimensions of responsibility* (pp. 97–113). J. B. Metzler.

[17]Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

[18]Sullivan, H., & Schweikart, S. (2019). Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA Journal of Ethics*, *21*, E160–E166. https://doi.org/10.1001/amajethics.2019.160. An anonymous reviewer made us aware of a recent article by Daniel W. Tigard (Tigard, D. W. (2020). There is no techno-responsibility gap. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00414-7), who emphasizes the ambiguity of the concept "responsibility" and suggests that one of its dimensions—answerability—might well be localizable in technological systems (through a focus on humans and their interactions with these systems), even if such systems contain autonomous artifacts. Our enumeration of competencies and skills of physicians could be understood as a way of concretizing the forward-looking component of Tigard's broader idea of answerability in technological systems.

[19]van de Poel, I. (2011). The relation between forward-looking and backward-looking responsibility. In N. A. Vincent, I. van de Poel, & J. van den Hoven (Eds.), *Moral responsibility* (pp. 37–52). Springer; van de Poel, I., & Sand, M. (2018). Varieties of responsibility - Two problems of responsible innovation. *Synthese*. https://doi.org/10.1007/s11229-018-01951-7

[20]Steutel, J. W. (1997). The virtue approach to moral education: Some conceptual clarifications. *Journal of Philosophy of Education*, *31*(3), 395–407. https://doi.org/10.1111/1467-9752.00064

[21]Mark Coeckelbergh insinuates forward-looking responsibility when suggesting that developers and users must close their knowledge-gaps regarding AI systems to become answerable when using them (Coeckelbergh, M. (2019). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 26, 2051–2068, https://doi.org/10.1007/s11948-019-00146-8). Yet, Coeckelbergh's "answerability" condition is a structural requirement that remains normatively vague; since bad answers are answers too, a positive enumeration of the knowledge that has to be gained to give good answers is necessary. Our proposal makes concrete suggestions regarding the knowledge and competencies required in the context of medical AI.

What do physicians have to learn, take care of, be aware of to successfully and responsibly interact with medical AI? Outlining these positive requirements for medical AI has so far been omitted, thereby missing the chance to guide the implementation of medical AI systems in more desirable directions.

3 | FORWARD-LOOKING RESPONSIBILITY IN THE CLINIC: EPA STANDARDS

In medicine, there is an increasing adoption of entrustable professional activities (EPAs) that describe the competencies that medical residents have to learn during their time in training.[22] Radiology is a particularly interesting field of specialization, because of the rapid development of AI in this field and because of these professionals' long-standing experience with highly complex technologies such as MRIs. EPAs are units of professional practice, defined as tasks or responsibilities that the trainee gets entrusted with for unsupervised execution once he or she has attained sufficient competence.[23] Making entrustment decisions for unsupervised practice requires observed proficiency, commonly on multiple occasions. EPAs involve clinical skills and abilities and more general facets of competence, such as understanding one's own limitations and knowing when to ask for help. Five levels of supervision are distinguished in assessing whether a trainee can be trusted with a certain EPA. These are: (a) observation but no execution, even with direct supervision; (b) execution with direct, proactive supervision; (c) execution with reactive supervision, i.e., on request and quick availability; (d) supervision at a distance and/or post hoc; (e) supervision provided by the trainee to more junior colleagues. These levels of supervision also indicate the growing responsibility of the trainee.

EPAs for radiology set a standard for the required competencies and help to assess whether residents are capable of assuming certain responsibilities. The following list of EPAs focuses on the abilities of physicians to have knowledge about (evidence-based) protocols and guidelines, to check for missing information, to ensure that the examination or procedure is conducted safely and correctly, regarding image interpretation and the ability to make differential diagnoses, to report clearly and accurately in order to inform referring physicians well and the assessment of one's own knowledge base ability to recognize limitations and mistakes (Table 1). We agree with Deitte et al., who consider this list of EPAs as a starting point for EPA development rather than being exhaustive or final.[24]

TABLE 1 Entrustable professional activities (EPAs) for radiologists

- Selects triages/protocols
- Collaborates as a member of an interprofessional team
- Interprets exams and prioritizes a differential diagnosis
- Communicates results of exams
- Recommends appropriate next steps
- Obtains informed consent and performs procedures
- Manages patients after imaging and procedures
- Formulates clinical questions and retrieves evidence to advance patient care
- Behaves professionally
- Identifies system failures and contributes to a culture of safety and improvement

As EPAs prompt professionals to anticipate or recognize possible mistakes and dangerous situations, and thereby to prevent harm, they elaborate forward-looking responsibilities and exceed a narrow focus on accountability. It is, however, notable that the list of EPAs relates solely to failures in human-human centered workflows. *Technical* competencies are not explicitly mentioned, which is surprising, given that radiology is one of the more technical specializations within medicine. These EPAs, for instance, do not require radiologists to understand and be able to *explain* how technologies such as MRI systems produce images, while this is an often-pushed requirement in the current medical AI debate. A number of further requirements that we will outline below are equally neglected: Arguably, radiologists working with AI also need to identify systems' failure rooted in the technical artifact and in their interaction with these technologies. These are substantial shortcomings of current EPAs that do not accommodate the shifting roles of radiologists in light of the increasing implementation of medical AI.

4 | TECHNICAL COMPETENCIES FOR MEDICAL AI

Jha and Topol are amongst the few authors who have alluded to forward-looking responsibility suggesting that medical AI may lead to a shift in the roles and responsibilities of radiologists and pathologists.[25] While they assert that AI systems will allow for more interaction with patients, they also suggest that radiology and pathology are likely to converge to form a new profession—that of the "information specialist"—a job mainly concerned with the interpretation of data. An information specialist's education will strongly focus on the

[22]We thank an anonymous reviewer for making us aware of other enumerations of professional competencies more specifically focused on radiologists such as the "Specialty Training Requirements in Diagnostic Radiology" of the Royal College of Physicians and Surgeons of Canada. Given their broader vision of "good professional conduct" in health care, we believe that EPAs lend themselves well to incorporate our suggested ideas of increased awareness and sensitivity and they can guide educatory practices not only of radiologists but also of other health care professionals. We are convinced that the suggested, extended list of EPAs greatly complements other frameworks of professional competencies that serve other purposes and are used in other stages of the educational development.

[23]Ten Cate, O. (2013). Nuts and bolts of entrustable professional activities. *Journal of Graduate Medical Education*, *5*(1), 157–158. https://doi.org/10.4300/JGME-D-12-00380.1

[24]Deitte, L. A., Gordon, L. L., Zimmerman, R. D., Stern, E. J., McLoud, T. C., Diaz-Marchan, P. J., & Mullins, M. E. (2016). Entrustable professional activities: Ten things radiologists do. *Academic Radiology*, *23*(3), 374–381. https://doi.org/10.1016/j.acra.2015.11.010

[25]Jha, S., & Topol, E. J. (2016). Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA*, *316*(22), 2353–2354. https://doi.org/10.1001/jama.2016.17438

acquisition of mathematical skills. Jha and Topol also suggest that these professionals will no longer need to learn how to interpret ultra-scans—a task that can fully be taken over by machines. We agree that the interpretation skills of physicians should be fostered rather than being regressed. Considering their shifting roles, we need to determine, however, which competencies physicians need to maintain, which tasks can be taken over by machines and which new competencies have to be acquired by physicians working with these machines. If clinicians should serve as supervisors and custodians for medical AI systems to ensure accountability, they also need to sustain many of their current capabilities such as those to assess ultra-scans and MRI images.

Elsewhere, Topol underscores his conviction that human interaction and patient contact in radiology and pathology will increase.[26] Physicians will be able to spend more time with the patient, thus being able to undertake a longer initial diagnostic conversation to assess whether an imaging procedure is actually necessary ("gatekeeping"; p. 122). Topol further suggests that radiologists should become "master explainers" and "integrate and explain medical results" facilitated by medical AI for the patient (p. 121). This envisioned shift in responsibility is remarkable: Topol does not seem to be worried that currently radiologists and pathologists do not regularly have such intimate doctor-patient interactions and that they might lack both the training and the experience to assume these novel responsibilities. Topol's vision is far off the reality in these fields. While current EPAs in radiology include "communicat[ing] results of exams" (see Table 1), this is mainly done by oncologists and other clinical specialists, who are much more experienced in communicating with patients as a result. Radiologists typically report to other clinicians and often appear in the procedure only when it comes to billing, as Topol also notes (p. 122). Some of these radiologists, who may have started their job with a more analytic interest in mind, might not be very enthusiastic or willing to embrace such new human-centered responsibilities. Balancing the increasing technical complexities of medical AI while simultaneously mastering the increasing human element is an upcoming challenge in these professions.

Thomas Ploug and Soren Holm suggested an understanding of "explainability" of medical AI in terms of "contestability," which is much in line with the formulation as outlined in the GDPR. They argue that such contestability provides opportunities for individuals to counter automated decisions: "[…] AI decision-making must be explainable to a degree that makes it possible for an individual to contest the decision of the system."[27] Ploug and Holm draw on patients' privacy rights and their right to defend themselves against harm from which particular duties (which are types of forward-looking responsibilities) arise for doctors. They suggest that the patients' rights for privacy and data sovereignty oblige health care professionals (they do not solely address physicians) to provide information about the data sources that serve as input for medical AI. Second, the authors assert the following:

> individuals have a right to protect themselves against discrimination, and therefore should be granted a right to contest bias in AI diagnostics. Exercising the right to contest bias requires that individuals have access to information about 1) the character of the dataset on which the model is built, 2) how the data were categorised by humans, and 3) the character and level of testing the AI model has undergone.

Third, patients have a "right to contest the performance of AI diagnostics," which requires sharing "1) information about the performance of the AI model, and 2) information about the tests used to determine the performance." Lastly, Ploug and Holm assert "the right to contest the division and organisation of diagnostic labour," granting that

> [e]xercising this right requires 1) information about the role of AI in the diagnostic process, 2) information about the role of HCPs in the diagnostic process. Challenging the organisation and division of diagnostic labour also requires 3) information about the objective/legal responsibility for diagnostic procedures.

Several aspects are noteworthy here: First, taken together these four points substantiate a number of new responsibilities for health care professionals. While Ploug and Holm understand contestability as an ex post activity, their phrasing suggests that health care professionals should be *prepared* to provide requested information, which is a forward-looking task. It remains, however, unclear who precisely is responsible for these tasks: radiologists, nurses, or administrative staff? They are all potential addressees of such duties. Second, the rights mentioned above—the right to privacy and contestability—seem reducible to a more basic right, the "right to defend against harm," as Ploug and Holm call it, as privacy violations and lack of contestability might cause harm (mentally and physically). Interestingly though, the right to defend against harm does not necessarily oblige anyone else to prevent the harm from occurring in the first place. The phrasing locates the active part on the side of the patient: *If* harm is forthcoming, the patient is allowed to fend it off. Yet, this does not specify any responsibility for the physician. One might argue charitably that responsibilities of physicians are implied in such patient rights; however, for purposes of improving the future implementation of AI systems, a more detailed explication of such responsibilities is important. Third, Ploug and Holm focus largely on information exchange with patients, thereby leaving a large class of practices unaddressed that could equally cause damage, e.g., treatment that is administered based on (undetected) flawed diagnostic outputs. Ploug's and Holm's list remains incomplete with regard to

[26]Topol, op cit. note 3.

[27]Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, *107*, 101901. https://doi.org/10.1016/j.artmed.2020.101901

newly emerging technical competencies of clinicians in the diagnostic process. Fourth, while contestability is in line with two important patient rights, the "burden" of *understanding* the medical system, data and the subsequent decision seems to lie primarily on the patients' shoulders. While contestability should be promoted, the patients are no experts and are therefore not in the best position to assess the data, the data set, and the usage of the algorithm. It is reasonable to educate physicians using medical AI systems about the initial data sets, their meaningful applications and about their limitations, but also to teach them how to convey such information to non-experts, which is a separate and intricate skill that has to be practiced and learned.

5 | FORWARD-LOOKING RESPONSIBILITIES FOR MEDICAL AI IN IMAGE-BASED DIAGNOSTICS

The previous discussion has shown that one can expect a fundamental shift in the roles and responsibilities of physicians due to the implementation of medical AI. Being a competent operator of such systems, however, demands more from physicians than becoming information specialists. It requires a more general awareness of the fallibility of these systems and the various ways in which their utilization might fail. It is obvious that patients' rights are not the only source for such forward-looking responsibilities: If we adopt a more value pluralistic view on medical practice, we see that also other principles such as patients' autonomy, beneficence and non-maleficence can ground forward-looking responsibilities.[28] These values deserve preservation and promotion and not only through compliance with duties and obligations: Forward-looking responsibilities can also be expressed in terms of competencies and virtues, as mentioned above.[29] In the following, we provide an extension of the list of EPAs to include some of the most neglected competencies and virtues relevant for the ethical implementation of medical AI. The following suggestions often intersect and should be understood as an elaboration rather than a complete set of relevant competencies of physicians dealing with medical AI:

1. *The duty to report uncertainty (sensitivity/specificity rates) to the patients:* If medical AIs regularly outperform humans in detecting certain diseases and classifying images, we will face situations in which humans are incapable of perceiving and ostensibly showing what led the AI to identify a certain data piece as malign tissue. If the radiologist ought to proceed and make a treatment suggestion based on such an AI diagnosis, she will have to justify her choice and her reliance on the AI. She is epistemically dependent on the AI system and has to critically assess the AI's accuracy. Today, only a few medical AI systems have been tested in real world clinical settings.[30] Thus, extrapolation of performance levels from previous performances in experimental settings contains uncertainty, which is relevant information for both patients and physicians. In the absence of a sufficient explanation for the AI's output for this particular patient, the patient also has to be informed about previous accuracy levels obtained in experimental settings and how radiologists extrapolated from those results.
2. *Understanding and critically assessing whether AI outputs are reasonable given a certain diagnostic procedure:* Physicians need to critically assess what output values are reasonable given certain input values. In order to recognize when AI systems provide flawed outputs, physicians need to understand the range of plausible outputs to be expected given certain input data. Physicians need to have a general, yet meaningful understanding of the origins of the data, the purpose for which it was gathered, curated and analyzed, in order to assess whether the system is rendering a reasonable outcome. To illustrate this point somewhat exaggeratedly: Physicians should become skeptical when a medical AI suggests prostate cancer based on a brain scan.
3. *Knowing and understanding the input data and its quality:* Physicians should be able to assess the data being used in the AI system and know which type of data is being used for a particular procedure. This responsibility requires an understanding of the limitations of these data sets. Previous research has shown that by applying AI in a real world clinical setting a "mismatch between training and operational data can be inadvertently introduced [...] by deficiencies in the training data, but also by inappropriate application of a trained ML system to an unanticipated patient context."[31] If radiologists remain cautious of the quality of the input data, they might sustain a higher level of awareness for spurious outputs by an AI based on flawed or low-quality inputs.
4. *Awareness of one's own experience and skill decline:* It has been shown that increased training and experience increase accuracy amongst radiologists.[32] It can be expected that this training and experience will diminish once AI systems are put into practice and regularly take over the task of image analysis. If radiologists and pathologists ought to become critical custodians or supervisors who exercise oversight (which is reasonable given AI's task

[28]Beauchamp, T. L., & Childress, J. F. (1994). *Principles of biomedical ethics*. Oxford University Press.

[29]Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

[30]Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E., Ioannidis, J. P. A., Collins, G., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, *368*, m689. https://doi.org/10.1136/bmj.m689

[31]Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, *28*(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

[32]Esserman, L., Cowley, H., Eberle, C., Kirkpatrick, A., Chang, S., Berbaum, K., & Gale, A. (2002). Improving the accuracy of mammography: Volume and outcome relationships. *Journal of the National Cancer Institute*, *94*(5), 369–375. https://doi.org/10.1093/jnci/94.5.369; Nodine, C. F., Kundel, H. L., Mello-Thoms, C., Weinstein, S. P., Orel, S. G., Sullivan, D. C., & Conant, E. F. (1999). How experience and training influence mammography expertise. *Academic Radiology*, *6*(10), 575–585. https://doi.org/10.1016/S1076-6332(99)80252-9

specificity, see [5] below) to ensure accountability, they should be careful of deskilling and "automation bias." Studies have indicated that "automation-bias," meaning an overreliance on automated systems,[33] and "deskilling" of clinicians, referring to a decline of skills due to a lack of regular training, can result from increased automation of certain tasks. Radiologists will have to expose themselves to practice to not unlearn the necessary skills for analyzing medical images, even when AI systems are reliable and accurate.

5. *Awareness and understanding of task specificity:* Current diagnostic AI systems are usually trained to identify one type of disease and can perform very specific tasks very well. While radiologists can search for a number of different conditions when assessing medical images, the currently developed algorithms can typically only fulfill a particular task. This means that an algorithm that can detect prostate cancer accurately is typically unable to detect other forms of cancerous cells.[34] Being aware of this task specificity is crucial when applying these systems and for communicating the results to patients and other specialists. If a medical AI system does not identify a malignancy, radiologists must be wary that this does not mean that a patient is disease-free.
6. *Assessing, monitoring, and reporting output development over time:* In order to properly assess whether the diagnosis for a particular patient is accurate, clinicians need to monitor the reliability of the AI system over time. Such monitoring is important for two reasons: Future algorithms might be capable of learning on a case-by-case basis in real world settings. They might, in other words, be able to improve and learn "on the job."[35] Such devices have benefits since their accuracy levels improve while already being employed to improve medical practice. Although, devices that continuously learn are currently not being developed[36]—it is important for radiologists to be alert and consider performance progression over longer times. Even "static" AI systems will likely require software updates and (slight) reprogramming. There is no guarantee that the implemented learning progresses instead of declines, which is why performance variations have to be monitored over time. Second, the potential for hacking of AI systems that might be constantly connected to networks for updates and re-programming is much greater than with stationary physical artifacts such as MRI scans.[37] Manipulation of AI systems can occur undetected without any physical traces. A systematic malfunctioning based on external software manipulation can only be detected if the operating radiologists are aware of such a possibility, continuously monitor the performance of the AI system and cautiously assesses the reasonableness of AI outputs (see [2]).

TABLE 2 Extension of entrustable professional activities (EPAs) for medical artificial intelligence (AI)

1. Reporting and informing about sensitivity rates and experimental performance
2. Understanding reasonable output
3. Understanding input data (e.g., relationship between image quality and accuracy rate)
4. Awareness of impact of utilizing medical AIs on one's own skills and capacities
5. Awareness of task specificity of the medical AI
6. Assessing, monitoring and reporting of outputs over time

In summary, we advocate extending current EPA standards with these six competencies summarized in Table 2 to accommodate the specific challenges that arise through medical AI.

6 | CONCLUSIONS

In the present paper, we have argued that the current ethical debate about medical AI is predominantly focused on the design and technical features of these systems. Constructive proposals suggest ways to minimize bias in the process of "training" these algorithms, implementing certain values such as explainability in their design and improving their accuracy. We have argued that medical AI systems remain ethically fragile in practice if their use is not additionally supported by physicians who are competent and skilled when interacting with these technologies. Unlike liability and accountability, which stand at the fore of the current debate, these responsibilities are forward-looking. Medical AI causes a shift in the roles and practices of radiology, which also induces a need to explicate health care professionals' newly emerging duties, competencies, and responsibilities. We have shown that current EPAs fail to mention the important technological competencies and skills that are pivotal for the responsible implementation of medical AI. We have argued that emerging forward-looking responsibilities can be justified both in relation to patients' rights as well as with reference to other well-established principles and values in biomedical ethics such as the principles of beneficence and non-maleficence. Based on a critical reading of the current literature, we have established six key competencies that are crucial for the responsible use of AI in a medical context (Table 2). We submit that these competencies are important additions to existing EPAs and should guide health care professionals' education in the future. Our enumeration has two additional benefits:

[33]Cabitza et al., op cit. note 14.

[34]Hosny, A., Parmar, C., Coroller, T. P., Grossmann, P., Zeleznik, R., Kumar, A., Bussink, J., Gillies, R. J., Mak, R. H., & Aerts, H. J. W. L. (2018). Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Medicine*, *15*(11), e1002711. https://doi.org/10.1371/journal.pmed.1002711

[35]Keane & Topol, op cit. note 2.

[36]Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., Ashrafian, H., Beam, A. L., Collins, G. S., Darzi, A., Deeks, J. J., El Zarrad, M. K., Espinoza, C., Esteva, A., Faes, L., Ferrante di Ruffano, L., Fletcher, J., Golub, R., Harvey, H., Haug, C., … & Yau, C. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *The Lancet Digital Health*, *2*(10), e549–e560. https://doi.org/10.1016/S2589-7500(20)30219-3

[37]Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., & Rueckert, D. (2020). Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, *7*, 25–25. https://doi.org/10.3389/fcvm.2020.00025; Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, *363*(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

It can guide self-assessments—residents who do not meet these requirements shall refrain from using these systems unsupervised. Furthermore, our thoughts on professional responsibilities can positively inform how such systems should be designed, as there is a complementary relation between design questions and professional responsibilities.

The list of EPAs is a starting point and can—and should—be extended in numerous ways. This has to be done while remaining aware of the risk of overburdening physicians with too much responsibility. Future practice will also have to show whether these tasks are better off in the hands of other health care professionals or might be in some way distributed in interdisciplinary teams.[38] This, however, would not undermine the importance of our list of suggested competencies and skills. As a first step towards ethical medical AI, we submit these recommendations for implementation.

Acknowledgments

This work was supported by the Netherlands Institute for Advance Study in the Humanities and Social Sciences (NIAS-KNAW), which enabled us to closely collaborate as a NIAS-Lorentz Theme Group on "Accountability in Medical Autonomous Expert Systems: Ethical and Epistemological Challenges for Explainable AI." We are grateful to all Fellows at NIAS and our Theme Group members Sander Beckers and Giuseppe Primiero for their constructive feedback and support. Karin Rolanda Jongsma's contribution to this paper was partially funded by the Dutch Science Organization (NWO), RAIDIO project with grant number 406.Di.19.089.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Martin Sand https://orcid.org/0000-0001-8167-4581
Juan Manuel Durán https://orcid.org/0000-0001-6482-0399
Karin Rolanda Jongsma https://orcid.org/0000-0001-8135-6786

[38]Pesapane et al., op. cit. note 6.

AUTHOR BIOGRAPHIES

Martin Sand is a Lecturer in Ethics of Technology at TU Delft. Before becoming a member of the theme group on "Accountable and Explainable Medical AI" at the Netherlands Institute for Advanced Study (NIAS), he undertook a postdoctoral Marie Skłodowska-Curie project on "Moral Luck in Science and Innovation." He is intrigued by the complex relationship between luck and (forward-looking) responsibility and the role of utopias in technological development.

Juan Manuel Durán is an Assistant Professor in Ethics of Technology and Philosophy of Science at TU Delft, who received the Herbert A. Simon Award (IACAP) for outstanding research in computing and philosophy in 2019. In 2020, he coordinated the theme group "Accountable and Explainable Medical AI" at the Netherlands Institute for Advanced Study (NIAS). His interests lie in the intersection of philosophy of science, philosophy of technology, and computer science.

Karin Rolanda Jongsma is an Assistant Professor of Bioethics at the University Medical Center Utrecht. Her research focuses on the ethics of AI and the ethics of patient and public involvement. She was a member of the theme group on "Accountable and Explainable Medical AI" at the Netherlands Institute for Advanced Study (NIAS) and is the coordinating-researcher of RAIDIO: Responsible Artificial Intelligence in Clinical Decision-Making.

How to cite this article: Sand M, Durán JM, Jongsma KR. Responsibility beyond design: Physicians' requirements for ethical medical AI. *Bioethics*. 2022;36:162–169. https://doi.org/10.1111/bioe.12887