

---

## Perspective

# Trust and medical AI: the challenges we face and the expertise needed to overcome them

Thomas P. Quinn<sup>1</sup>, Manisha Senadeera<sup>1</sup>, Stephan Jacobs<sup>1</sup>, Simon Coghlan<sup>2</sup>, and Vuong Le<sup>1</sup>

<sup>1</sup>Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia and <sup>2</sup>Centre for AI and Digital Ethics, School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

contacttomquinn@gmail.com

Received 21 August 2020; Editorial Decision 6 October 2020; Accepted 12 October 2020

### ABSTRACT

Artificial intelligence (AI) is increasingly of tremendous interest in the medical field. However, failures of medical AI could have serious consequences for both clinical outcomes and the patient experience. These consequences could erode public trust in AI, which could in turn undermine trust in our healthcare institutions. This article makes 2 contributions. First, it describes the major conceptual, technical, and humanistic challenges in medical AI. Second, it proposes a solution that hinges on the education and accreditation of new expert groups who specialize in the development, verification, and operation of medical AI technologies. These groups will be required to maintain trust in our healthcare institutions.

**Key words:** AI, machine learning, challenges, ethics

---

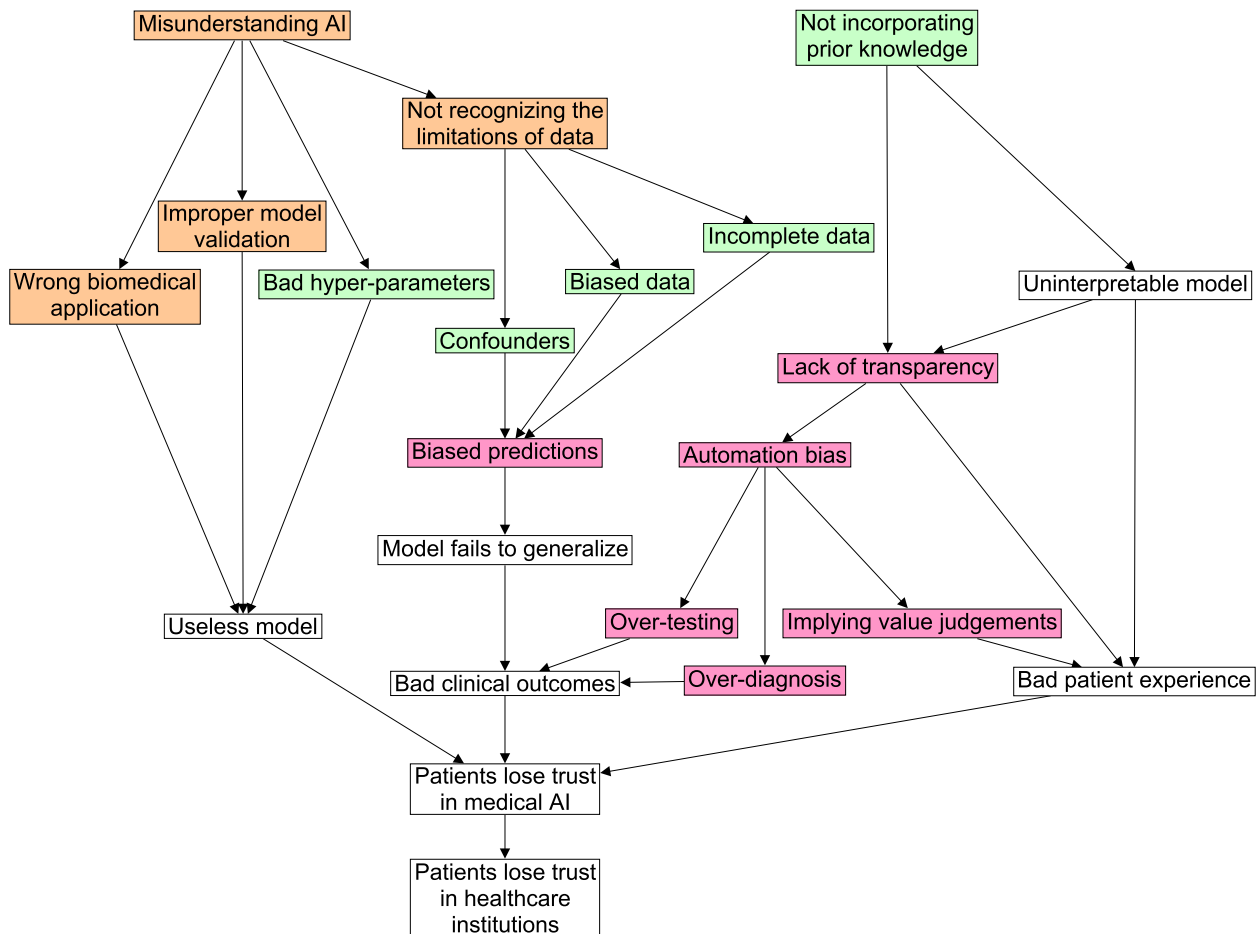
## TRUST AND MEDICAL AI

Trust underpins successful healthcare systems.<sup>1</sup> Artificial Intelligence (AI) both promises great benefits and poses new risks for medicine. Failures in medical AI could erode public trust in healthcare.<sup>2</sup> Such failure could occur in many ways. For example, bias in AI can deliver erroneous medical evaluations,<sup>3</sup> while deliberate “adversarial” attacks could undermine AI unless detected by explicit algorithmic defenses.<sup>4</sup> AI also magnifies existing cyber-security risks, potentially threatening patient privacy and confidentiality.

Successful design and implementation of AI will therefore require strong governance and administrative mechanisms.<sup>5</sup> Satisfactory governance of new AI systems should span the period from design and implementation through to repurposing and retirement.<sup>5</sup> In 2019, McKinsey Company reviewed changes needed to manage algorithmic risk in the banking sector.<sup>6</sup> Its advice hinges on AI’s sheer complexity: just as the development of an algorithm requires deep technical knowledge about machine learning, so too does the

mitigation of its risks. McKinsey & Company discuss the need to involve 3 expert groups: (1) the group developing the algorithm, (2) a group of validators, and (3) the operational staff.

These groups are also needed in the healthcare sector to overcome the following 3 key challenges in AI: (1) *conceptual challenges* in formulating a problem that AI can solve, (2) *technical challenges* in implementing an AI solution, and (3) *humanistic challenges* regarding the social and ethical implications of AI. This article offers concise descriptions of these challenges, and discusses how to ready expert groups to overcome them. Recognizing these challenges and readying these experts will put the medical profession in a good position to adapt to the changing technological landscape and safely translate AI into healthcare. Conversely, failure to address these challenges could erode public trust in medical AI, which could in turn undermine trust in healthcare institutions themselves (see [Figure 1](#)).



**Figure 1.** This figure shows how key challenges in medical AI relate to one another and to clinical care. If these challenges go unaddressed, the consequences could act concertedly to erode trust in medical AI, which could further undermine trust in our healthcare institutions. Node color represents the type of challenge: conceptual (orange), technical (green), or humanistic (pink). Uncolored nodes represent consequences.

## THE CHALLENGES WE FACE

### Conceptual challenges

Before we can translate AI into the healthcare setting, we must first identify a problem that AI can solve given the data available. This requires a clear conceptual understanding of both AI and medical practice. Conceptual confusion about AI's capabilities could undermine its deployment. Currently, AI systems cannot reason as human physicians can. Unlike physicians, AI cannot draw upon “common sense” or “clinical intuition.” Rather, machine learning (the most popular type of AI) resembles a signal translator in which the translation rule is learned directly from the data. Nevertheless, machine learning can be powerful. For example, a machine could learn how to translate a patient's entire medical record into a single number that represents a likely diagnosis, or image pixels into the coordinates of a tissue pathology. The nascent field of “machine reasoning” may one day yield models that connect multiple pieces of information together with a larger body of knowledge.<sup>7</sup> However, such reasoning engines are currently far from practically usable.

Any study involving AI should begin with a clear research question and a falsifiable hypothesis. This hypothesis should state the AI architecture, the available training data, and the intended purpose of the model. For example, a researcher might implicitly hypothesize that a long short-term memory (LSTM) neural network trained on

audio recordings of coughs from hospitalized pneumonia patients could be used as a pneumonia screening tool. Stating the hypothesis explicitly can reveal subtle oversights in the study design. Here, the researcher wants an AI model that diagnoses pneumonia in the community, but has only trained the model on patients *admitted* for pneumonia. This model may therefore miss cases of mild pneumonia, and thus fail in its role as a general screening tool. Meanwhile, model verification requires familiarity with abstract concepts like overfitting and data leakage. Without this understanding, an analyst could draw incorrect conclusions (notably, to conclude that a model *does* work when it *does not*).

It is equally important to conceptualize the nature of the medical problem correctly. Although analysts might intend for a model to produce reliable results that match the standards set by human experts, this is impossible for problems in which no standard exists (eg, because experts disagree about the pathophysiology or nosology of a clinical presentation). Even when a standard does exist, models can still recapitulate errors or biases within the training data.

### Technical challenges

AI is a dynamic and evolving field, and (like medicine itself) could be considered as much art as science. This makes AI much harder to implement and use appropriately than other technologies that come

with a user manual. For example, while LSTM is widely used for sequential data, its specific application to electroencephalogram (EEG) signals requires the analyst to carefully tune dozens of *hyperparameters*, such as the sampling rate, segment size, and number of hidden layers. These all have a major impact on performance, yet there is no universal “rule-of-thumb” to follow.

AI benefits from 2 sources of information: (1) prior knowledge as provided by the domain expert, and (2) real-world examples as provided by the training data. With the first source of information, the model designer encodes expert knowledge into the model architecture, optimization scheme, and initial parameters, which all guide how the model learns. This is hard to do when the problem at hand is complex or ill-defined, as is the case in healthcare, where physician reasoning is not easily expressed as a set of concrete rules.<sup>8</sup>

With the second source of information, a generic model is fit to the observed data. This can deal nicely with ambiguities by discovering elaborate statistical patterns directly from the training data and can also help update imperfect expert knowledge embedded within the algorithm. However, data-oriented models have problems too, especially when applied to healthcare, where data can be scarce or incomplete (eg, owing to differences in disease prevalence or socioeconomic factors). Such factors intensify the risk of covariate shift, confounder overfitting, and other model biases,<sup>9</sup> thus reducing the trustability of purely data-oriented models.

### Humanistic challenges

Patients are not mere biological organisms, but human beings with general and individualized needs, wishes, vulnerabilities, and values.<sup>10</sup> The human dimension of healthcare involves a unique professional–patient relation imbued with distinctive values and duties. This relation is widely regarded as requiring a *patient-centered approach* which respects patient autonomy and promotes informed choices that align with patient values.<sup>11</sup> Other values include the duties of privacy, confidentiality, fairness, and care, as well as the promotion of benefit (beneficence) and avoidance of harm (nonmaleficence).<sup>12</sup> Medical AI must align with these values.

Many AI models are “black boxes” that (for proprietary or technical reasons) cannot explain their recommendations.<sup>13</sup> This lack of transparency could conceivably damage epistemic trust in the recommendations and diminish autonomy by requiring patients to make choices without sufficiently understanding the relevant information.<sup>14</sup> The use of black boxes also makes it difficult to identify biases within models that could systematically lead to worse outcomes for under-represented or marginalized groups<sup>15</sup>—an important limitation given that such biases can even be present for theoretically fair models.<sup>16</sup> Meanwhile, some models tend to rank treatment options from best to worst, implying value judgements about the patient’s best interests.<sup>17</sup> For example, rankings could prioritize the maximization of longevity over the minimization of suffering (or *vice versa*). If practitioners fail to incorporate the values and the wishes of a specific patient into their professional decisions, the AI system may paternalistically interfere with shared decision-making and informed choice.<sup>18,19</sup> A patient may even wish to follow a doctor’s opinion without any machine input.<sup>20</sup> Trust could be damaged if patients discover that healthcare workers have used AI without seeking their informed consent.

One problem for any powerful AI (interpretable or not) is that practitioners may come to over-rely on it, and even succumb to automation bias.<sup>21</sup> Over-reliance, whether conscious or unconscious, can lead to harmful (maleficent) patient outcomes due to flawed

health decisions, overdiagnosis, overtreatment, and defensive medicine.<sup>22</sup> Concern has also been raised about the incremental replacement of human beings with AI systems. For example, robot carers may soon look after older adults at home or in aged care.<sup>23</sup> This may deprive people of the empathetic aspects of healthcare that they want and need.<sup>24</sup>

Another challenge involves the legal regulation of medical AI, especially with regard to attributing liability in the case of a catastrophic failure. The topic of legal liability is especially challenging for black-box models, where it can be difficult to know whether, when, and which models apply.<sup>25</sup> A lack of clarity about legal liability could negatively influence the responsible use of medical AI, for example by promoting defensive medicine<sup>26</sup> or slowing the adoption of validated tools.<sup>27</sup>

To address the conceptual, technical, and humanistic challenges of AI in medicine, 3 expert groups are required: developers, validators, and operational staff (see Table 1). The section below discusses how to ready these groups, and asks that “we”—as in all of us involved in designing, implementing, introducing, or using medical AI systems—work collectively towards solving the relevant challenges.

## THE EXPERTS WE NEED

### The group developing the algorithm

This group must understand the technical details of AI systems, but also how these details influence outcomes for patients. As such, this group should involve not only AI practitioners but also healthcare professionals, patient advocates, and medical ethicists who together enable design processes that are flexibly sensitive to individual patient values.<sup>16,17,28</sup>

### How to ready this group

In the short term, we need to prioritize transdisciplinary research collaborations. Computer scientists need guidance from medical experts to choose healthcare applications that are medically important and biologically plausible. Medical experts need guidance from computer scientists to choose prediction problems that are conceptually well-formulated and technically solvable.

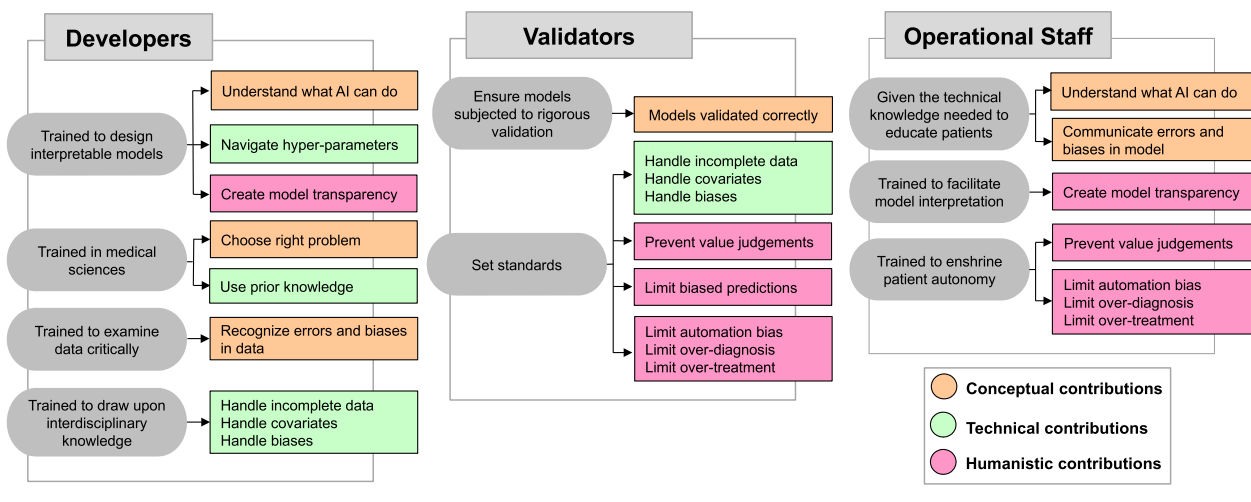
Although medical experts (often lacking in AI skills) and computer scientists (often lacking in data) may already work together on some projects, the scope of the challenges presented by medical AI requires more than just a multidisciplinary, or even an interdisciplinary, partnership. The development of AI requires a truly *transdisciplinary* approach, where each discipline broadly understands the challenges that affect the other disciplines and where the combined understanding is more than the sum of its parts. Transdisciplinary partnerships will require transdisciplinary education.

In the long term, we will need transdisciplinary training programs that teach computer science alongside health science, complete with accreditation through undergraduate and postgraduate degrees in *digital medicine*. Both sciences rely on a precise vocabulary not readily understood by outsiders, necessitating the involvement of experts who specialize in digital medicine specifically. These degrees should also require coursework in medical ethics.

### A group of validators

This group similarly needs to understand the technical details of AI systems in order to validate their performance in day-to-day work. Transdisciplinary collaboration will result in new knowledge pro-

**Table 1.** This table summarizes how accredited expert groups—developers, validators, and operational staff—can help overcome the key challenges in medical AI. Node color represents the type of challenge: conceptual (orange), technical (green), or humanistic (pink).



duction, and AI models must be constantly monitored, audited, and updated as medical knowledge advances.

### How to ready this group

In the short term, we need to apply the validation systems already available to enforce methodological rigor and safeguard patient care. This includes peer review, which should require that multiple disciplines critique the conceptual and technical design of AI systems plus their humanistic implications. We should also subject AI algorithms to the same rigorous standards we apply to evidence-based medicine<sup>29</sup>—for example, by using randomized clinical trials to evaluate model performance in terms of *clinical endpoints* and not just predictive accuracy.

In some cases, medical AI has been benchmarked against clinician performance. However, a recent systematic review identified limitations in how medical AI is evaluated. Few studies used randomization, and only 9/81 nonrandomized studies were prospective.<sup>30</sup> Moreover, studies rarely made their data and code available and often poorly adhered to reporting standards.<sup>30</sup> These findings raise concerns that medical AI devices may not be given the same level of scrutiny as other medical devices. Yet, even if they were, black-box medical AI is qualitatively different than other medical devices. Unlike say, a CT scanner, the inner workings of a deep learning model cannot be easily described: it can be impossible to know how a model made its decision or to troubleshoot why it failed.<sup>31</sup> For this reason, it is important to mind the *AI chasm* (ie, the gap between soundly designed algorithm and its meaningful clinical application)<sup>32</sup> and to ensure that AI is judged accordingly.

In the long term, we will need formal institutions that are empowered to audit whether AI has been developed and deployed responsibly, giving “AI safety” the same scrutiny we give drug safety. Since validation requires a strong understanding of systems-level healthcare operations, some have suggested the development of so-called “Turing stamps” to formally validate AI systems,<sup>33</sup> as well as a greater involvement of official regulators, like the FDA.<sup>34</sup> Whether the validators are government bodies or independent firms will ultimately depend on the

laws and customs of the jurisdiction. In either case, validation will require validators of the right kind: experts who hold a transdisciplinary perspective and understand the conceptual, technical, and humanistic challenges associated with medical AI.

### The operational staff

The operational staff includes any professional who works within the healthcare system. They provide an interface between developers and validators as well as between AI and patients. Experience shows that computer-based recommendations may be explicitly ignored by operational staff when they find the recommendations obscure or unhelpful—with potentially disastrous consequences.<sup>35</sup> Operational staff can help minimize not only the risks associated with ignoring AI, but also the risks associated with over-relying on it.

### How to ready this group

In the short term, we must take staff concerns about AI safety very seriously. This includes IT staff who oversee AI systems and monitor for privacy and data security breaches. We should also encourage clinicians to use continuing medical education allowances to attend workshops and seminars on AI and AI ethics.

In the long term, we should equip healthcare workers with literacy in AI by teaching them about the conceptual, technical, and humanistic challenges as part of the professional medical curriculum to *all medical students*. However, the intricacies of AI in medicine will additionally require opportunities for specialization. “Digital medicine” must become an *applied discipline*, complete with coursework and accreditation, in that it becomes its own specialty much like “pathology” or “radiology.” We need *digital medicine specialists*—digital doctors and digital nurses—who will work to calibrate patient expectations, listen and adapt to patient preferences and values, enshrine patient autonomy and decision-making capacity, and clearly communicate AI predictions alongside its limitations. We also need these experts to liaise with developers and validators in order to roll out new technology safely.

As noted by Park et al,<sup>36</sup> medical students do not necessarily require proficiency in the computer sciences. However, we do believe that they should acquire an adequate understanding of the conceptual, technical, and humanistic challenges associated with medical AI. This knowledge would enable them to critically examine AI in their workplace. On the other hand, *digital medicine specialists* would likely benefit from coursework on algorithm design and implementation. Like other specialties, however, most training would likely come from hands-on experience, for example through researching, validating, evaluating, and deploying medical AI systems as part of a residency training program.

## FINAL REMARKS

AI is a potentially powerful tool, but it comes with multiple challenges. In order to put this imperfect technology to good use, we need effective strategies and governance. This will require creating a new labor force which can develop, validate, and operate medical AI technologies. This in turn will require new programs to train and certify experts in digital medicine, including a new generation of “digital health professionals” who uphold AI safety in the clinical environment. Such steps will be necessary to maintain public trust in medicine through the coming AI age.

## FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## AUTHOR CONTRIBUTIONS

TPQ prepared a structured outline of the paper. TPQ, MS, SJ, SC, and VL contributed sections. All authors helped revise the manuscript and approved the final draft.

## CONFLICT OF INTEREST STATEMENT

None declared.

## REFERENCES

- Pellegrino ED, Veatch RM, Langan J. *Ethics, Trust, and the Professions: Philosophical and Cultural Aspects*. Washington, DC: Georgetown University Press; 1991.
- Powles J, Hodson H. Google DeepMind and healthcare in an age of algorithms. *Health Technol* 2017; 7 (4): 351–67.
- Howard A, Borenstein J. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Sci Eng Ethics* 2018; 24 (5): 1521–36.
- Ma X, Niu Y, Gu L, et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recogn* 2020; 107332.
- Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc* 2020; 27 (3): 491–7.
- Babel B, Buehler K, Pivonka A, Richardson B, Waldron D. Derisking machine learning and artificial intelligence. Technical report, McKinsey & Company, 2019. <https://www.mckinsey.com/business-functions/risk/our-insights/derisking-machine-learning-and-artificial-intelligence> Accessed September 1, 2020.
- Bottou L. From machine learning to machine reasoning. *Mach Learn* 2014; 94 (2): 133–49.
- Barnett GO. The computer and clinical judgment. *N Engl J Med* 1982; 307 (8): 493–4.
- Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; 17 (1): 195.
- Ramsey, P Jonsen, AR WF. *The Patient as Person: Explorations in Medical Ethics*. New Haven, CN: Yale University Press; 2002.
- Emanuel EJ, Emanuel LL. Four models of the physician-patient relationship. *JAMA* 1992; 267 (16): 2221–6.
- Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*. Oxford, England: Oxford University Press, 2001.
- Alvarez-Melis D, Jaakkola TS. Towards robust interpretability with self-explaining neural networks. In Proceedings of 32nd Conference on Neural Information Processing Systems (NeurIPS); 2018; Montréal, Canada.
- Grote T, Berens P. On the ethics of algorithmic decision-making in health-care. *J Med Ethics* 2020; 46 (3): 205–11.
- Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc* 2020; 27 (7): 1173–85.
- DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc* 2020.
- McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. *J Med Ethics* 2019; 45 (3): 156–60.
- Elwyn G, Frosch D, Thomson R, et al. Shared decision making: a model for clinical practice. *J Gen Intern Med* 2012; 27 (10): 1361–7.
- Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philos Technol* 2020; 1–23.
- Ploug T, Holm S. The right to refuse diagnostics and treatment planning by artificial intelligence. *Med Health Care Philos* 2020; 23 (1): 107–14.
- Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012; 19 (1): 121–7.
- Carter SM, Degeling C, Doust J, Barratt A. A definition and ethical evaluation of overdiagnosis. *J Med Ethics* 2016; 42 (11): 705–14.
- Sparrow R. Robots in aged care: a dystopian future? *AI Soc* 2016; 31 (4): 445–54.
- Parks JA. Lifting the burden of women’s care work: should robots replace the “human touch”? *Hypatia* 2010; 25 (1): 100–20.
- Ii P, Nicholson W. *Regulating Black-Box Medicine*. SSRN Scholarly Paper ID 2938391, Social Science Research Network, Rochester, NY, March 2017.
- Sekhar MS, Vyas N. Defensive medicine: a bane to healthcare. *Ann Med Health Sci Res* 2013; 3 (2): 295–6.
- Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018; 15 (11): e1002689.
- Umbrello AFD, S. Bellis A Value-Sensitive Design Approach to Intelligent Agents. SSRN Scholarly Paper ID 3105597, Social Science Research Network, Rochester, NY, 2018.
- Evans D. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *J Clin Nurs* 2003; 12 (1): 77–84.
- Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; 368.
- Gunning D, Aha D. DARPA’s Explainable Artificial Intelligence (XAI) program. *AIMag* 2019; 40 (2): 44–58.
- Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *Npj Digit Med* 2018; 1 (1): 1–3.
- Dalton-Brown S. The ethics of medical AI and the physician-patient relationship. *Camb Q Healthc Ethics* 2020; 29 (1): 115–21.
- FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): Discussion Paper, 2019. <https://www.fda.gov/media/122535/download> Accessed September 1, 2020.
- Leveson NG, Turner CS. An investigation of the Therac-25 accidents. *Computer* 1993; 26 (7): 18–41.
- Park SH, Do K-H, Kim S, Park JH, Lim Y-S. What should medical students know about artificial intelligence in medicine? *J Educ Eval Health Prof* 2019; 16: 18.