Ethical Considerations of Artificial Intelligence via Neural Networks Applied to Medical Applications

Ternent, James
jwternent@wpi.edu

Thompson, Maximilian
mthompson2@wpi.edu

May 12, 2020

Abstract

This project examines the possible ethical implications of emerging machine learning technologies in medicine, as technological complexities underlying such ethical implications need to be translated into a language accessible to the broader public. The project also provides a description of machine learning that is intended to inform those who are not familiar with the technology. After analyzing some of the benefits of present and potential applications of machine learning in medicine, we call attention to issues such as explainability and interpretability.

1 Introduction

Artificial intelligence (AI) and, more particularly, machine learning (ML) is becoming increasingly prevalent in medical fields. The regulations proposed by the FDA[1] place a large emphasis on procedural artificial intelligence which leaves the capabilities of machine learning models relatively less examined. We wish to add to this discussion in the hopes of extending the ethical reasoning and guidance with regards to neural network implementations, so as to achieve more safety in applications of these devices and to greater understand the realm of applicability of these devices.

Machine learning comprises a subset of artificial intelligence that closely matches what most commonly comes to mind when artificial intelligence is mentioned in public conversation. The difference between procedural artificial intelligence and machine learning is detailed in section 1.2, but, in essence, procedural artificial intelligence takes input data and outputs a response by following a strict set of rules or procedures, while machine learning takes input data and outputs a response determined by a set of facets of interest and weights that may change and shift in response to each input. Frequently machine learning models undergo a period of "training", during which weights may

shift wildly, with the rate of change slowing down until it reaches a period of relatively little change.

We also provide an overview of ethical considerations within the context of medical devices in section 2 before continuing to weigh the pros and cons of machine learning. We initially intended take survey of a number of nursing students to gauge their general opinion, understanding, and expectations of machine learning. Due to how overworked nurses can be[2] and how much of an impact machine learning has had and will have on the nursing population[3], we believe that their viewpoints are and important part of the conversation which we wish to highlight. However, we had to modify this plan due to the ongoing pandemic at the time of writing.

1.1 Review of Terminology

First we must define some terms used commonly in the field of machine learning that may be misleading to those without a background in it. The term "feature" refers to what amounts to an interesting property or value in the input data represented by a function explicitly defined by the developers of the artificial intelligence. Within the context of machine learning, these functions are what is modified by changes in the weights of the model. Additionally, the term area under the curve (AUC) is used to express the quality of predictions derived from the relation of two values referred to as sensitivity and selectivity. Sensitivity is the rate of positive predictions (predictions that claim something to be true). Selectivity is the rate of negative predictions (predictions that claim something to be false). An important caveat is that the rate of positive predictions includes both correct predictions *and* false positives. The same is true for selectivity and false negatives.[4, 5]

Explainability of machine learning models is an important ongoing research field. An explanation of a model seeks to convey why a model made the decisions it did and why it outputted what it did, and explanations have two key attributes: completeness (how accurately the explanation conveys the workings of the machine learning model), and interpretability (how easy it is for humans to understand the explanation). The two are often at odds, and finding the solution to this problem is an important field of study[6]—especially regarding the validation of machine learning model outputs.[7] Unfortunately for the sake of clarity for non-experts, the term interpretability is not only applied to explanations, but also to the machine learning models themselves. When used to describe a model, the term refers to how easy it is to identify and understand the mechanisms that drive the machine. Essentially explainability seeks to answer how the machine works, while interpretability regards why the machine did what it did.[6] More in depth definitions of these concepts are given in section 1.2.

1.2 Review of AI and ML

Artificial Intelligence in its basic form is a computer program that takes in an input or inputs and produces an output. An example would be if a program took in how many hours it had been since someone washed their hands and determined whether they were due for washing their hands or not. The Artificial Intelligence would have a threshold that would trigger a recommendation to wash their hands. For the purpose of this example, we will set it at two hours. If it had been half an hour since someone washed their hands, it would not tell them to wash their hands, but if it had been five hours since they washed their hands, the Artificial Intelligence would recommend washing their hands. However, we know time is not the only factor in whether someone should wash their hands. To better improve this theoretical Artificial Intelligence, a second input could be added such as whether the person is about to touch or consume food. In general, people should wash their hands if they are about to touch or consume food. The updated logic in the Artificial Intelligence is that it will recommend washing hands if the time since the last time someone washed their hands is greater than two hours or if they are about to touch or consume food. Only one condition has to be true to recommend a washing hands, but both could be true. The more inputs or parameters that the Artificial Intelligence takes in, the better it will get at giving a correct output as long as the researcher determines the correct thresholds for the Artificial Intelligence.

Eventually as the Artificial Intelligence gets more complex, two problems present themselves. First, not all problems that an Artificial Intelligence tries to solve have a simple linear relation between their inputs and the output. Second, the researcher may not know the exact relation between the inputs and the output. This is where a subset of Artificial Intelligence called Machine Learning (ML) comes into play. A common type of ML is called a Neural Network (NN) which more or less simulates a brain though in a vastly oversimplified manner. The brain has many neurons that are connected together by axons, dendrites, and synapses. Axons are used to transport the electrical signals for the neurons. Synapses receive the signals at the end of the axons. Each synapse has a strength that determines how much the signal matters whether from another neuron or an input from the body. A neuron can have multiple inputs. The neuron sums all the signals although some matter more than the others based on their synaptic strength. Once the neuron reaches a certain sum, it fires sending signals to other neurons or the body. The synaptic strengths can be changed over time based on whether the response generated by the neurons was correct for a particular situation. Over time, the synaptic strengths for a neuron are optimized to provide a correct response most of the time. Neural Networks work in a similar way. Each NN has nodes and weights. Nodes are like the neurons in the brain and weights are like the synaptic strength. The most basic NN has two layers. A layer of input nodes that takes the inputs in and passes them onto the next layer, the output layer. The hidden layer takes in the input, multiplies them by their weights, and sums them. An activation function maps the sum of the adjusted inputs to the desired output. The way

the NN is trained through using data with known inputs and outputs. An algorithm known as error backpropagation is used to adjust the weights. In simple terms, it compares the output using the current weights, compares it to the desired output, and adjusts the weights until there is a little to no error.[8] The weights can either increase or decrease based on the error backpropagation algorithm. Higher weights means an input is more important to determining the outcome, but lower weights means that whatever input is not as important. The weights allow the Neural Network to decide the features, what is important to the output. This is one of the things that makes NN's so powerful. The example of washing hands can also be modeled with a NN. It would have two input nodes. One would take in the number of hours since someone had washed their hands and the other would take in whether someone was about to touch or consume food. The input nodes send the inputs to the output node where the inputs would be multiplied by their weights and summed. At first the weights would be randomly assigned or assigned to an estimate based off of previous knowledge of the researcher. An activation function would then be used to turn the sum into a yes or no of whether someone should wash their hands. Training data with known inputs and outputs would be used with error back propagation to adjust the weights until the error was eliminated. More inputs can also be added to improve the NN, but the real power of NN comes by adding a layer or layers between the input and output layers. Nodes can also be added to these layers. The more nodes the better. Once there is more than one layer, the connections between the nodes of separate layers can be chosen. The nodes of the previous layer do not have to connect to all the nodes of the next layer. More layers, nodes, and connections will make the NN better, but as the complexity increases, the time that it takes for the NN to produce an output increases. At a certain point of adding layers, nodes, and connections, the improvement of the success rate will be marginal in comparison with increase in time to run the NN. Neural Networks are used to model nonlinear problems linearly which can be very useful in the medical field.[9]

Artificial Intelligence exists on a scale between human driven data analysis and machine driven analysis. Closer to humans doing the analysis of data is the example of the procedural Artificial Intelligence for washing hands. Researchers define the rules for washing hands that make the decision of whether someone should wash their hands. These rules are hard coded into the Artificial Intelligence. Somewhere in the middle of this scale is the Neural Network version of the hand washing Artificial Intelligence. An algorithm determines the relationship between the parameters and the output, but the researchers chose two parameters which they felt were important. At the far end of the machine driven side, algorithms do most of the work. A perfect example of this is image recognition. The only human input is a researcher that annotates the images with the correct response. The algorithm figures out what makes a dog image different from a cat image for example with no human intervention. The algorithms needed are often called deep learning algorithms due to the multiple layers of the Neural Networks they create. Deep learning algorithms have the power to find the connection between vast amounts of data and the proper output, help-

ing humans classify something that is not easily understandable. However the more control that is given to algorithms, the less you can guarantee accuracy or fairness. They more or less become black boxes.[10]

Although there are these risks, algorithms or machine learning can vastly help when it comes to medicine. Artificial Intelligence in medicine mainly focuses around a few disease types: cancer, nervous system diseases, and cardiovascular diseases. These diseases have hit humanity the hardest which is why they are the focus of researchers today. In each of these diseases, diagnosis imaging has been the focus of research. This is where machine learning thrive, finding the connection between symptoms, data from other non-invasive tests and what life altering disease a patient may have. A great example of the use of machine learning in medicine is in early stroke detection. A movement detection device was used to monitor a patient's movements. The device would record both the normal movement of a person and the movement before a stroke started. By using machine learning, the device was able to tell when a person's movement patterns differed significantly from their normal ones.[11]

2 General Ethics of Medical Devices and ML

Artificial Intelligence has the power to save many lives in the medical field, but researchers must be careful to not cause any unnecessary risks to patients and the medical professionals using Artificial Intelligence. In their book Principles of Biomedical Ethics, Beauchamp and Childress outline four principles: autonomy, beneficence, nonmaleficence, and justice. Each principle is used to help medical professional make decisions on treating patients.

The first principle autonomy refers to the ability to decide for oneself free from control of others and with sufficient level of understanding as to provide for meaningful choice. Also, the person deciding should have capacity to make the choice. Next, beneficence is the principle of weighing the risks against benefits in order to get the best result. This applies to not only the patient, but also to society in general. The benefits to society should be considered when weighing the risks. Nonmaleficence can be summed up with primum non nocere, first do no harm. A medical professional should do no harm. Finally, justice addresses the question of who receives healthcare resources as they are in scarce supply.[12]

Artificial Intelligence is primarily being developed as a diagnostic tool to help medical professionals. In order to develop these tools, researchers need access to large amounts of data such as Electronic Health Records. Machine learning algorithms use this data to define features. However, the Electronic Health Record (EHR) of a person is private information for use by doctors to treat said person. This brings into question is the ethics of the use of a patient's Electronic Health Records.

When using EHR, the first and most important principle, autonomy, should be thought of. Will patients allow the use of their EHR's to develop Artificial Intelligence that could save many people? Patients will need to be informed of what their EHR's are being used for and what the risks are. The risk of using

EHR's should be weighed against the benefits. The main risk of using EHR's is having patient information leaked and used against them. This risk can be minimized by removing identifiable information from the EHR's before being used for research.

Patients should know that an algorithm defined what was important in determining their future diagnosis and how much input a doctor has in the final diagnosis. Any risk that has appeared in testing needs to be communicated to the patients. They need to be informed to make a decision. The beneficence of the Artificial Intelligence needs to be examined. Do the benefits justify the risks? At what point are the risks too great? These are questions researchers and health professionals need to ask themselves. One thing doctors principle by is to do no harm. If they use an Artificial Intelligence, can they guarantee that the patients will receive more benefits than the possible risks they are taking on?

The final question of medical Artificial Intelligence is the justice of them. Medical Artificial Intelligence could become life saving tools, but the likelihood is that they will only be able to be used in developed countries. Less developed countries could miss out on them and fall further behind the world in health.

Ethics are something that need to be considered especially with new unproven technology. Medical Artificial Intelligence has the power to change the world, but they should not be released without considering its ethical implications.

3 Impact of Expectations of ML in Medicine

There are unrealistically high hopes for machine learning in medicine[13, 14], and there are just as many unrealistic fears to complement them[15]. We chose to conduct our own survey of nursing students to gather opinions on their hopes and concerns regarding machine learning in medicine, as well as hopes and concerns. As a frequently understaffed and overworked profession[2], we believe that among healthcare professionals they stand to gain the most from advances in medical machine learning, and thus would provide valuable insight. Unfortunately, due to the ongoing COVID-19 pandemic at the time of writing, we felt that it would be inappropriate to administer such a survey. We believed that it may cause undue distress for participants given the extenuating circumstances most find themselves in.

A number of surveys summarized by Vayena *et al.* show a clear divide between public and professional opinion on medical machine learning: "63% of the adult population [of the United Kingdom] is uncomfortable with allowing personal data to be used to improve healthcare and is unfavorable to artificial intelligence systems replacing doctors and nurses in tasks they usually perform." However, drawing from a German survey, they express that a large portion of medical students believe wholeheartedly that machine learning will improve medicine. Though they are not eager enough to embrace it with little caution, expressing that they are "skeptical that it will establish conclusive

diagnoses."[16] This opinion is roughly mirrored by another United States survey cited by the authors of the article.[17]

These statistics illustrate a clear disconnect between public and professional opinions, that warrants further discussion. Each viewpoint has the potential for profound impact on the perception of developing technologies and together they highlight the how ethical issues can arise from the disconnect itself. For instance (as further discussed in section 5.2), the disconnect between an acceptable explanation for a machine learning expert and an acceptable explanation for a patient can cause distress and possibly even harm.

3.1 Data and Privacy

Chen and Asch argue that the public hopes for the predictive power of machine learning in medicine vastly outstrip the realistic potential for the emerging technology. The limitations of data sources (such as electronic health records) can introduce bias[18] due to the fact that they were only ever intended to be used (in their current state) for clinical care and billing. Chen and Asch propose adding additional data sources such as browser history to diversify and improve predictive power. However, they are quick to add that even that makes some glaring assumptions regarding how closely the future will resemble the present. It is incredibly unlikely that machine learning models will ever be able to predict something as broad and distant as the date of a catastrophe—just the risk of one occurring within a given period of time. Regardless, they are still hopeful for the future of machine learning, showing greater concern for the potential backlash when the predictive power of machine learning doesn't meet these expectations.[13]

Chen and Asch make an assumption regarding the availability of numerous unfettered sources of user information in an era where large scale data breaches have become almost commonplace with, as reported by CNET, 5,183 data breaches within 2019 alone.[19] With bills such as the General Data Protection Regulation (GDPR) bringing well deserved attention to the storage and processing of personal data people are significantly more skeptical of what personal data is being collected and how it is being used.[16] While this does promote the ethical handling of personal data, it may create greater pressure on researchers to identify a concrete reason for including a data stream; this has the potential to introduce bias through the choice of which data streams are available to a given machine learning model and which are deemed unnecessary.[18]

These developments raise important questions: how much loss of privacy are the potential improvements to medical decision making given by machine learning worth? How much privacy is the public willing to part with regardless of the value? Though promises may be made regarding maintaining privacy through safe storage of data, it is evident that data breaches have become more a question of when rather than a question of if.

Shameer *et al.* express hope regarding the future of machine learning and cardiology, while acknowledging that there are a number of issues that must be overcome. Additionally, they note that machine learning is more likely to be a

complement to standard statistical analysis rather than a replacement.[20] This falls roughly in line with what is expressed in the article by Vayena *et al.*[16], and the same trend can be seen in a number of articles.[21, 22]

This demonstrates a level of caution that will likely be conducive to greater safety in developing machine learning technologies, but can the same be said regarding ethical growth in future devices?

4 Benefits of Medical AI

Artificial Intelligence can improve society's overall health. Researchers are mainly focused on using artificial intelligence for cancer, neurological diseases, and cardiological diseases. It is used in healthcare in such places as clinical practice and translational research.

Currently the focus of artificial intelligence development in clinical practice is for diagnostic purposes. One of the most successful areas in diagnostics is automated image-based diagnosis. Medical professionals such as ones in radiology, ophthalmology, dermatology, and pathology rely on image based diagnoses. In radiology, applications of artificial intelligence such as the detection of lung nodules, the diagnosis of pulmonary tuberculosis and breast mass identification have reached expert level diagnostic accuracy. This also applies to dermatologists, whose job is largely based around inspection of the skin and its irregularities. Typical skin melanoma has visual features that separate it from benign lesions. Dermatologists have even developed a guideline known as ABCDE. A stands for the asymmetry of a lesion, B for irregular borders, C for color variegation, D for a diameter of 6 mm or greater, and E for enlargement of the surface of a lesion. With the exception of E, every other guideline can be accessed from a photo. A neural network was trained on 129,450 images to identify skin melanoma. The accuracy of the neural network was greater than the average dermatologists based a comparison between the assessments of the neural network and twenty-one dermatologists. The training for the neural network was computationally intensive, but could be deployed as a mobile app granting in home access to expert level screening. In ophthalmology, retinal cameras are used to capture the retina, optic disc,and macula. This is used to detect and monitor diseases such as DR, glaucoma, neoplasms of the retina and age-related macular degeneration. Artificial intelligence helps when it comes to diabetic patients. Diabetic patients are at risk for DR and are recommended to be screened every year. Usually, an ophthalmologist will examine and interpret the photos. This creates a large workload for ophthalmologists for something that is a part of a routine check up for many people. However, a neural network was created and trained on 128,175 images and had similar performance to ophthalmologists.[23] The second area artificial intelligence can help out in clinical practice is in administration. The business of healthcare has become more complex with healthcare infrastructure being stretched thin due to administrative burdens and resourcing constraints. Artificial intelligence can perform repetitive and routine tasks such as patient data entry and automated review of laboratory data and imaging results. If

machine learning algorithms are attached to electronic health records, they can assist clinicians and administrators retrieve context related patient data allowing patients to be better treated. Artificial intelligence can also assist in the logistics of hospitals. They can predict hospital stay time from pre-admission data provided by patients, enabling more efficient use of hospital resources. Artificial intelligence can improve the health of patients through patient monitoring. With the adoption of smartphones and fitness monitoring devices, it is now possible to have access to details of patients' sleep patterns, blood pressure, heart rate, and other measures that was not possible before. These measures when analyzed by medical professionals can tell how healthy or unhealthy someone is, but there is nowhere close to enough people to analyze that much data. Artificial intelligence when trained can analyze this data and extract actionable information.[24]

Artificial Intelligence can also be very useful in translational research, the research of turning biomedical research into new therapies, medical procedures, or diagnostics. Machine learning has been used for biomarker discovery. Biomarkers are measurable indicators of the severity or presence of a disease. Biomarker discovery depends on the identification of unrecognized correlations between thousands of measurements and phenotypes. It is almost impossible for researchers to manually analyze the vast amount of data to find the correlations necessary to define a biomarker. Many of the biomarker panels that have been generated by machine learning have outperformed selected by experts or traditional statistical methods. Artificial intelligence can be used to predict clinical outcomes. Electronic health records with the use of machine learning predict mortality, readmission, and length of hospital stay. If data from health insurance claims are used, the mortality of elderly patients can be predicted.[24]

Artificial Intelligence can vastly improve the healthcare provided to patients. Through machine learning, artificial intelligence has the power to find connections that were not possible before or at least very resource intensive. It can also bring expert level medical opinions to the average person's fingertips. The benefits of artificial intelligence could take the health of the world to another level.

5 Low Explainability in Complex Systems

A key issue lies in two important facets of machine learning: explainability of machine learning models and interpretability of their results. Luo *et al.* define explainability as the ability to "summarize the reasons for the behavior of [machine learning] algorithms" and interpretability as the ability to "comprehend what a model did."[25] Though the difference between the two is subtle, the distinction is still an important one to make; methods that may increase interpretability may not yield results regarding the overall explainability of the model. However, improvements to explainability increase interpretability by definition.

More often than not machine learning models are a black box—a machine

learning model with very little explainability—that takes in datasets and outputs values with some prescribed significance. It is difficult for non-machine learning experts understand the risks and potential failure points of such a model, and methods of addressing this lack of explainability are still being debated.[26, 7] This in conjunction with a typical person's initial trust placing little to no emphasis on the actual functionality of the model[27] can make for sudden and unexpected tragedy within medical fields.

For example a CEHC study in the mid 90's found a rule-based trained to incorrectly identified pneumonia patients with asthma as having lower risk due to the lower mortality rate following urgent movement to intensive care.[28] If such an algorithm were used in a real hospital environment, it is unlikely that there would be enough resources or time to devote to interpreting the results to identify this. However, if the algorithm were explainable to the layman this issue would be much more readily identifiable.

Improving explainability of algorithms and interpretability of results isn't as simple as, for instance, writing a more in depth manual. The burden falls to researchers and developers to improve these important aspects of machine learning, and not those in charge of documenting. As Coley points out: "the architecture of deep learning models should inherently enable some degree of interpretability as has been done in natural language processing."[29]

To make matters more complicated, it is important to consider who the audience is when determining if an algorithm is explainable. If an algorithm can theoretically be explained but requires a degree in the subject to understand, can it really be considered explainable for medical professionals and patients? We believe that those whose lives are impacted by usage of machine learning techniques are ethically entitled to at the very least a cursory understanding of the given algorithm and the information it has procured.

5.1 Difficulty of Validation

There is immense difficulty of validating the predictions (before acting on them) of a machine learning model. This creates further risk by obfuscating potential error behind a highly technical veil for non-machine learning experts. Moreover, unknown error introduced in input datasets is difficult to detect without access to the data itself, as the machine learning model will continue to output seemingly valid predictions[7] Error could be introduced through any number of innocuous vectors (differences in storage systems between health centers, errors digitizing physical records, decay of electronic records over time) or through more malignant ones (such as a targeted attack). The ethical delicacy of handling electronic health records[30] can make impossible to directly examine new inputs for a source of error, and even if it were possible, the sheer magnitude of the task would be prohibitive. Redyuk *et al.* proposes a method of detecting dataset shifts (without direct examination) that may mitigate this issue in the future[7], but it still remains something important to discuss.

Shifts in datasets are not the only thing that may cause unexpected error in outputs. Poor training, though ideally easier to identify simply due to training

occurring well before live usage, is just as much of a risk. Unintentional biases may result from flawed training data[31] or training data that has intentionally been poisoned or otherwise tampered with[32]. In the case of healthcare, these issues may be incredibly difficult to detect once a machine learning model is being used in production—again due to the delicate nature of electronic health records. It may not be possible to verify that all training data is both intact and correct without crossing ethical or legal bounds.

The consequences of both unforeseen shifts in inputs and poor training may prove discriminatory or even life threatening. Even given some virtually infinite workforce to attempt to validate all results of machine learning models, doing so would require a level of precognizance that is currently impossible to attain.

5.2 Lack of Acceptable Explanation of Decision Making

Even if a decision could magically be made with complete and utter certainty at all times, the frequent lack of explainability in machine learning introduces a number of problems and risks at the patient care level. Communication is incredibly important in a doctor-patient relationship,[33] and obfuscating decision making behind a black box harms the ability for clinicians to effectively communicate with patients. When the medical professional does not know how a decision was made, how can they be expected to provide a satisfactory explanation for their patients?

Going forward, the need for an explanation should drive development of architectures geared toward providing an acceptable explanation with minimal loss of accuracy. One architecture of note is detailed by Shachor *et al.* Their proposed "mixture of views" architecture is an expansion of an already used "mixture of experts" architecture which connects number of neural networks together to provide the goal functionality.[34] The points of connection between the neural networks provides opportunity for insight into what would otherwise be one big black box. By understanding what factors are considered, providing an explanation of a neural network's decision would become possible.

However, exposing the facets of a problem considered by a neural network highlights an important ethical consideration: what potential aspects of a decision are suitable for an ethical explanation, and on what level should unethical aspects be removed or recontextualized? For instance, suppose skin tone were an emergent feature in a machine learning model; this would clearly be an unacceptable facet of a patient to consider. Would the only proper solution be to accept any potential loss in accuracy and eliminate this feature from the model, or would it be ethical to allow it to remain and simply highlight other features in an explanation given to an inquiring patient? Would being able to statistically prove identical quality of decision making for majority and minority groups through comparisons of things such as the area under the curve and demonstrating equal representation in training data sets through augmentation of minority data[35] affect whether or not such considerations are ethical? How will the ethical acceptability of given terms affect the architecture of machine learning applications moving forward?

5.3 Diffusion of Liability

Though largely considering more concrete applications of machine learning (such as self driving cars), Reed *et al.*'s 2016 article on the legal responsibility and liability (with regards to English law) points out that there are several parties that could be considered at fault should harm come to a patient due to the complex nature of development and usage of machine learning models. Moreover they posit that it would be incredibly difficult to establish the fault of one party and significantly more expensive to establish the fault of several parties in unison.[36]

The difficulty of proving liability has the potential to do significant harm. Should a defective algorithm result in patient harm, the blame might not even reach the manufacturers of the algorithm due to the difficulty of proving so—allowing said algorithm to see continued usage. Inversely, blame may falsely be pinned on the manufacturer giving a clinician an easy way to avoid being held accountable for medical malpractice. These issues are only aggravated by "black box" models making it more difficult to validate the output of the algorithm and assign appropriate blame.

Methods of responding to these issues are being developed. For instance the GDPR includes a "right to explanation" of algorithmic decisions.[37] However, as Watcher *et al.* point out the current state of the GDPR (as of 2017) does not actually ensure a "right to explanation" and instead grants something more akin to a "right to be informed."[38] While a true "right to explanation" would be vastly more preferable, it is still a step in the right direction enacted in a very real and actionable way.

6 Conclusion

Machine learning is a growing part of health care, and is already seeing significant use within some fields.[20] There are many ways we can benefit from advancements in machine learning, however machine learning as a whole is often difficult for non-machine learning experts to understand.[26] Improvements in explainability will likely raise a number of ethical conundrums that will significantly affect the shape the architecture of machine learning applications takes. Additionally, while steps are being made to address this on a legal level[37] it is clear that we are not quite there yet.[38] We are left with an important question: how will machine learning develop within the following years and how will it change as we address the ethical concerns brought forth in the present?

References

[1] Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (samd). 2019.

[2] Peter Van Bogaert, Herman Meulemans, Sean Clarke, Karel Vermeyen, and Paul Van de Heyning. Hospital nurse practice environment, burnout, job outcomes and quality of care: test of a structural equation model. *Journal of Advanced Nursing*, 65(10):2175–2185, 2009.

[3] Nathaniel Ham, Amir Dirin, and Teemu H. Laine. Machine learning and dynamic user interfaces in a context aware nurse application environment. *Journal of Ambient Intelligence and Humanized Computing*, 8(2):259–271, 2017.

[4] Sarang Narkhede. Risks of ai – what researchers think is worth worrying about — Towards Data Science, 2018. [Online; accessed 8-March-2020].

[5] Rae Hodge. An antibody test for the novel coronavirus will soon be available — The Economist, 2020. [Online; accessed 18-April-2020].

[6] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, Oct 2018.

[7] Sergey Redyuk, Sebastian Schelter, Tammo Rukat, Volker Markl, and Felix Biessmann. Learning to validate the predictions of black box machine learning models on unseen data. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, HILDA'19, New York, NY, USA, 2019. Association for Computing Machinery.

[8] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[9] David Fumo. A gentle introduction to neural networks series — part 1 — Towards Data Science, 2017. [Online; accessed 5-April-2020].

[10] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[11] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243, 2017.

[12] Dana J Lawrence. The four principles of biomedical ethics: a foundation for current bioethical debate. *Journal of Chiropractic Humanities*, 14:34–40, 2007.

[13] Jonathan H. Chen and Steven M. Asch. Machine learning and prediction in medicine - beyond the peak of inflated expectations. *The New England journal of medicine*, 376(26):2507–2509, Jun 2017. 28657867[pmid].

[14] Andrew L. Beam and Isaac S. Kohane. Big Data and Machine Learning in Health Care. *JAMA*, 319(13):1317–1318, 04 2018.

[15] Bianca Nogrady. The real risks of artificial intelligence — BBC, 2016. [Online; accessed 28-Febuary-2020].

[16] Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLoS medicine*, 15(11):e1002689–e1002689, Nov 2018. 30399149[pmid].

[17] D. Pinto dos Santos, D. Giese, S. Brodehl, S. H. Chon, W. Staab, R. Kleinert, D. Maintz, and B. Baeßler. Medical students' attitude towards artificial intelligence: a multicentre survey. *European Radiology*, 29(4):1640–1646, 2019.

[18] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, Nov 2018. 30128552[pmid].

[19] Rae Hodge. 2019 data breach hall of shame: These were the biggest data breaches of the year — cnet, 2019. [Online; accessed 5-April-2020].

[20] Khader Shameer, Kipp W Johnson, Benjamin S Glicksberg, Joel T Dudley, and Partho P Sengupta. Machine learning in cardiovascular medicine: are we there yet? *Heart*, 104(14):1156–1164, 2018.

[21] Nilay Shah, Ewout Steyerberg, and David Kent. Big data and predictive analytics: Recalibrating expectations. *JAMA*, 320, 05 2018.

[22] Joy T. Wu, Franck Dernoncourt, Sebastian Gehrmann, Patrick D Tyler, Edward T Moseley, Eric T Carlson, David W Grant, Yeran Li, Jonathan Welt, and Leo Anthony Celi. Behind the scenes: A medical natural language processing project. *International Journal of Medical Informatics*, 112:68 – 73, 2018.

[23] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

[24] Sandeep Reddy, John Fox, and Maulik P Purohit. Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, 112(1):22–28, 2019.

[25] Yi Luo, Huan-Hsin Tseng, Sunan Cui, Lise Wei, Randall K. Ten Haken, and Issam El Naqa. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR—Open*, 1(1):20190021, 2019.

[26] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5686–5697, New York, NY, USA, 2016. Association for Computing Machinery.

[27] Keng Siau and Weiyu Wang. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2):47–53, 2018.

[28] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[29] Connor Coley. A graph-convolutional neural network model for the prediction of chemical reactivity. *The Royal Society of Chemistry*, 2019.

[30] Roger Allan Ford, W Price, and II Nicholson. Privacy and accountability in black-box medicine. *Mich. Telecomm. & Tech. L. Rev.*, 23:1, 2016.

[31] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:8869–8879, 2017.

[32] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1893–1905, Nov 2015.

[33] Jennifer Fong Ha and Nancy Longnecker. Doctor-patient communication: a review. *Ochsner Journal*, 10(1):38–43, 2010.

[34] Yaniv Shachor, Hayit Greenspan, and Jacob Goldberger. A mixture of views network with applications to multi-view medical imaging. *Neurocomputing*, 374:1–9, 2020.

[35] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*, pages 1–11. Springer, 2018.

[36] Chris Reed, Elizabeth Kennedy, and Sara Silva. Responsibility, autonomy and accountability: legal liability for machine learning. *Queen Mary School of Law Legal Studies Research Paper*, (243), 2016.

[37] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[38] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.