PriMera Scientific Engineering
Volume 3 Issue 5 November 2023
DOI: 10.56831/PSEN-03-084
ISSN: 2834-2550

PriMera Scientific Publications

Unpacking the Bias Challenges of Deep Learning in Clinical Applications: A Critical Explorer of the Impact of Training

Type: Research Article
Received: September 19, 2023
Published: October 20, 2023

Citation:
Fred Wu., et al. "Unpacking the Bias Challenges of Deep Learning in Clinical Applications: A Critical Explorer of the Impact of Training". PriMera Scientific Engineering 3.5 (2023): 03-10.

Fred Wu[1]* and Colmenares-Diaz Eduardo[2]

[1]*West Virginia State University, Institute, USA*

[2]*Midwestern State University, Wichita Falls, USA*

***Corresponding Author:** Fred Wu, West Virginia State University, Institute, USA.

Abstract

The field of artificial intelligence (AI) in healthcare is rapidly expanding worldwide, with successful clinical applications in orthopedic disease analysis and multidisciplinary practice. Computer vision-assisted image analysis has several U.S. Food and Drug Administration-approved uses. Recent techniques with emerging clinical utility include whole blood multicancer detection from deep sequencing, virtual biopsies, and natural language processing to infer health trajectories from medical notes. Advanced clinical decision support systems that combine genomics and clinomics are also gaining popularity. Machine/deep learning devices have proliferated, especially for data mining and image analysis, but pose significant challenges to the utility of AI in clinical applications. Legal and ethical questions inevitably arise. This paper proposes a training bias model and training principles to address potential harm to patients and adverse effects on society caused by AI.

Introduction

Artificial Intelligence (AI) has been referred to as the "fourth industrial revolution" due to its transformative impact on society worldwide. In essence, AI is a field that combines computer science, engineering, and related disciplines to develop machines capable of performing tasks that require intelligence in humans. These tasks can include anything from recognizing speech and visual images to learning from and adapting to new information. To achieve these goals, AI relies on various techniques, such as machine learning, which enables algorithms to solve problems and make predictions based on large amounts of data without explicit programming. Deep learning, a subset of machine learning, goes further by using multiple layers of artificial neural networks to address complex issues from unstructured data, similar to the functioning of the human brain. With potential economic gains estimated at $15.7 trillion by 2030, investments in AI are growing rapidly. The advent of AI applications has the potential to become one of the most significant developments in the history of medicine, with implications for all medical specialties and healthcare service users. According to the World Health Organization (WHO), the market for AI in the medical field is expected to reach $150 billion by 2026. Among the areas projected to benefit are robot-assisted surgery at $40 billion, virtual nursing

assistants at $20 billion, administrative workflow assistance at $18 billion, fraud detection at $17 billion, reduction of dosage error at $16 billion, connected machines at $14 billion, clinical trial participant identifier at $13 billion, preliminary diagnosis at $5 billion, automated image diagnoses at $3 billion, and cybersecurity at $2 billion.

As AI continues to emerge, ethics has emerged as a critical concern in its development and deployment across various industries. The Belmont Report, first published in 1979, outlined the fundamental principles of clinical research ethics in the United States, including respect for persons, beneficence, and justice. In a recent study published in Science in October 2019 (191), researchers uncovered significant racial bias in an algorithm widely used in the US healthcare system to guide health decisions. The algorithm relies on cost as a proxy for needs, rather than illness. However, the US healthcare system spends less money on Black patients than on white patients with the same level of need. As a result, the algorithm erroneously assumed that white patients were sicker than equally sick Black patients, leading to a significant racial bias. The researchers estimated that this bias reduced the number of Black patients receiving additional care by over 50%. This example underscores the importance of addressing biases in AI and mitigating them from the outset to avoid discrimination based on factors such as race, gender, age, or disability. Biases can arise not only in the algorithm but also in the data used to train it.

To our knowledge, only three studies over the past three years have specifically focused on the ethical challenges related to the implementation of AI in healthcare. One such study by Char et al. (2018) examined the ethical concerns arising from potential errors generated by algorithms and their impact on decision-making. The same study also raised concerns about algorithms becoming a repository of collective medical knowledge. Another study by Guan (2019) investigated the role of governments in safeguarding ethical values associated with the emergence of AI in healthcare. The author emphasized the importance of ethical auditing by governments and defining the responsibilities of stakeholders in ethical governance systems. Gerke et al. (2020) identified legal challenges posed by AI in healthcare in the United States and Europe.

In this article, we will examine the reasons for bias in deep learning, specifically in the training process of deep learning and the fine-tuning process of reinforcement learning. We will also provide some suggestions for addressing these issues. The remainder of the paper is organized as follows: Section 2 discusses the ethical challenges and AI bias in various clinical fields, Section 3 examines the bias challenges in deep learning training and reinforcement learning fine-tuning, Section 4 presents our recommendations for addressing these issues, and finally, Section 5 provides a summary and conclusion.

Bias and Challenges of Artificial Intelligence in Clinical application

Data collection challenges

The accuracy of AI models is significantly influenced by the collection of training data, and this is especially challenging in the medical domain due to the need to protect patient privacy and ensure data fairness. For example, in 2019, Facebook employed AI to collect data from users' postings and then forecast their mental health and proclivity for suicide. As a result, Facebook collected and kept users 'mental health information without their knowledge or consent. Another case health records threat was reported by the journalist Alder (2020). Accordingly, the stage data, personal and health information of more than 2.5 million US patients was published online by an AI firm named Cense AI. The data were openly accessible through the internet and required no credentials to retrieve. These data had been temporarily stored into a storage repository before becoming deposited into the AI system, according to the author. Most of the time, AI-based systems solutions can violate the privacy rights of the patients. AI-based applications raise some concerns about user agreements. A contract that a person agrees to without a face-to-face dialog is contrary to the generally informed process of consent (Klugman et al., 2018). Most of the time, people do not take time and regularly violate user agreements (Friedman et al., 2000). Accordingly, some concerns may arise as to what kind of data should be gathered by AI developers and practitioners? For what purposes patients' data can be processed, used and shared? Can patients have the right to withdraw their data?

Breast imaging challenges

Certain systems may contain inherent latent biases, especially if these have been developed on datasets that underrepresent certain populations (with a lack of diversity in age, ethnicity and socioeconomic background) and therefore lack the ability to generalize. Outcomes based on pre-existing inequalities could be exacerbated by the skewed outcome being fed back into the algorithm, creating negative reinforcement, thus limiting the fairness of an algorithm. This can lead to algorithmic decisions that amplify discrimination and health inequalities. The data used in testing should therefore encompass a representative relevant population and the components of the dataset used explicitly reported alongside the results. A recent paper provides an example of such documentation, where an AI-CAD mammography algorithm trained on data from South Korea, USA and UK primarily using data from GE machines, achieved the best performance compared with other algorithms (sensitivity (81.9%) at the reader specificity (96.6%)), when tested on data from Sweden on only Hologic machines, demonstrating generalizability. Algorithms also have the ability to "learn on the fly", that over time become more biased due to "performance drift", thus potentially limiting their generalizability. "Learning on the fly" could potentially be beneficial to adjust algorithms to the local systems in which they are being used but this will also require close observation through regular audits to monitor for detrimental "performance drift".

AI algorithms challenges of breast cancer care

There is a common belief that AI is neutral and can be neither good nor bad in itself. In our view t this viewpoint is problematic since every algorithm encodes values. Certainly, AI has the potential to produce both positive and negative outcomes. But every algorithm will encode values, either explicitly, or more commonly in the era of 'new AI', implicitly. To give an example from breast screening: like any breast screening program, a deep learning algorithm may prioritize minimizing false negatives over minimizing false positives or perform differently depending on the characteristics of the breast tissue being imaged, or for women from different sociodemographic groups. Pre-AI, the performance of screening programs was a complex function of several factors, including the technology itself (e.g. what digital mammograms are capable of detecting) and collective human judgement (e.g. where programs set cut-offs for recall). Post new AI, these same factors will be in play, but additional factors will be introduced, especially arising from the data on which AIs are trained, the way the algorithm works with that data, and the conscious or unconscious biases introduced by human coders. The 'black box' problem in deep learning introduces a critical issue: explainability or interpretability. If an algorithm is explainable, it is possible for a human to know how the algorithm is doing what it is doing, including what values it is encoding. At present, Less explainable algorithms seem to be more accurate, and it is not clear whether accuracy and explainability must inevitably be traded off, or whether it is possible to have both. Fundamentally, machine learning systems are 'made of' data. By exposure to massive datasets, they develop the ability to identify patterns in those datasets, and to reproduce desired outcomes; these abilities are shaped not just by their coding, but also by the data they are fed. There is now extensive evidence from fields including higher education, finance, communications, policing and criminal sentencing that feeding biased data into machine learning systems produces systematically biased outputs from those systems; in addition, human choices can skew AI systems to work in discriminatory or exploitative ways. It is already well-recognized that both healthcare and evidence-based medicine are biased against disadvantaged groups, not least because these groups are under-represented in the evidence base. AI will inevitably reinforce this bias unless explicit human choices are made to counter it.

AI bias in Surgery

In surgical AI, bias mainly stems from training data. A taxonomy of biases in machine learning can be divided into two categories: technical or computational sources, and inappropriate use or deployment of algorithms and autonomous systems. Statistical bias in training data falls under the first category. Adhering to the iconic aphorism - 'garbage in, garbage out' - bias in the input will result in a biased model. For example, existing medical datasets have had much higher ratios of adult males of Caucasian origin (i.e., an over-representation bias) than exists in the actual population. A lack of diversity in sampling manifests in biased data and, without special controls in place, therefore results in biased models that may not behave as expected for under-represented groups. Other sources of technical or computational bias include algorithmic focus and processing bias. The focus bias emanates from the differential usage of information in training AI systems. For example, developers may deliberately include or exclude certain features (i.e., types of inputs)

when training a model, thereby causing it to deviate from the statistical standard if those attributes have a strong main effect on the outcome, or interaction effects with other variables. Algorithmic processing bias occurs when the algorithm itself is biased, as in the use of statistically biased estimators, which may result in a significant reduction of model variance on small sample sizes (i.e., the bias-variance trade off). Thus, developers may embrace algorithmic processing as a bias source in order to mitigate or compensate for other types of biases. The potential effects of biases emanating from technical and computational sources, of AI in surgery, could have direct effects on patient safety and system integrity. For example, training data bias could dramatically impact a preoperative risk stratification prior to surgery.

Training Bias in Deep Learning and Fine-tuning Bias in Reinforcement Learning

Training bias in deep learning refers to the situation where the algorithm is systematically biased towards certain types of data or features during training, leading to reduced performance on unseen data. This bias can arise due to a variety of reasons, such as the choice of training data, the selection of features, or the optimization algorithm used during training. The main manifestation of training bias is sampling bias. Sampling bias occurs when there is an underrepresentation or overrepresentation of observations from a segment of the population. Such bias, which is sometimes called selection bias, or population bias, may result in a classifier that performs bad in general, or bad for certain demographic groups. One example of underrepresentation is a reported case where a New Zealand passport robot rejected an Asian man's eyes because 'subject eyes are closed. A possible reason could have been that the robot was trained with too few pictures of Asian men, and therefor made bad predictions on this demographic group. There are many reasons for sampling bias in a dataset. One kind is denoted self-selection bias and can be exemplified with an online survey about computer use. Such a survey is likely to attract people more interested in technology than is typical for the entire population and therefor creates a bias in data. Another example is a system that predicts crime rates in different parts of a city. Since areas with more crimes typically have more police present, the number of reported arrests would become unfairly high in these areas. If such a system would be used to determine the distribution of police presence, a viscous circle may even be created. Survivorship bias occurs when the sampled data does not represent the population of interest, since some data items 'died'. One example is when a bank's stock fund management is assessed by sampling the performance of the bank's current funds. This leads to a biased assessment since poorly performing funds are often removed or merged into other funds. Another example is that GPT 4 did know who won 2022 word cup (See Figure 1). The reason is GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its pre-training data cut off in September 2021 and does not learn from its experience. Training bias can be a major problem in deep learning, as it can lead to poor performance on real-world data. To avoid training bias, it is important to use a diverse and representative training dataset, and to carefully consider the choice of features and optimization algorithm used during training. Additionally, it is important to monitor the performance of the model on test data, and to adjust the training process as needed to avoid overfitting and improve generalization performance.

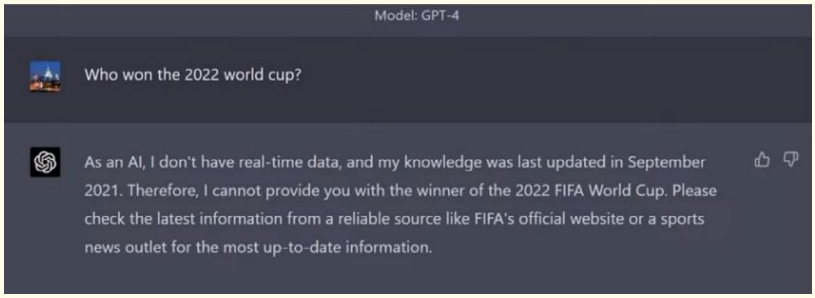

Figure 1: Answer from GPT 4.

Fine-tuning bias in reinforcement learning refers to the situation where the agent becomes biased towards the specific environment and tasks it has been fine-tuned on, leading to reduced performance on new, unseen environments or tasks. This bias can arise when the agent is fine-tuned on a particular task or environment using a small amount of data, and then deployed in a new, different environment where it may not perform as well. Figure 2 shows the three-step training process of the instructGPT. The third step, which involves fine-tuning of RL, may have potential biases. In the first two steps, biases may also arise due to the involvement of labelers, such as when they rank outputs from best to worst.

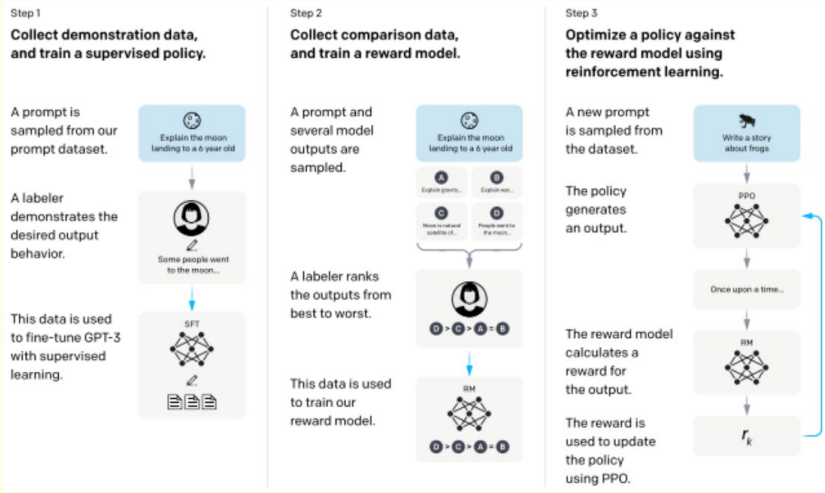

Figure 2: A diagram illustrating the three steps of InstructGPT.

These rankings may differ depending on the labelers' knowledge backgrounds, political stances, and life experiences. To avoid fine-tuning bias in reinforcement learning, it is important to use a diverse and representative training dataset, and to carefully consider the choice of pre-training data, fine-tuning procedure, and evaluation metrics used during the training process. Additionally, it is important to monitor the performance of the agent on new environments and tasks, and to adjust the training process as needed to improve generalization performance.

Proposed model and principles

To address the bias of deep learning, we should firstly identify potential sources of bias. This could include biased data sets, biased algorithms, or biased features. Next, collect diverse and representative data. Ensure that training data is diverse and representative of the population that are trying to model. The third is to Pre-process and clean data. Pre-processing and cleaning your data can help remove any biases that may be present in your data set. For example, we could remove features that are highly correlated with sensitive attributes such as race or gender. The forth step is regularly test and monitor AI model, regularly test and monitor model to ensure that it is not exhibiting any bias. This can include testing for fairness, accuracy, and other metrics. Form an expert committee composed of experts in various fields, like computer experts, medical experts and sociologists, etc. They stipulate test standards and select standard test samples. Another test committee composed of more diversity persons is used. They test AI model by the sample specified by the expert committee.

Suppose the output of the standard is y and the output of the test is $\hat{y}$. Results of testing bias b.

$$b = \sum_{i=1}^{N}(y_i^2 - \hat{y}_i^2) \qquad (1)$$

In generally, the smaller *b* is the better. If it falls within an acceptable range, the bias of the model can also be accepted to meet the regulations. Otherwise, it is necessary to adjust the training samples to correct the model parameters until *b* meets the requirements. Overall, addressing AI bias requires a combination of technical expertise and ethical considerations. By following these steps and continually monitoring and improving a AI model, only that can help ensure that it is fair, accurate, and unbiased.

In order to ensure the correct use of AI in the clinical application, the training of deep learning models should comply with the following principles:

Safety is one of the principles for AI in the clinical application. When training AI model, priority must be given to maintaining and demonstrating evidence of patient safety and quality of care. In the service of safety and patient confidence some amount of transparency must be ensured. While in an ideal world all data and the algorithms would be open for the public to examine, there may be some legitimate issues relating to protecting investment/intellectual property and also not increasing cybersecurity risk. Third party or governmental auditing may represent a possible solution.

Fairness is another principle of training for AI in the clinical application. Unlike bias, the fairness of a machine learning model is judged against a set of legal or ethical principles, which tends to vary depending on the local government and culture. In addition to diagnostic prediction, machine learning algorithms are now being applied to operational aspects of health care delivery, such as decisions regarding admissions and triage, as well as determining the cost of insurance premiums that a patient should pay. All these applications have the ability to produce unfair outcomes with respect to demographic groups; therefore, it is necessary to have a framework for quantitatively assessing the fairness of such decisions. When collecting data sets for training and validation purposes, both the risks and benefits associated with health data collection need to be shared equitably across different populations, giving attention to not disadvantaging marginalized groups.

Transparency and accountability another principle of training for AI in the clinical application. AI systems should be auditable, comprehensible and intelligible by "natural" intelligence at every level of expertise, and the intention of developers and implementers of AI systems should be explicitly shared. If an AI system fails or causes harm, we should be able to determine the underlying reasons, and if the system is involved in decision-making, there should be satisfactory explanations for the whole decision making process. This process should be auditable by the healthcare providers or authorities, thus enabling legal liability to be assigned to an accountable body.

Conclusion

The application of AI in clinical settings should prioritize the promotion of well-being, minimize harm, and ensure equitable distribution of benefits and risks. To achieve this, AI systems must be transparent and dependable, with a focus on mitigating bias in decision-making. Human designers or operators should remain responsible and accountable for the outcomes. In this article, we explore the various forms of bias that may be present in AI models used in clinical settings and present solutions to address these issues. We propose safety, fairness, transparency, and accountability as key principles to guide the training of AI models and reduce bias in clinical decision-making.

References

1. Schwab K. "The Fourth Industrial Revolution: what it means and how to respond". World Economic Forum (2016).
2. AI in the UK: ready, willing and able? United Kingdom: authority of the house of lords (2018). (Intelligence SCoA, editor).
3. Ravi D., et al. "Deep learning for health informatics". IEEE J Biomed Health Inform 21.1 (2017): 4-21.
4. LeCun Y, Bengio Y and Hinton G. "Deep learning". Nature 521.7553 (2015): 436-44.
5. Price Waterhouse Cooper. Sizing the prize: What's the real value of AI for your business and how can you capitalize? (2017).
6. Murphy K., et al. "Artificial intelligence for good health: a scoping review of the ethics literature". BMC medical ethics 22.1 (2021): 14.
7. Brady A P and Neri E. "Artificial intelligence in radiology—ethical considerations". Diagnostics 10.4 (2020): 231.

8. WHO Consultation, Development of guidance on ethics and governance of artificial intelligence for health, Geneva, Switzerland (2019).
9. Gibney E. "The battle for ethical AI at the world's biggest machine-learning conference". Nature 577.7792 (2020): 609-609.
10. Jackson BR., et al. "The ethics of artificial intelligence in pathology and laboratory medicine: principles and practice". Academic Pathology 8 (2021): 2374289521990784.
11. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; Licence: CC BY-NC-SA 3.0 IGO (2021).
12. Kooli C and Al Muftah H. "Artificial intelligence in healthcare: a comprehensive review of its ethical concerns". Technological Sustainability (2022).
13. DeCamp M and Lindvall C. "Latent bias and the implementation of artificial intelligence in medicine". J. Am. Med. Informatics Assoc 27 (2020): 2020-2023.
14. Salim M., et al. "External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms". JAMA Oncol 6 (2020): 1581-1588.
15. Pianykh OS., et al. "Continuous learning AI in radiology: implementation principles and early applications". Radiology 297 (2020): 6-14.
16. ChenIY., et al. "Ethical machine learning in health. ligence in medicine". J. Am. Med. Informatics Assoc (2020).
17. NHSX. "A buyer's guide to AI in health and care". Radiology 286 (2018): 800-809.
18. Hosny A., et al. "Artificial intelligence in radiology". Nat. Rev. Cancer 18 (2018): 500-510.
19. Hickman SE, Baxter GC and Gilbert FJ. "Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations". British journal of cancer 125.1 (2021): 15-22.
20. Carter SM. "Valuing healthcare improvement: implicit norms, explicit normativity, and human agency". Health Care Anal 26.2 (2018): 189-205.
21. Cabitza F, Rasonini R and Gensini GF. "Unintended consequences of machine learning in medicine?". J Am Med Assoc 318.6 (2017): 517-518.
22. Holzinger A., et al. What do we need to build explainable AI systems for the medical domain?. arXiv:1712.09923 (2017).
23. Cowls J and Floridi L. "Prolegomena to a white paper on an ethical framework for a good AI society". SSRN (2018).
24. O'Neil C. Weapons of math destruction. London: Penguin Random House (2016).
25. Rogers WA. "Evidence based medicine and justice: a framework for looking at the impact of EBM upon vulnerable or disadvantaged groups". J Med Ethics 30.2 (2004): 141-5.
26. Carter SM., et al. "The ethical, legal and social implications of using artificial intelligence systems in breast cancer care". The Breast 49 (2020): 25-32.
27. Latrice G Landry., et al. "Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice". Health Affairs 37.5 (2018): 780-785.
28. Adewole S Adamson and Avery Smith. "Machine Learning and Health Care Disparities in Dermatology". JAMA Dermatology 154.11 (2018): 1247-1248.
29. Stuart Geman, Elie Bienenstock and René Doursat. Neural networks and the bias/variance dilemma. Neural computation 4.1 (1992): 1-58.
30. David Danks and Alex John London. "Algorithmic bias in autonomous systems". In IJCAI (2017): 4691-4697.
31. Rudzicz F and Saqur R. Ethics of Artificial Intelligence in Surgery (2020).
32. Rice University David M. Lane. "Chapter 6 research design - sampling bias". in Online Statistics Education: A Multimedia Course of Study, Rice University.
33. Alexandra Olteanu., et al. "Social data: Biases, methodological pitfalls, and ethical boundaries". Frontiers in Big Data (2019).
34. CNN World (2016).
35. Anupam Datta., et al. "Proxy non- discrimination in data-driven systems". CoRR, abs/1707.08120 (2017).
36. Burton G Malkiel. "Returns from investing in equity mutual funds 1971 to 1991". The Journal of Finance 50.2 (1995): 549-572.

37. Hellström T, Dignum V and Bensch S. "Bias in Machine Learning--What is it Good for?". arXiv preprint arXiv:2004.00686 (2020).
38. Ouyang L., et al. "Training language models to follow instructions with human feedback". arXiv preprint arXiv:2203.02155 (2022).
39. Gerke S, Minssen T and Cohen G. "Chapter 12—Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare". Artificial Intelligence in Healthcare; Bohr, A., Memarzadeh, K., Eds.
40. Currie G, Hawk KE and Rohren EM. "Ethical principles for the application of artificial intelligence (AI) in nuclear medicine". European Journal of Nuclear Medicine and Molecular Imaging 47 (2020): 748-752.
41. Fletcher RR, Nakeshimana A and Olubeko O. "Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health". Frontiers in Artificial Intelligence 3 (2021): 561802.
42. High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI (2019).
43. Floridi L., et al. "AI4People - An ethical framework for a good AI society: opportunities, risks, principles, and recommendations". Minds Mach 28.4 (2018): 689-707.
44. D'Antonoli TA. "Ethical considerations for artificial intelligence: An overview of the current radiology landscape". Diagnostic and Interventional Radiology 26.5 (2020): 504-511.