Published in final edited form as:

PET Clin. 2022 January ; 17(1): 1–12. doi:10.1016/j.cpet.2021.09.007.

Trustworthy Artificial Intelligence in Medical Imaging

Navid Hasani[1], **Michael A Morris**[2], **Arman Rhamim**[3], **Ronald M Summers**[4], **Elizabeth Jones**[4], **Eliot Siegel**[5], **Babak Saboury**[6]

[1]Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health (NIH), 9000 Rockville Pike, Building 10, Room 1C455, Bethesda, MD 20892, USA; University of Queensland Faculty of Medicine, Ochsner Clinical School, New Orleans, LA 70121, USA.

[2]Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health (NIH), 9000 Rockville Pike, Building 10, Room 1C455, Bethesda, MD 20892, USA; Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore Country, Baltimore, MD, USA.

[3]Department of Radiology, BC Cancer Research Institute, University of British Columbia, 675 West 10th Avenue, Vancouver, British Columbia, V5Z 1L3, Canada; Department of Physics, BC cancer Research Institute, University of British Columbia, Vancouver, British Columbia, Canada.

[4]Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health (NIH), 9000 Rockville Pike, Building 10, Room 1C455, Bethesda, MD 20892, USA.

[5]Department of Radiology, BC Cancer Research Institute, University of British Columbia, 675 West 10th Avenue, Vancouver, British Columbia, V5Z 1L3, Canada.

[6]Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health (NIH), 9000 Rockville Pike, Building 10, Room 1C455, Bethesda, MD 20892, USA; Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore Country, Baltimore, MD, USA; Department of Radiology and Nuclear Medicine, University of Maryland Medical Center, 655 W. Baltimore Street, Baltimore, MD 21201, USA; Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA, USA.

Abstract

Trust in artificial intelligence (AI) by society and the development of trustworthy AI systems and ecosystems are critical for the progress and implementation of AI technology in medicine. With the growing use of AI in a variety of medical and imaging applications, it is more vital than ever to make these systems dependable and trustworthy. Fourteen core principles are considered in this article aiming to move the needle more closely to systems that are accurate, resilient, fair, explainable, safe, and transparent: toward trustworthy AI.

Keywords

Ethics of AI; Machine learning; Trustworthiness; Trustworthy artificial intelligence

babak.saboury@nih.gov.

Introduction

The question is no longer whether artificial intelligence (AI) will impact the future of medicine but instead “by whom, how, where, and when this beneficial or harmful impact will be felt.” [1] The rate of development of new and enhanced AI-based technologies is accelerating and permeating every industry. AI has enormous potential to improve human life and the environment around us; however, we must tread carefully ahead in order to realize the opportunities it provides and avoid potential pitfalls.

Acknowledgment of the current and future benefits of AI systems in health care safety, quality, [2] equity, and access [3] is the first step toward developing clear plans to harness its potential. In medical imaging, AI has aided and will be aiding physicians in evaluation of disease progression, prediction, and/or assessment of treatment effectiveness, and tracking disease patterns over time as discussed by Hasani and colleagues in “ Artificial Intelligence in Lymphoma PET Imaging: A Scoping Review (Current Trends and Future Directions),” in this issue. AI also plays an important role in improving the effectiveness of imaging workflow and efficiency of time-consuming tasks such as segmentation. [4, 5, 6] Yousefirizi and colleagues evaluated the role of AI in segmentation in oncological PET imaging. [7]

At the same time, AI can pose certain risks and a slew of unexpected ethical, [8] legal, [9] and societal [10] challenges that, if not addressed properly, may substantially limit its value. We are becoming increasingly aware that AI systems might be fragile. [11] Graffiti on a stop sign may trick a machine learning (ML)-based classification system into not identifying a stop sign. [12] Additional noise in an picture of a benign skin lesion tricks an AI classification system into identifying the lesion as cancerous. [13] A small section of an image of a cat has been classified as guacamole with 100% confidence by a Google AI-based image recognition algorithm. [14]

So, how can we deliver on the promise of AI’s advantages while also dealing with circumstances that have life-or-death repercussions for individuals in medical settings? How can we develop “reliable AI”?

AI progression has been hindered due to limitations in computer performance the complicated nature of AI research including exaggerated claims, confusion, and issues of public trust. This necessitates tight collaboration among scientists at various phases of translational research. [15, 16] One topic that continues to be controversial is the trustworthiness of AI. Trust is an important element in the implementation of AI medical devices (AIMDs) into routine practice. One measure of a tool’s “trustworthiness” is the desire of physicians and patients to rely on it in a dangerous scenario. [17]

The aviation sector has served as a paragon for safety of its passengers and often inspires efforts to improve patient safety and reduction of medical errors. [18] The recent tragic losses of two Boeing 737 MAX aircraft can teach us lessons on AI systems and how they may be improved as deployed in medical imaging. [19] Such performance failures highlighted the fact that an AI system’s output is only as good as its inputs, and therefore the correctness of AI input data is as crucial as the AI’s ability to interpret those inputs. The quality assurance and control mechanisms must encompass single or a few isolated algorithms as well as the entire system. [20]

There are presently more than 70 frameworks and lists of AI ethical principles. [21, 22] The abundance of such guidelines creates inconsistency and confusion among stakeholders over the most acceptable document. [23] Although the trustworthiness of AI can be an element of ethical principles, not everything related to trustworthiness is a matter of ethics. Specifically, some of the concerns surrounding trustworthiness that are more practical components of clinical practice implementation cannot be addressed in ethical standards and frameworks.

Thus, in this article, we highlight the importance of addressing trustworthiness in the era of medical AI devices, and we suggest a set of essential, but not exhaustive, requirements for an AI system to be considered trustworthy.

Trust and trustworthiness from the theoretic standpoint

Human interaction is predicated on trust. The entire fabric of our daily lives, of our social order, is based on trust, [24 (p443)] and lack of it would paralyze societies and individuals by inaction. Humans are wired with an innate need to trust and to be trusted by those with whom they engage. [25] When trust is misplaced or abused, the trustor may incur significant costs [26]; hence, trusting entails taking a chance and a willingness to be vulnerable. [27, 28] In medicine, patients put trust in their physicians and health care providers when the cost may be the difference between life, death, or disability.

Trust is an attitude we have toward people devices, or systems (ie, an AI-based medical software or device) that we believe are trustworthy. Trustworthiness on the other hand is a characteristic, not an attitude. This dichotomy refers to the notion that someone trustworthy may or may not be trusted, and that someone who is trusting may trust someone who is not trustworthy. Trust and trustworthiness are thus distinct notions. In an ideal world, what we trust will also be trustworthy, and those trustworthy will be trusted. To trust someone or something means to (1) be vulnerable to what the trustee does to the extent that interests are entangled with the trustee’s performance; (2) rely on the individual or technology to be competent to accomplish the intended goal; and (3) rely on them to be willing to perform the intended task. [17]

Following Annette Baier, [29] a widely shared assumption among philosophers is that to trust someone or something is to rely on them to deliver what was expected. [17, 30] However, trust is not mere reliance, as a violation of that trust leads to a sense of betrayal rather than mere disappointment. [31] By trusting, we run the danger of losing precious things we entrust to others, in this case, our health. [17] Karen Jones [32] proposes a different perspective on trust, arguing that there is an emotional component to trust, which means that the attitude central to trust is not simply belief, but also a perceived optimism toward the proposition that the trustee would do what they are trusted to do for the right reasons.

So, *why do we need trust* in society and health care when formal constructs, such as contracts, instructions, and standard operating procedures, are available? These devices and contracts have limitations and are insufficient considering the complex and ever-changing needs of the world, societies, and, on a smaller scale, health care systems. As suggested by the findings of Giddens, [33] the need for trustworthiness does not arise from a lack of driving force but rather a lack of complete information. In addition, a fundamental core of contracts between parties is the element of trust in the counterparties and those engines of enforcement (ie, Food and Drug Administration [FDA], health systems). According to Adam Smith's theory of moral sentiments, people are linked by strong relationships of sympathy, empathy, and trust, and it is on top of this bond that markets and systems within a society may exist. [34]

To address the issue of trust, we must deal with two elements of uncertainty and vulnerability. We can either deal with uncertainty by acquiring more information or by managing the vulnerabilities, which means finding ways to mitigate harm and future actions of the trustee (in this case AI).

Importance of trustworthiness in artificial intelligence–enabled medicine: dissemination and implementation science

Trust has been at the heart of the patient-caregiver relationship from humankind's earliest forays into health care, when shamans, priests, and medicine practitioners ministered to the sick. People choose to put themselves in the hands of others in their most vulnerable moments, trusting, or at least believing, that they would benefit and be relieved.

Although there is now improved regulation surrounding many medical claims, patient's trust is equally needed in today's scientific and technological environment. The rapid progress in medicine over the past 50 years, especially the exponential increases in the past 25 years, opened possibilities that could not have been conceived a few generations earlier.

AI is advancing at a tremendous speed, with new avenues for its routine application in preclinical, clinical, and administrative health care, as well as promising evidence of its benefits to existing practice. However, these systems are complicated and opaque. Judging and interpreting their outcomes as fair and trustworthy is challenging, and they have shown to be vulnerable to major errors. For example, "heatmaps" corresponding to components of an image that are most important in the decision-making process of an algorithm have demonstrated that AI frequently pays attention to parts of an image that are irrelevant or might be called out as "cheating" by a human (eg, learning the hospital marker and using that knowledge to "predict" pneumonia or using the presence of chest drains to "diagnose" pneumothorax). These kinds of incidents may well be harmful not only to the adoption of AI in medical care, but also to general patient trust in medicine and the technology used within this field.

There are several levels of trust depending on the degree of automation and the risk associated with the work performed. With increased automation and risk involved with the task, a higher level of trust is required. Trusting an algorithm to segment the kidney as a preliminary task for evaluation by a radiologist falls on one end of this spectrum, whereas trusting the algorithm to identify cancer and initiate chemotherapy falls on the other. Therefore, AI-based medical imaging systems can be classified into 5 categories based on their degree of automation, similar to the categorization set forth by Society of Automation Engineers (SAE) for automation of vehicles. [35, 36] This 6-layered trust model offers a novel perspective through which the heterogeneity of trust in AIMDs can be realized (Table 1).

Furthermore, the distinction between high-risk and low-risk computer-aided device (CAD) is reflected in the law. The FDA distinguishes between two types of CAD used in medical imaging: computer-aided detection (CADe) and computer-aided diagnosis (CADx). [37] The agency distinguishes between CADe, which is designed to simply highlight regions of interest, and CADx, which shows the likelihood of the disease's presence or specifies a disease type. [38] Because CADx presents a greater risk it may be regulated more stringently. Therefore, CADs should adhere to the regulatory and trust standards that are developed based on their category and the risks associated with their task.

There are limited recognized standards or methods to manage and test medical AI systems. [39] It has also been documented that these systems can operate unjustly, resulting in dangerous consequences. Unprepared and inequitable AI adoption and general application in medical services, on the other hand, may bring new obstacles, potentially triggering a chain of skepticism, distrust, criticism, budget reduction, and, ultimately, the third winter of AI. Therefore, appropriate implementation and dissemination of AI in health care necessitate *trustworthy applications*.

Trust is a challenging subject that has inspired several academic arguments in recent years. The conceptualization of what makes AI trustworthy, as of today, remains ambiguous and highly debated in research and practice. [40] To address this need, frameworks and guidelines for *ethical* AI, [10, 41] beneficial AI, [42] and trustworthy AI (TAI) [39, 40, 43] have been set forth to advance AI while minimizing the potential risks associated with it.

The technology sector has been a leader in seeking the implementation of TAI. Microsoft emphasized the significance of trustworthy software in its January 2002 “Trustworthy Computing” message to personnel, users, shareholders, and the rest of the information technology sector. [44] According to an internal Microsoft white paper, security, privacy, dependability, and commercial integrity are the four pillars around which trust is founded. [44]

Others have proposed using the Formal Methods approach of computer science for achieving TAI. [45, 46] In this approach, TAI requires a shift away from conventional computer systems' deterministic approach and toward a more probabilistic nature. [47] To create end-user trust, this method uses data science and formal verification in which properties are established over a wide domain for all inputs or behaviors of a particular distributed or concurrent system. [46] On the other hand, the verification mechanism discovers a counterexample, such as an input value for which the program delivers an inaccurate outcome that does not meet the necessary characteristic. This process can provide useful insights for further improving the system. Formal verification provides the advantage of obviating the requirement to test each input value or action one by one, which may be a challenging task for vast (or infinite) state spaces. Similar methods for the development of AIMDs are necessary.

Key requirements to promote trustworthy artificial intelligence systems

When we discuss the topic of the trustworthiness of AI in medicine, it is entirely from the perspective of the patient. For AI to be trustworthy, it needs to be implemented through generalized trust and relational trust. Generalized trustworthiness will encourage the patient to consent to or seek AI-augmented medical care while a relational trust will be developed over time and enables maintenance of trustworthiness after the patient's initial encounter with an AIMD.

Several components can promote the trustworthiness of the AIMD and all processes and individuals who are a part of the AI Ecosystem. In what follows, we list 14 core principles and requirements toward TAI; these are listed in Fig. 1 and elaborated in the following sections.

Transparency

Transparency promotes informed decision making and is a key component in building trustworthy AI systems. As a result, "black box" AI systems that do not place a strong focus on various indicators of transparency (data use transparency, clear disclosures, traceability, auditability, and understandability) should be avoided in clinical settings as much as possible.

Conceptually there are 2 types of opacity in medical AI systems that can influence trustworthiness: (1) lack of transparency, and (2) epistemic opacity, [48] which we describe next.

Data transparency indicates that data subjects are aware of how their health records are used for AI system profiling and decision-making processes. In this regard, AIMD's public confidence and integrity may be jeopardized. Although transparency is essential, one major concern for developers is the risk of harmful usage or patient privacy violations. [49] Vendors should provide the characteristics of the training and testing data used for validation, as well as how an AI system's influence is verified for the labeled claim (purpose, criteria, and limits).

When using a decision support system, a clear distinction must be made on what is conveyed by the AI and the information communicated by the clinician. AI systems should have mechanisms for recording and identifying whether data, AI models, or rules were utilized to generate certain AI outcomes (auditability and traceability). To provide a mechanism to assess and challenge AI system outputs, the influence of the input on the output must be reported in such a manner that medical professionals and patients can understand the relationship.

Epistemic opacity refers to the inability of developers or users (health care providers) to understand how an AI system arrives at a certain outcome. Autonomous systems engage in actions that are difficult to comprehend or predict from users' perspectives, although there is a plethora of tools to probe the algorithm. For instance, Zeiler and Fergus [50] created a visualization approach that provides insight into the function of intermediate feature layers and classifier operations in a convolutional network model. Yet, reducing epistemic opacity and understanding internal rules used in the decision-making processes of evolving AI systems continues to be a challenge. This aspect of AI's black box nature can complicate quality assurance and interpretability or restrict clinician and patient input in the decision-making process. [51]

Explainability

The issue of explanatory opacity refers to the inability to understand and elucidate how and why the system made a particular decision. [48] This differs from epistemic opacity because not only do users need to comprehend technical aspects of decision making, but they should also be able to explain them in plain terms. [10] But does one need a deep grasp of data science, physics, statistics, and epidemiology to understand and describe the residual bias and confounding that may exist in AIMDs? At the very least, there must be enough training materials and disclaimers for health care workers on how to use the system properly.

Amann and colleagues [52] conducted an ethics-based assessment utilizing the Principles of Biomedical Ethics (beneficence, no-maleficence, autonomy, and justice) to establish the necessity for explainability in AI systems used in healthcare. Reportedly, to maximize the well-being of patients (beneficence) and prevent harm (nonmaleficence) as well as trustworthiness, physicians should generally understand and be able to explain the AI decision-making processes. The issue of explainability may not be equally important in different industries; the stakes are far greater in the health care industry, as explainability allows physicians to assess a system's suggestions based on their clinical judgment and expertise. Thus, an explainable system would empower physicians and patients, promoting autonomy, trust, and informed decision-making. Otherwise, parties or physicians may not fully trust the AIMD suggestions and outcomes, especially when their own opinion is different from that of AI. [52]

Transparent mechanisms of risk management and accountability should be in place in case of any adverse events. According to the principles of safety-critical systems, vendors and physicians should be accountable for their claims and the extent that AIMD is involved in patient care. For an AI system to be just, clinicians and operators need to be able to explain and understand the system, as they are ultimately accountable and therefore responsible for addressing if an AI system is for some reason unjust or biased. [53]

Clinicians around the world representing diagnostic radiology and nuclear medicine must advise the scientific community and industry to commit to moving toward "explainable AI" as much as possible. Necessary resources should be allocated to prioritize this aim as a component of AIMD products. [54, 55, 56, 57] One strategy to achieve this is by creating a second AI system that tries to explain and analyze what the first AIMD decision was based on. This AI may not be able to explain how the AIMD came to the decision, but it can show what factors were weighed. Overall, a concerted research effort is needed in the frontier of explainable AI for medical applications.

Technical Robustness

Readily available data can be used to train and test the model, whereas unseen data are data that the model must (or is expected to) operate on without having previous encounters with it. The primary aim of a model is to be capable of function and analyzing novel inputs, often with some level of certainty, based on the data it was trained and tested on.

A key aspects of AI systems' robustness is their ability to reproduce the claimed performance accurately and reliably with a certain degree of confidence reported to the user (ie, the physician). Additionally, the system must be generalizable to the claimed user population. These aspects of AI's technical robustness must be regularly monitored through various standardized quality control measures.

Safety and Security

AI-based medical devices and systems in health care must incorporate strategies to minimize any potential harm due security breaches according to the principles of safety-critical systems. [58, 59] As such, AIMDs must comply with all existing cybersecurity requirements, and their inherent vulnerabilities, such as model evasion or data poisoning, should be thoroughly evaluated prior to clinical deployment. Health systems and vendors must be transparent regarding the measures taken to mitigate and resolve potential AI vulnerabilities. [60]

Predetermined Change Control Plan

Machine learning systems can be highly iterative and adaptive which may result in product performance improvement or changes over time. AI developers and vendors should anticipate such alterations and create appropriate change control plans accordingly (ie, developing secondary AI-based control system to monitors and reports the changes of the original AI system). Strategies for controlling performance quality and assessing the robustness and safety of the updated AI system should be clearly anticipated and protocoled. Recommendations for retraining systems, performance evaluation, and procedure updates should be included in a well-documented algorithm change protocol. Such measures will enhance quality control and enable organizational and regulatory oversight. [61]

Diversity, Bias-Awareness, Nondiscrimination, and Fairness

Although AI has many benefits for humanity, one of the most serious issues arising from its increased usage is its potential to entrench and perpetuate prejudice and discrimination. [62] The performance of AI medical devices can be impacted if the input training or testing data is flawed (ie, incomplete or skewed data) or if the performance monitoring methods are suboptimal. [13, 63] These factors may result in AI-enabled biases, subsequent prejudices, and unintentional discrimination against a group of patients. As a result, in accordance with the Universal Design Principles, any potential bias that could lead to prejudice should be carefully addressed and eradicated from AI systems during the conceptualization and deployment stages. [64]

Socially created biases are common in current AI-based systems in health care. [53] Another form of bias, in addition to bias in training input data, is an overemphasis of particular features (ie, skin color or locality) by AI model developers.

Developers must openly document any efforts made to minimize, and thereby quantify, unfair effects in their models. Second, regulated firms must develop specific, good faith justifications for the models they eventually embrace. [65] AI system performance should be generalizable to all patients suffering from a particular condition regardless of extraneous personal characteristics. [56, 66] Patients who are underrepresented or suffer from rare diseases should not be excluded from AI systems development or evaluation [see Hasani and colleagues' article, " Artificial Intelligence in Medical Imaging and its Impact on the Rare Disease Community: Threats, Challenges, and Opportunities," in this issue]. Appropriate validation testing on standardized sets that include a diverse patient population, including rare or unusual presentations of disease, is critical to evaluate the presence of bias in results regardless of the training data used. In recent months, the US Federal Trade Commission has shown an increased interest in AI fairness, openly suggesting that the agency should broaden its monitoring of potentially biased AI. [67]

AI solutions should be created with clinical settings in mind, as well as designed and implemented to accommodate various cultural and organizational norms. Furthermore, such solutions should consider extending access and including those with disabilities or rare diseases.

Human Agency

AI systems in clinical settings should not only enhance the workflow of the care team but also further enable the patient and the care team to make informed decisions and clearly communicate those decisions with others, as in Freidman's fundamental theorem of informatics. [14, 68] This will further empower the autonomy of both parties while limiting potential for automation bias. Patients and physicians should understand the extent to which AIMD is integrated in care delivery and the scope of physician's oversight.

Oversight

Appropriate supervision techniques should be used, which may be accomplished using methodologies such as "humans in the loop," "humans on the loop," and "humans in charge." [39] Such approaches will ensure human values are being considered. According to the World Health Organization, AI systems should be thoroughly regulated post-market by independent professional credentialing authorities in a way similar to the way in which medical practitioners are certified and recertified. [69] The approval and auditing processes should not only consider the level of the risk associated with the AI claim, but also the level of learning [70] (supervised or unsupervised) and characteristics such as explainability, transparency, and accountability. AI systems should be monitored and categorized according to their degree of automation and autonomy. Similar to the AI categories set forth by SAE, medical imaging can categorize AIMDs into categories such as (1) no automation, (2) physician assistance, (3) partial automation, (4) conditional automation, (5) high automation, and (6) full automation. [71]

The investigation and validation process should include the AI technology's assumptions, operating procedures, data characteristics, and output decisions. Regular tests and assessments should be conducted in a transparent manner and with sufficient breadth to account for variances in algorithm performance based on race, ethnic origin, gender, age, and other important human traits. Such testing and assessments should be subjected to rigorous, independent monitoring to verify their safety and effectiveness. Medical institutions, hospital systems, and other related organizations should frequently disclose information regarding how choices concerning the deployment of AI technologies were made and how the technology will be assessed on a periodic basis. Its applications, recognized limits, and degree of involvement in decisions should also be considered, all of which can also permit third-party audits and supervision.

Stakeholder Engagement

A comprehensive collaboration and coordination system involving all stakeholders which may include patients, clinicians, insurers, health systems, research investigators, manufacturers, and regulatory agencies is of paramount importance if our goal is to integrate sustainable and trustworthy AI systems in patient care. Active engagement of all stakeholders will enable and mediate transparency, inclusiveness, trust, and accountability, all of which further enhance long-term sustainability of AI systems in clinical practice. Continuous engagement allows stakeholders to provide regular feedback and voice any potential concern at each stage of design, development, and implementation.

Sustainability of Societal Well-being

Deployment of AI into the health care system must be with careful consideration of its potential impact on the social well-being, trust in the health care system, and the physician-patient relationship. [72, 73] As such, AI solutions should strive to enhance social interaction within the care team and between the physician and the patient. To achieve this goal, all health care providers who interact with the AI or are impacted by AI's implementation into their workflow should be given an opportunity have an active voice throughout the life cycle of the AI system. Professional societies and health care training programs should take necessary measures to ensure AI related skills and knowledge is incorporated into the education curricula and board examinations of appropriate health care workers.

Privacy and Data Governance

In 2020, there were 29 million health care records breached, [74] demonstrating the widespread theft of patient electronic Protected Health Information (PHI), social security numbers, and private financial information. Deployment of not fully secured AI systems to this environment could pose a risk. However, AI can also help health systems safeguard against cyber threats. AI should have procedures in place to ensure that patient data are kept secure and private. Safeguarding devices at all stages is critical, especially if intercepting or modifying data may affect device functionality. Additional AI can be added to AIMDs for cybersecurity purposes. [75] Cybersecurity AI has the potential to not only distinguish between regular network traffic and harmful hacker activity but can also respond quickly to stop the attack from spreading. Only the bare minimum of personal information should be used (data minimization). A declaration on the methods used to accomplish privacy-by-design, such as encryption, pseudo anonymization, aggregation, and de-identification should be included. [76] To achieve this goal, standardized protocols and guidelines should be recognized and routinely used to safeguard patient privacy and data handling. [77], [78]

Accountability

The model's capacity to justify its judgments to the system's users is referred to as model accountability. This entails accepting responsibility for all decisions taken, regardless of whether they were correct or resulted in errors or unanticipated outcomes. Mechanisms for guaranteeing accountability and redress should be in place when adverse events occur. AI medical device manufacturers must be held liable for the claims made by their AI systems. Additionally, clinicians and health systems should be held accountable for the proper integration and deployment of the AI technology into the workflow and delivery of medical services. According to the principles of safety-critical systems, the capacity to independently audit the root cause of a failure in an AI system is vital. Individuals or groups who report real concerns must be protected in accordance with risk management standards.

Supportive Context of Implementation

Developer protection, customer protection, and legal protection are all important considerations. The "supporting context of implementation" is critical for establishing confidence in the AI ecosystem. Patients should be able to seek legal advice and representation if necessary. This strengthens their feeling of agency, and as a consequence, individuals may be more responsive to the innovative intervention, knowing they would be protected in the event of an unforeseen event. This significantly speeds the spread of novel technology.

To prevent this technology from dying prematurely while it is still in its infancy, governments may adopt methods similar to those utilized decades ago to preserve the vaccine industry. In the 1980s, there were a slew of lawsuits filed against vaccine manufacturers. Because of the general anticipated risks associated with lawsuits, there was widespread anxiety that vaccine developers would leave the field. To entice developers, the US government established a federally regulated financial resource, funded by vaccine taxes, to award judgments for injuries caused by certain adverse responses. Similar supportive strategies for appropriate AI use in health care could not only enforce regulations but also foster innovation and advancement toward safe AI deployment.

Promoting Systems for Experimenting Trustworthiness Properties

Health systems, AI developers, and other key stakeholders must collaborate to improve their grasp of psychological, sociologic, and cultural trustworthiness properties. Cultures across the globe often have a variety of value systems and fundamental beliefs that may contribute to the diversity value systems that deem an AIMD trustworthy. Thus, we must discern and implement trustworthy qualities in AI systems in order for them to function across cultural and socioeconomic differences.

Summary

Trust in AI by society and the development of trustworthy AI systems and ecosystems are critical for the progress and implementation of AI technology in medicine. With the growing use of AI in a variety of medical and imaging applications, it is more vital than ever to make these systems dependable and trustworthy. Fourteen core principles are considered in this article aiming to move the needle more closely to systems that are accurate, resilient, fair, explainable, safe, and transparent—toward *trustworthy AI*.

Acknowledgments and Disclosures

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), Clinical Center. The opinions expressed in this publication are the author's own and do not reflect the view of NIH, the Department of Health and Human Services, or the US government.

References

1. Floridi L, Cowls J, Beltrametti M, et al. : AI4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 2018; 28: pp. 689–707.
2. Davenport T, Kalakota R: The potential for artificial intelligence in healthcare. Future Healthc J 2019; 6: pp. 94–98.
3. Matheny M, Thadaney S, Ahmed M, et al.: Artificial intelligence in health care: the hope, the hype, the promise, the peril.2019.National Academy of MedicineWashington (DC) Available at: https://nam.edu/wp-content/uploads/2019/12/AI-in-Health-Care-PREPUB-FINAL.pdf
4. Kapoor N, Lacson R, Khorasani R: Workflow applications of artificial intelligence in radiology and an overview of available tools. J Am Coll Radiol 2020; 17: pp. 1363–1370. [PubMed: 33153540]
5. Nikpanah M, Xu Z, Jin D, et al. : A deep-learning based artificial intelligence (AI) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic MRI. Clin Imaging 2021; 77: pp. 291–298. [PubMed: 34171743]
6. Weisman AJ, Kieler MW, Perlman S, et al. : Comparison of 11 automated PET segmentation methods in lymphoma. Phys Med Biol 2020; 65: pp. 235019. [PubMed: 32906088]
7. Yousefirizi F, Jha AK, Brosch-Lenz J, et al. : Toward High-Throughput Artificial Intelligence-Based Segmentation in Oncological PET Imaging. PET Clin 2021 Oct; 16: pp. 577–596. [PubMed: 34537131]
8. Char DS, Abràmoff MD, Feudtner C: Identifying ethical considerations for machine learning healthcare applications. Am J Bioeth 2020; 20: pp. 7–17.
9. Ganapathy K: Artificial intelligence and healthcare regulatory and legal concerns. TMT 2021;
10. Geis JR, Brady AP, Wu CC, et al. : Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. Radiology 2019; 293: pp. 436–440. [PubMed: 31573399]
11. Zou J, Schiebinger L: AI can be sexist and racist — it's time to make it fair. Nature 2018; 559: pp. 324–326. [PubMed: 30018439]
12. Eykholt K, Evtimov I, Fernandes E, et al. : Robust physical-world attacks on deep learning models. arXiv [csCR]. Available at: http://arxiv.org/abs/1707.08945 Accessed September 26, 2021
13. Finlayson SG, Bowers JD, Ito J, et al. : Adversarial attacks on medical machine learning. Science 2019; 363: pp. 1287–1289. [PubMed: 30898923]

14. Brown M: A Google algorithm was 100 percent sure that a photo of a cat was guacamole. Availabel at: https://www.inverse.com/article/56914-a-google-algorithm-was-100-percent-sure-that-a-photo-of-a-cat-was-guacamole Accessed September 12, 2021

15. Kaul V, Enslin S, Gross SA: History of artificial intelligence in medicine. Gastrointest Endosc 2020; 92: pp. 807–812. [PubMed: 32565184]

16. Toosi A, Bottino AG, Saboury B, et al. : A brief history of AI: how to prevent another winter (a critical review). PET Clin 2021; 16: pp. 449–469. [PubMed: 34537126]

17. McLeod C: Trust.Zalta E.N.The Stanford encyclopedia of philosophy. Fall 2020. Metaphysics Research Lab.2020.Stanford UniversityCA, USA: Availabel at: https://plato.stanford.edu/archives/fall2020/entries/trust/

18. Helmreich RL: On error management: lessons from aviation. BMJ 2000; 320: pp. 781–785. [PubMed: 10720367]

19. Federal Aviation Administration: Summary of the FAA's review of the Boeing 737 MAX. Available at: https://www.faa.gov/foia/electronic_reading_room/boeing_reading_room/media/737_RTS_Summary.pdf Accessed September 15, 2021

20. Mongan J, Kohli M: Artificial intelligence and human life: five lessons for radiology from the 737 MAX Disasters. Radiol Artif Intell 2020; 2: pp. e190111. [PubMed: 33937819]

21. AI ethics guidelines global inventory. Available at: https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/ Accessed August 23, 2021

22. Winfield A, Profile V my C.: Alan Winfield's Web Log. Available at:

23. Floridi L: Establishing the rules for building trustworthy AI. Nat Mach Intell 2019; Available at: https://www-nature-com.ezproxy.library.uq.edu.au/articles/s42256-019-0055-y

24. Rotter JB: Generalized expectancies for interpersonal trust. Am Psychol 1971; 26: pp. 443–452.

25. Hawley K: Trust: a very short introduction.2012.OUP OxfordOxford, UK Available at: https://play.google.com/store/books/details?id=8KTrSrCfhkIC

26. Covey SR, Merrill RR: The speed of trust: the one thing that changes everything.2008.Simon and SchusterNew York, NY Available at: https://play.google.com/store/books/details?id=31Qe_e61Y10C

27. Kramer RM: Trust and distrust in organizations: emerging perspectives, enduring questions. Annu Rev Psychol 1999; 50: pp. 569–598. [PubMed: 15012464]

28. Misztal B: Trust in modern societies: the search for the bases of social order.2013.John Wiley & SonsHoboken, NJ, USA Available at: https://play.google.com/store/books/details?id=zfIdAAAAQBAJ

29. Baier A: Trust and antitrust. Ethics 1986; 96: pp. 231–260. Available at: http://www.jstor.org.ezproxy.library.uq.edu.au/stable/2381376

30. Hawley K: Trust, distrust and commitment. Nous 2014; 48: pp. 1–20.

31. Goldberg SC: Trust and reliance 1.Oxfordshire UK:Taylor Francis IncThe Routledge handbook of trust and philosophy.2020.Routledgepp. 97–108.

32. Jones K: Trust as an affective attitude. Ethics 1996; 107: pp. 4–25. Available at: http://www.jstor.org.ezproxy.library.uq.edu.au/stable/2382241

33. Giddens A: The consequences of modernity.1990.Stanford University Press Available at: https://play.google.com/store/books/details?id=oU99QgAACAAJ

34. Evensky J: Adam Smith's theory of moral sentiments: on morals and why they matter to a liberal society of free people and free markets. J Econ Perspect 2005; 19: pp. 109–130.

35. SAE International releases updated visual chart for its "levels of driving automation" standard for self-driving vehicles. Available at: https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles Accessed September 17, 2021

36. Jaremko JL, Azar M, Bromwich R, et al. : Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. Can Assoc Radiol J 2019 May; 70: pp. 107–118. [PubMed: 30962048]

37. Center for Devices, Radiological Health: CADe devices applied to radiology images and device data - 510(k) Sub. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/computer-assisted-detection-devices-applied-radiology-images-and-radiology-device-data-premarket Accessed September 17, 2021

38. Learning from Experience: FDA's treatment of machine learning. Available at: https://www.mobihealthnews.com/content/learning-experience-fda%E2%80%99s-treatment-machine-learning Accessed September 17, 2021

39. Kaur D, Uslu S, Durresi A: Requirements for trustworthy artificial intelligence – a review.Barolli LLi KEnokido T et al.Advances in Networked-based information systems.2021.Springer International PublishingNY, USA:pp. 105–115.

40. Thiebes S, Lins S, Sunyaev A: Trustworthy artificial intelligence. Electron Mark 2021; 31: pp. 447–464. Available at: https://link-springer-com.ezproxy.library.uq.edu.au/article/10.1007/s12525-020-00441-4

41. Kohli M, Geis R: Ethics, artificial intelligence, and radiology. J Am Coll Radiol 2018; 15: pp. 1317–1319. [PubMed: 30017625]

42. Beneficial Artificial Intelligence: Harvard Business Review.2019. Available at: https://hbr.org/podcast/2019/06/beneficial-artificial-intelligence Accessed August 24, 2021

43. Bærøe K, Miyata-Sturm A, Henden E: How to achieve trustworthy artificial intelligence for health. Bull World Health Organ 2020; 98: pp. 257–262. [PubMed: 32284649]

44. Gates B: Gates: Trustworthy Computing.. Published January 17, 2002; Available at: https://www.wired.com/2002/01/bill-gates-trustworthy-computing/. Accessed August 26, 2021

45. Vassev E: Safe artificial intelligence and formal methods.Leveraging applications of formal methods, verification and validation: foundational techniques.2016.Springer International PublishingNY, USA:pp. 704–713.

46. Trustworthy AI. Available at: https://datascience.columbia.edu/news/2020/trustworthy-ai/ Accessed August 26, 2021

47. Rajendran JJV, Sinanoglu O, Karri R. Building Trustworthy Systems Using Untrusted Components: A High-Level Synthesis Approach. IEEE Trans Very Large Scale Integr VLSI Syst. 2016;24(9):2946–2959.

48. Ferretti A, Schneider M, Blasimme A: Opening the new data protection black box. Available at: https://www.forbes.com/sites/korihale/2018/05/22/ai Accessed September 15, 2021

49. Pesapane F, Volonté C, Codari M, et al. : Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 2018; 9: pp. 745–753. [PubMed: 30112675]

50. Zeiler MD, Fergus R: Visualizing and understanding convolutional networks. arXiv [csCV]. Available at: http://arxiv.org/abs/1311.2901 Accessed September 15, 2021

51. Quinn TP, Jacobs S, Senadeera M, et al. : The three ghosts of medical AI: can the black-box present deliver? arXiv [csAI]. Available at: http://arxiv.org/abs/2012.06000 Accessed September 15, 2021

52. Amann J, Blasimme A, Vayena E, et al. : Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020; 20: pp. 310. [PubMed: 33256715]

53. Obermeyer Z, Powers B, Vogeli C, et al. : Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366: pp. 447–453. [PubMed: 31649194]

54. Mudgal KS, Das N: The ethical adoption of artificial intelligence in radiology. BJR Open 2020; 2: pp. 20190020. [PubMed: 33178959]

55. Currie G, Hawk KE, Rohren EM: Ethical principles for the application of artificial intelligence (AI) in nuclear medicine. Eur J Nucl Med Mol Imaging 2020; 47: pp. 748–752. [PubMed: 31927637]

56. Currie G, Hawk KE: Ethical and legal challenges of artificial intelligence in nuclear medicine. Semin Nucl Med 2020; 11:

57. Geis JR, Brady AP, Wu CC, et al. : Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. Can Assoc Radiol J 2019; 70: pp. 329–334. [PubMed: 31585825]

58. Knight JC. Safety-critical systems: challenges and directions. In: Proceedings of the 24th International Conference on Software Engineering. ICSE 2002. 25 May 2002:547–550. doi:10.1109/icse.2002.1007998

59. Grant ES. Requirements engineering for safety critical systems: An approach for avionic systems. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC);Oct. 14–17, 2016:991–995. doi:10.1109/CompComm.2016.7924853

60. Lathrop B The Inadequacies of the Cybersecurity Information Sharing Act of 2015 in the Age of Artificial Intelligence. Hastings LJ. 2019;71:501.

61. US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf. Available at: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf Accessed September 15, 2021

62. Zuiderveen Borgesius F: Discrimination, artificial intelligence, and algorithmic decision-making. Available at: https://pure.uva.nl/ws/files/42473478/32226549.pdf Accessed August 26, 2021

63. Fletcher RR, Nakeshimana A, Olubeko O: Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. Front Artif Intell 2020; 3: pp. 561802. [PubMed: 33981989]

64. Odukoya EJ, Kelley T, Madden B, et al. Extending “Beyond Diversity”: Culturally Responsive Universal Design Principles for Medical Education. Teach Learn Med. 2021;33(2):109–115. [PubMed: 33792455]

65. Burt A: How to fight discrimination in AI. Harvard Business Review. Available at: https://hbr.org/2020/08/how-to-fight-discrimination-in-ai Accessed August 26, 2021

66. Allen B, Dreyer K: The role of the ACR Data Science Institute in advancing health equity in radiology. J Am Coll Radiol 2019; 16: pp. 644–648. [PubMed: 30947901]

67. Aiming for truth, fairness, and equity in your company’s use of AI. Available at: https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai Accessed September 19, 2021

68. Friedman CP: A “fundamental theorem” of biomedical informatics. J Am Med Inform Assoc 2009; 16: pp. 169–170. [PubMed: 19074294]

69. Health Ethics & Governance: Ethics and governance of artificial intelligence for health. Available at: https://www.who.int/publications/i/item/9789240029200 Accessed July 1, 2021

70. Angehrn Z, Haldna L, Zandvliet AS, et al. : Artificial intelligence and machine learning applied at the point of care. Front Pharmacol 2020; 11: pp. 759. [PubMed: 32625083]

71. Driving AI adoption: what radiology can learn from self-driving vehicles. Available at: https://www.radiologytoday.net/archive/WebEx0918.shtml Accessed September 12, 2021

72. Cassell P: The Giddens reader.1993.Macmillan International Higher EducationLondon, UK Available at: https://play.google.com/store/books/details?id=0kldDwAAQBAJ

73. Harvey DL: Agency and community: a critical realist paradigm. J Theory Soc Behav 2002; 32: pp. 163–194.

74. HIPAA Journal: 2020 Healthcare data breach report: 25% increase in breaches in 2020. Available at: https://www.hipaajournal.com/2020-healthcare-data-breach-report-us/ Accessed August 26, 2021

75. AI in healthcare: protecting the systems that protect us. Wired. Available at: https://www.wired.com/brandlab/2020/04/ai-healthcare-protecting-systems-protect-us/ Accessed September 12, 2021

76. Parker W, Jaremko JL, Cicero M, et al. : Canadian Association of Radiologists White Paper on de-identification of medical imaging: part 2, practical considerations. Can Assoc Radiol J 2021; 72: pp. 25–34. [PubMed: 33140663]

77. Artificial intelligence — Overview of trustworthiness in artificial intelligence. ISO (the International Organization for Standardization). Available at: https://www.iso.org/obp/ui/ Accessed September 12, 2021

78. Ethically Aligned Design. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Available at: https://standards-ieee-org.ezproxy.library.uq.edu.au/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf. Accessed September 15, 2021.

Key points

- Trust has been at the heart of the patient-caregiver relationship from humankind’s earliest forays into health care.
- Artificial intelligence (AI) systems, rapidly emerging and increasingly used, are complicated and remain largely opaque.
- We are becoming increasingly aware that AI systems might be fragile and unjust.
- Incidences of broken trust by AI systems will be harmful not only to the adoption of AI in medical care but also to general patient trust in medicine and technologies used within this field.
- We discuss 14 core principles and key requirements to enable and promote trustworthy AI systems.

Clinics care points

- Trustworthy AI is not just based on the trustworthiness of an AI Medical Device; it encompasses the entire ecosystem of AI development, production, implementation, and oversight, as well as all the social institutes protecting the wellbeing and rights of stakeholders.
- The fourteen key concepts and standards for trustworthy AI outlined in this article are all significant, complement one another, and should be implemented and evaluated throughout the AI system's life cycle.
- Addressing the existing and future benefits of AI systems through the lens of trust, safety, quality, fairness, and access is the first step toward devising specific plans to harness the full potential of trustworthy AI in medical imaging.
- Clinicians should understand and be able to explain the AI decision-making processes in general to nurture the trustworthiness of AI and enhance patient well-being (beneficence) and prevent harm (non-maleficence).
- Broken trust by AI systems will be harmful not only to the adoption of AI in medical treatment but also to patient trust in medicine and the technology utilized in the profession.

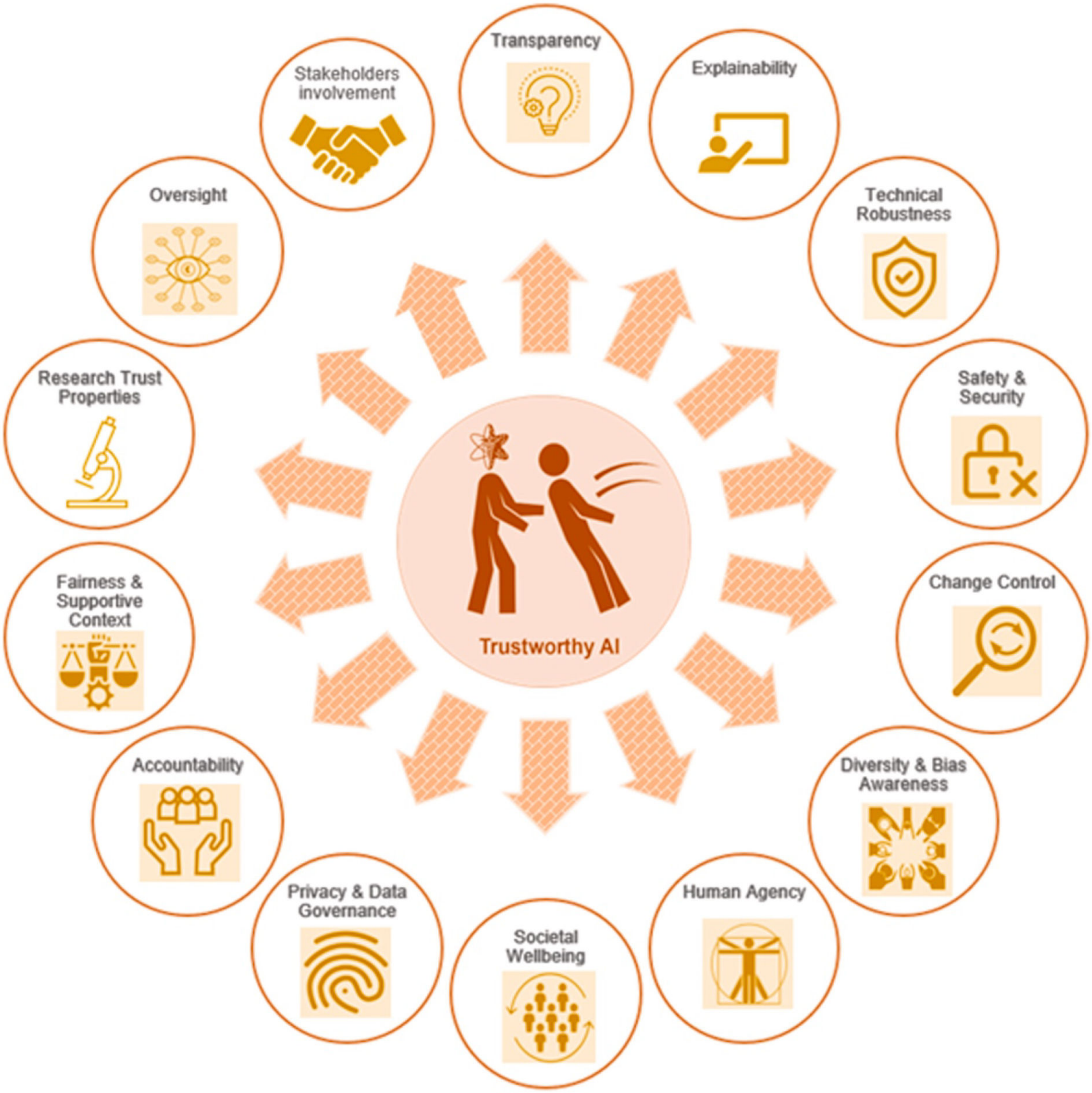

Fig. 1.

The 14 core principles and requirements for TAI: the principles are all significant, complement one another, and should be applied and assessed over the entire life cycle of the AI system.

Table 1

Six levels of automation based on the Society of Automation Engineers (SAE) model and the version appropriate to AI-based medical imaging tools

Adapted from SAE International Releases Updated Visual Chart for Its "Levels of Driving Automation" Standard for Self-Driving Vehicles. Accessed September 17, 2021. and Jaremko et al. Canadian Association of Radiologists (CAR) Artificial Intelligence Working Group. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. Can Assoc Radiol J. 2019 May;70(2):107–118. with permission.

Levels	SAE Model	Medical Imaging Version
0	**No automation** All driving tasks are carried out by humans.	**No automation** Interpretation/intervention is done solely by the radiologist.
1	**Driving Assistance** The car is equipped with a single automated system (ie, cruise control).	**Physician assistance** The radiologists are in charge of interpretation and intervention, while Al provides secondary oversight (ie, existing CAD software for mammography and lung nodules, worklist prioritization).
2	**Partial driving automation** The vehicle is capable of steering and acceleration. The human is still monitoring all tasks and has the ability to take control at any moment.	**Partial automation** The Al is in responsible for interpretation and intervention, with the radiologist providing secondary monitoring (ie, bone age prediction, chest x-ray pathology detection and report pre- population).
3	**Conditional driving automation** The vehicle is capable of detecting its surroundings. The car can do the majority of driving responsibilities, although human intervention is still necessary.	**Conditional automation** The Al is responsible for the interpretation and intervention only for specific indications, with the expectation that the radiologist will intervene if the results are positive or inconclusive (ie, automated triaging of normal cases where radiologist is expected to intervene if positive but not if negative).
4	**High driving automation** Under specific situations, the vehicle performs all driving tasks. Geofencing is essential. Human intervention is still a possibility.	**High automation** The Al is the lone interpreter/interventionist for a specific indication, with no expectation that the radiologist will intervene. Al can independently reach a differential diagnosis and care plan (ie, Al studies thyroid ultrasound and advises and/or performs a nodule biopsy).
5	**Full driving automation** The vehicle completes all driving duties under all situations. No need for human involvement or attention.	**Full automation** For all indications expected of a radiologist, the Al is the sole interpreter/ interventionist. Al can provide a differential diagnosis and make care recommendations on its own (ie, a chest x-ray request indicates "rule out pneumonia" Al reports a bone tumor with a differential diagnosis and recommendations for more imaging/consultation).

Abbreviations: AI, artificial intelligence; CAD, computer-aided design.