

# UC Riverside

## UCR Honors Capstones 2021-2022

### Title

A PHILOSOPHICAL ANALYSIS OF DIAGNOSTIC REASONING BY PHYSICIANS AND APPLICATION IN AI SYSTEMS

### Permalink

<https://escholarship.org/uc/item/8k15t2w6>

### Author

Ambat, Bigy

### Publication Date

2022-05-06

### Data Availability

The data associated with this publication are not available for this reason: N/A

A PHILOSOPHICAL ANALYSIS OF DIAGNOSTIC REASONING BY PHYSICIANS AND  
APPLICATION IN AI SYSTEMS

By

Bigy Ambat

A capstone project submitted for Graduation with University Honors

May 6<sup>th</sup>, 2022

University Honors

University of California Riverside

APPROVED

Dr. Erich Reck

Department of Philosophy

Dr. Richard Cardullo, Howard H Hays Jr. Chair

University Honors

## ABSTRACT

Oftentimes, philosophical ideas are at the fundamental root of various disciplines, especially in the sciences. This is also true for the field of medicine. In my capstone, I plan to examine the philosophy of medicine, a relatively new field in philosophy of science, and analyze its application into AI medical diagnosis technology. Throughout numerous studies, Deep Learning AI medical systems have been proven extremely accurate when analyzing CT scans, X-Rays, and other screening tests. While this is very promising for the medical field, computer science experts are having extreme difficulty analyzing the thought processes or rationale that these Deep Learning AI use to make their decisions. This is a well-known problem in computer science called the “black box” problem. To address this “black box problem” in the context of medical diagnosis, I will focus on the intrinsic philosophical basis behind the reasoning and logic that doctors use in screening tests. By analyzing various aspects of clinical reasoning, I will uncover philosophical ideas at the root of the diagnosis process conducted by physicians. Then, I will do an analysis on the reasoning processes involved in AI systems. Finally, I will suggest how diagnosis concepts involved in the physicians’ reasoning process can be integrated into AI diagnosis systems. Along such lines, philosophical ideas about reasoning provide a bedrock that computer scientists can use to further develop these medical Deep Learning AI systems, thereby improving the healthcare industry further.

## ACKNOWLEDGEMENTS

I am well beyond grateful for the mentorship of my faculty mentor Dr. Erich H. Reck. Despite being a philosophy minor, he agreed to mentor me throughout this capstone project. He has been an extraordinarily kind and generous individual throughout the course of this project. He has helped me gain a greater appreciation for philosophy, especially as it pertains to logic, science, and history. I am grateful for this opportunity of learning that has been presented to me.

I would also like to thank philosophy professor Kim Frost for his ideas and input regarding the earlier drafts of my capstone project.

I am thankful to UCR Honors for bestowing me with this capstone opportunity and facilitating my learning experience as an Honors student. Special thanks to Dr. Richard Cardullo for his support of the UCR Honors program.

Finally, I would like to thank my friends and family for their continued support throughout the course of this capstone project.

## INTRODUCTION

Deep Learning Artificial Intelligence (AI) Systems have started to become implemented into the medical system and are proving extremely effective in their diagnosis capabilities (Wilder, 2020). This is especially true when AI systems analyze CT Scans (Sathyakumar et al., 2020), X-Rays (Elkins et al., 2020), disease risk-assessment (Riihimaa, 2020), and ultrasound (Howard et al., 2020). They are also used in risk analysis of diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes (Xu et al., 2020).

While AI systems become more accurate in their assessments over time, they also become increasingly complicated systems to understand. In the computer science field this is referred to as the “black box problem.” In medicine, AI systems have become very accurate in their diagnostic capacities, but a lot of data is needed for AI systems to make their decisions. It is difficult to figure out the rationale (or what AI systems are “thinking”) when those accurate decisions are made. This problem has a lot of ethical and legal implications that must be solved before AI can become fully implemented into the medical field (Jorstad, 2020).

The general goal of this capstone is to examine the reasoning process of physicians and find philosophical and logical ideas at the core of that process. Later, an analysis of deep learning AI systems will reveal the ideas behind their reasoning processes. Analyzing Deep Learning AI in this way could have a significant impact on medicine in several ways. For instance, this work could further improve the diagnosis accuracy of AI systems, possibly even surpassing human capabilities. More particularly, this analysis could allow researchers to discover unique mechanisms that AI systems utilize to “reason” better than humans. These specific reasoning pathways could also be added to the physician reasoning “tool belt”, thus improving the accuracy of medical diagnosis by physicians. On the other hand, the analysis may reveal that physicians

excel in specific clinical situations over AI systems. In these situations, the reasoning mechanism used by the physicians could then be implemented to improve the capabilities of the AI systems. Both of these aspects could significantly improve medical diagnosis, thereby providing higher quality of care for those in a hospital. All of these aspects will play a role in what follows.

## METHODOLOGY

Here is the general structure of this capstone project. The discussion has 6 distinct parts: **1.** Applications of AI Systems in Medicine Today; **2.** Introduction of Physician Case Studies and their Context; **3.** The Architecture of Deep Learning AI Systems and AI Case Studies; **4.** Tools from Logic and the Philosophy of Science; **5.** Using Philosophy for Improving Deep Learning AI Systems; and **8.** A Functional Sub-Logical System for Improving Deep Learning AI.

The 1<sup>st</sup> section will provide an introduction to various applications of AI technology in medicine today. This will include a brief description of the broad abilities AI systems have that have been created for medical applications. The section will emphasize the efficacy of AI diagnoses and their comparison to human counterparts in various clinical scenarios.

In the 2<sup>nd</sup> section, it is vital to provide a general understanding of the diagnostic reasoning processes used by human doctors. To do this, it will be beneficial to consider various case studies that exhibit physicians' diagnostic reasoning protocols, especially in relation to CT Scans, X-Rays, and other medical imaging devices. The goal here is to offer initial descriptions of a few cases where human doctors draw conclusions based on evidence. To do that, references to some medical textbooks will be provided, such as *Felson's Principles of Chest Roentgenology* (L. Goodman, ed.). These textbooks include a wide range of clinical cases which help to understand the physicians' reasoning processes in diagnosis. The primary focus here will be on X-Rays and CT Scans, as they are the most prevalent kind of medical scans administered.

The 3<sup>rd</sup> section will concern the basics of AI technology and the fundamental processes that occur in AI systems. This will include an initial description of how AI systems draw conclusions from data. Next, several cases of AI implementation into the clinical setting will be discussed further, building on the 1<sup>st</sup> section. As we will see, initially this did not involve deep analysis or deep learning, while more recently this aspect has been added. A brief introduction to the philosophical literature relevant to AI will also be provided.

In the 4<sup>th</sup> section, it will be crucial to expand on the philosophical ideas and nuances that are relevant in this connection. Some ideas from logic and the philosophy of science will be brought in to further elaborate on the differences in the reasoning used by human doctors and AI systems. Philosophers of science have analyzed various aspects of scientific inquiry, especially of scientific methodology. Diagnosis in the clinical setting involves implicit use of scientific methodology, especially in differentiating between several plausible diagnostic conclusions. In addition, there are various logical concepts that apply to what human doctors or AI systems do. This includes: deductive inference, with some examples; inductive inference, again with some examples; abductive reasoning, i.e., “inference to the best explanation”; also causal reasoning, where the latter may involve either general laws or mechanisms/models.

For the 5<sup>th</sup> section, the essay will revisit the reasoning processes used by human doctor and AI system. Building on the previous sections, it will now be much easier to analyze these processes in a philosophical manner. That means re-describing, with refined philosophical concepts, what human doctors and AI systems do in various cases. This will make the differences between them clearer. It will also bring out the strengths and the limits of both sides. And it allows us to develop a new philosophical theory about the way AI systems “reason”.

Finally, with all this information in place, the goal of the 6<sup>th</sup> section is to use ideas from formal logic to construct a new symbolic system that mimics the reasoning of physicians. The suggestion will be to implement this system into the AI architecture to further enhance its diagnostic abilities and expand the scope of its application in the clinical environment.



## 1. APPLICATIONS OF AI SYSTEMS IN MEDICINE TODAY

Recent studies have shown that AI systems are competent in medical diagnostics, especially image based radiology systems, such as CT Scans, X-Rays, and Ultrasounds. In their review article, “Intersection of artificial intelligence and medicine: Tort liability in the technological age” (2020), Jorstad and his co-authors discuss the increasing prevalence of AI technology in medical diagnosis systems. They first survey the history of AI technology and how it eventually developed into Machine Learning systems. A subset of these systems, called Deep Learning AI Systems, has started to be implemented into medicine recently. They also describe how Computer Aided Systems (CADs) have already been implemented into the medical technology in most hospitals. Deep Learning AI Systems have further improved CAD systems and become more useful for physicians in their diagnosis process.

Throughout another review article, “Impact of machine learning and feature selection on type 2 diabetes risk prediction” (2020), Riihimaa and colleagues analyze the results of an AI system attempting to predict a patient’s risk for Type 2 Diabetes in the future. While other algorithms exist for risk assessment, those algorithms are not AI based. An AI based algorithm is effective in creating more accurate and more patient-specific predictions. The authors note how older Machine Learning systems are not as effective as modern Deep Learning AI. These older systems have an accuracy (when compared to a physician) of 70- 80%. This shows the importance of having Deep Learning AI systems, whose accuracy ranges from 80% to 99% (in comparison to a physician). The authors show how these kinds of systems can also be effective when testing genetic conditions and doing DNA analysis for predicting genetic diseases.

In a third review article, “Physician-assist automated AI lung cancer detection: A narrative review” (2020), Sathyakumar and colleagues analyze the effectiveness of AI systems in

lung cancer detection based on CT scans. In their study, an AI was allowed to “look” at a group of CT scans so as to determine if there is lung cancer detectable in the scans. Then, an expert physician was given the same group of CT scans and the physician gave a diagnosis of whether the CT scan showed signs of lung cancer or not. The AI Deep Learning system was also varied and updated each time with different kinds of learning algorithms. These algorithms were tested for their efficacy in detecting cancer and they were compared to an expert physician. Upon analyzing the results, the group found that the different AI algorithms were somewhere between 80 and 95% accurate in their diagnosis (when compared to a physician), with one kind of Deep Learning algorithm getting a 99% accuracy (when compared to a physician).

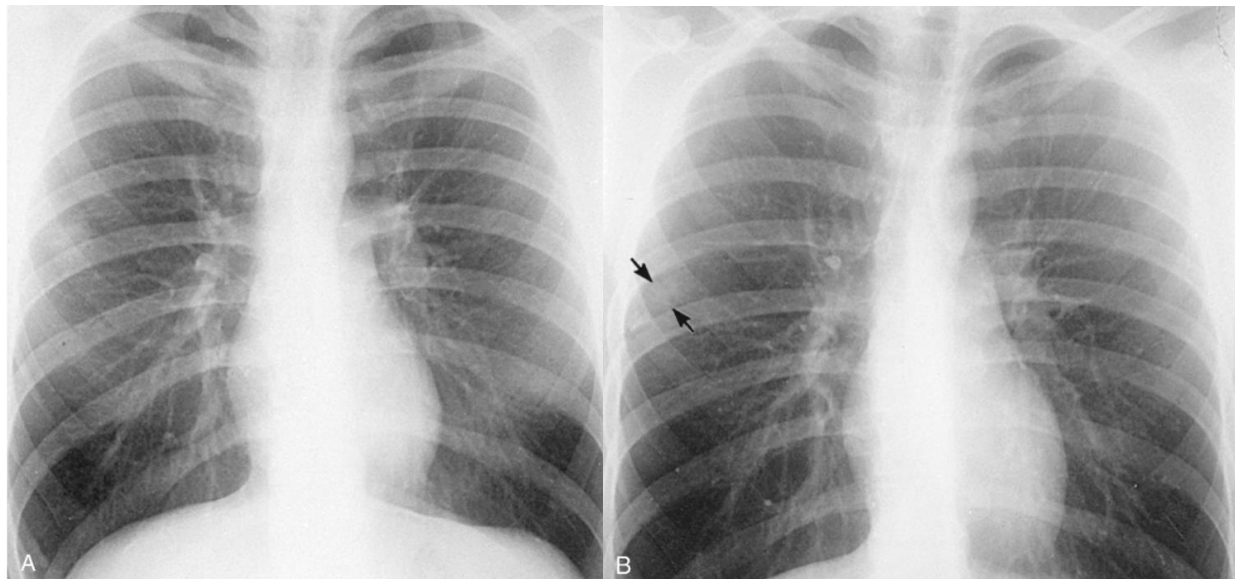
Let us consider one more review article, “General movement assessment by machine learning: Why is it so difficult? (2019), by Schmidt etc. In it the effectiveness of AI systems in properly diagnosing various Cerebral Palsy conditions in children is examined. In this case, the AI systems had a difficult time distinguishing between various types of Cerebral Palsy from short video clips of pediatric patients walking. The study shows that, as of now, this particular task is more effectively accomplished by a doctor. While AI systems excel in certain clinical situations, they struggle with distinguishing relevant features when analyzing moving objects. This area of AI technology needs further development before being fully implemented.

Clearly, AI systems have proven beneficial in a diverse set of clinical situations, from medical diagnosis to risk assessment. While many AI researchers are optimistic in their findings, they often struggle to understand how the AI system “thinks” when it makes decisions. Computer science provides unique parameters to extract data, but the significance of these parameters (or how exactly one should interpret them) remains unclear. Thus, it seems necessary to approach this issue in a new way: a philosophical manner. Since antiquity, philosophers have

been concerned with thinking, rationality, methodology, and epistemology. Perhaps utilizing ideas from the philosophical literature can provide the much-needed clarity and help us understand how AI systems, as well as human physicians, reason? My aim throughout the next three sections is to demonstrate the value of a philosophical approach to AI systems.

## 2. INTRODUCTION OF PHYSICIAN CASE STUDIES AND THEIR CONTEXT

Today Physicians use a variety of imaging technology for arriving at diagnoses about their patients. While other imaging technologies have also been implemented in the last 30 years (CT scans, CAT Scans), no other diagnostic scan is more fundamental than the X-Ray. Even though the X-Ray is a relatively old imaging system, it still produces a surprisingly fast, cheap, and reliable image. Thus, it is still an integral part of the physician's "diagnostic toolbelt" and is the most common imaging procedure prescribed. At the same time, an X-Ray scan is often the most difficult to analyze; only a physician with a trained eye can extract data from it. Often minute differences in saturation, texture, or opacity of an X-Ray distinguish between two similar disease processes. Consequently, the analysis of X-Rays requires strict attention to detail. To illustrate the difficulty of this task, below are two X-Rays of the same patient taken one year apart from each other. The X-Ray on the left represents a normal, unproblematic chest X-Ray result. The X-Ray on the right shows a lung nodule that is barely visible (indicated with the arrows).



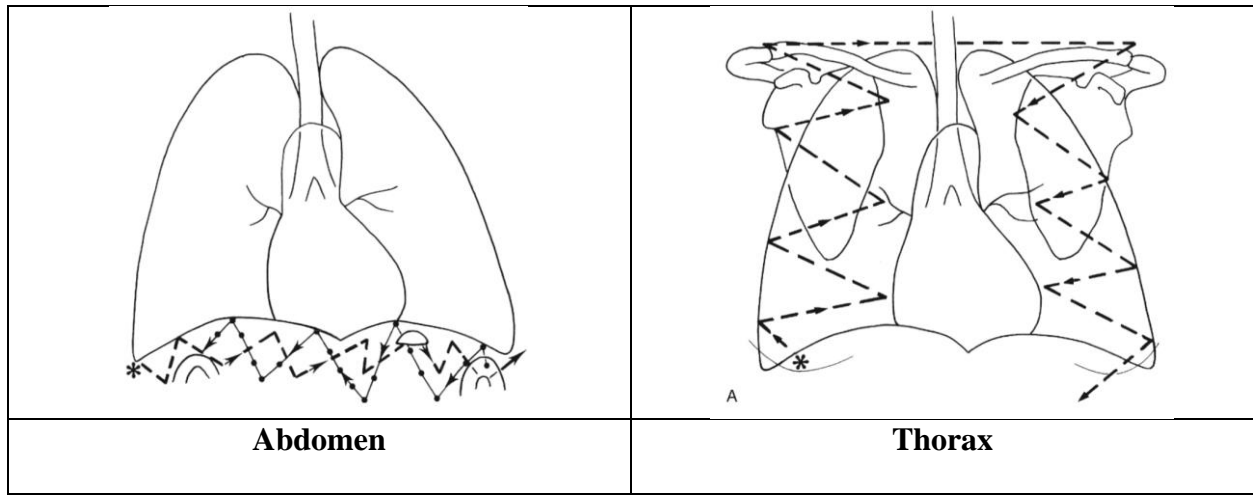
The chest X-Ray, like in our example, is the most common type of X-Ray ordered by physicians. Therefore, analysis on a physician's approach to it can reveal details about the intrinsic methodology of the diagnosis process. Furthermore, since human physicians struggle with learning how to read X-Rays, it is safe to assume that medical AI systems used for X-Ray diagnosis might approach the task differently (at least initially) when learning to extrapolate conclusions from X-Rays. For this reason, I intend to analyze specific reasoning processes and methodologies used in the analysis of X-Ray scans, specifically Chest X-Rays.

Before understanding the diagnostic features used to extrapolate diseases from an X-Ray, we need to consider the characteristics of a normal X-Ray and the general procedures used for reading an X-Ray accurately. An X-Ray beam contains photons which pass through the patient. Some photons are absorbed by the body while others are reflected back to the X-Ray receptor. As a result: "The differential absorption of radiation by different tissues or diseases is responsible for all radiographic shades of gray. Air, fat, soft tissue (muscle, fluid) and metal (bone) absorb progressively more radiation. The thicker the tissue, the more it absorbs" (Goodman 2021, 17). Generally speaking, bone and metal appear near white, tissues appear "greyish", and air appears black. A chest X-Ray is best taken with a full breath (inspiration) to insure greater contrast between tissues (which appear gray) and air (which appear black). The silhouettes of all major organ groups in the chest are then clearly visible (heart, diaphragm, lungs). Any opacity or unusual shapes of the silhouettes of these organs tend to indicate medical issues. Here is an example:

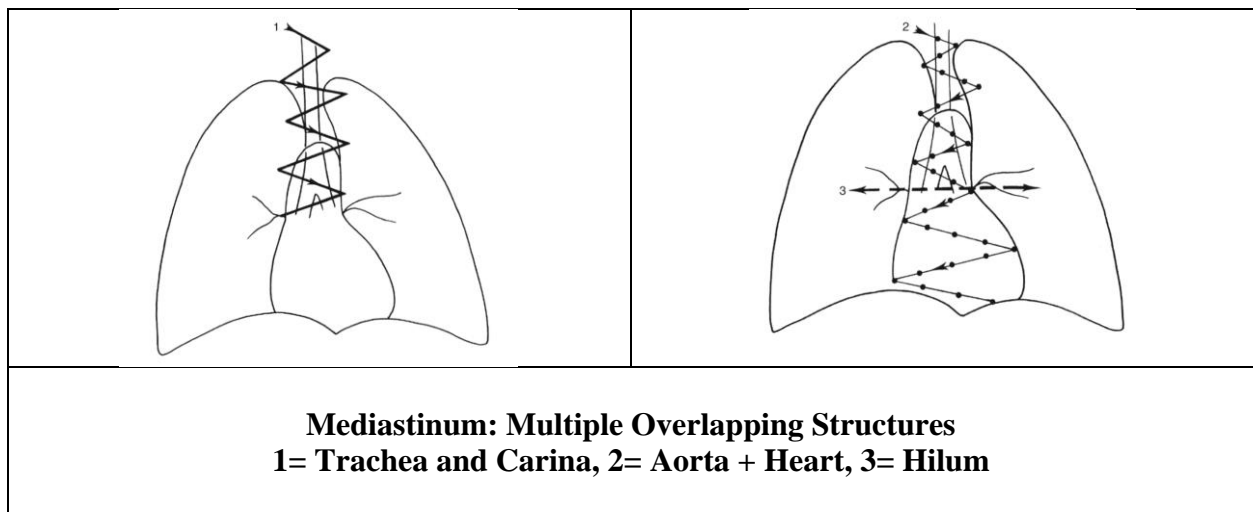


Physicians follow very specific procedures to ensure proper analysis of a Chest X-Ray:

“To maximize your accuracy, [a physician] must have an organized search pattern. Start reading every radiograph—chest or otherwise—by scanning the areas of least interest first, working toward the more important areas. [You] are less likely to miss secondary but important findings this way. For the chest x-ray, start in the lower abdomen, then look at the thoracic cage (soft tissues and bones), then the mediastinal structures, and finally the lung. Look at each lung individually, then compare left lung and right lung. A helpful mnemonic is ‘Are There Many Lung Lesions?’ (ATMLL)” (Goodman 2021, 41). The reasons for such specific pathways are obvious. They insure that important indicators of the X-Ray are not missed due to human error or negligence. Oftentimes, physicians stop searching for symptom indicators when one indicator is found. This is problematic since patients can have multiple, different disease processes involved in the symptoms they are experiencing. Missing multiple abnormalities in an X-Ray (even if they are not as significant as the initial abnormality discovered) constitutes negligence in the physician’s role and can detrimentally impact the patient’s treatment plan following their X-Ray.



Search patterns (starting with \* and ending with an arrow) to efficiently analyze various aspects of a chest X-Ray.



Having learned the basics of X-Ray analysis, we can consider some real-world cases where X-Rays are involved. The following are two corresponding cases studies, taken again

from *Felson's Principles of Chest Roentgenology*, which demonstrate the application of chest X-rays in the physician diagnosis process. Although details for the case studies are explained in the rest of this section, their fuller analysis will be conducted later, upon establishment of the philosophical principles behind physician and AI reasoning processes.



8 This is a middle-aged woman with a cough and fever for 3 days.



8

1. On [A](#) and [B](#), there is a 3-cm area of consolidation in the \_\_\_\_\_ lobe.

1. left lower

2. There is blunting of the \_\_\_\_\_ costophrenic angle.  
 a. left  
 b. right  
 c. both  
 d. neither

2. a. left

Note that the blunting is against the small ribs (less magnified), therefore on the left. See [Figs. C](#) and [D](#) to the right.



3. The history and x-ray findings support a diagnosis of

3. pneumonia. The x-ray is also compatible with a cancer, but the 3 day history suggests otherwise

### *First Case Study*

In this first case, a patient complains of cough and a fever for three days. For any issue related to coughing, the most reasonable step is to conduct a chest X-Ray. Blunting of the left costophrenic angle is the sharp angle of the bottom left most part of the lung. This particular feature is very difficult to see by a normal chest X-Ray. Hence a lateral chest X-Ray (from the side of the patient) is also conducted to ensure a more specific location of the consolidation. After analysis, the physician is left with two diagnostic options: either cancer or pneumonia. The physician takes the patient history into account before making a decision. The physician reasons that it is extremely unlikely that this illness is a cancer diagnosis because of the timeframe of the patient complaint. If the patient had been experiencing these symptoms for over three weeks, it would have given the doctor a stronger reason to conduct a cancer test to before making a final diagnosis.

**11** This is a young male with mild shortness of breath for several months.

**11**



1. In Fig. 11A, the hilar structures are \_\_\_\_\_.

- a. large
- b. normal
- c. small

1. a. large

The most frequent cause of bilateral large hila in a young patient is \_\_\_\_\_.

- a. big pulmonary arteries
- b. adenopathy
- c. idiopathic

b. adenopathy

2. In Fig. 11B, the lungs show \_\_\_\_\_.

- a. cysts
- b. nodules
- c. fibrosis and distortion

2. b. nodules

3. The best diagnosis would be \_\_\_\_\_.

- a. sarcoidosis
- b. lymphoma
- c. miliary tuberculosis

3. a. sarcoidosis

### *Second Case Study:*

In this second case, a young male patient arrives complaining of mild shortness of breath for several months. Upon examining X-Ray, the hilar structures appear enlarged. Furthermore, there are nodules that can clearly be seen surrounding this hilar structure. Based on the patient's age and history, the physician links the patients large hilar structures with the large lymph nodes detected on the patient during the physical exam. As the patient is young, the physician takes the likelihood for a chronic issue to be low. It is more likely that an acute illness is affecting the patient. The physician identifies various "dots" near the hilar structures. The grain-like texture of these "dots" are nodules, a feature of some sort of lung inflammation. Based on these symptoms (especially the swollen lymph nodes), the physician correctly diagnoses the patient with sarcoidosis.

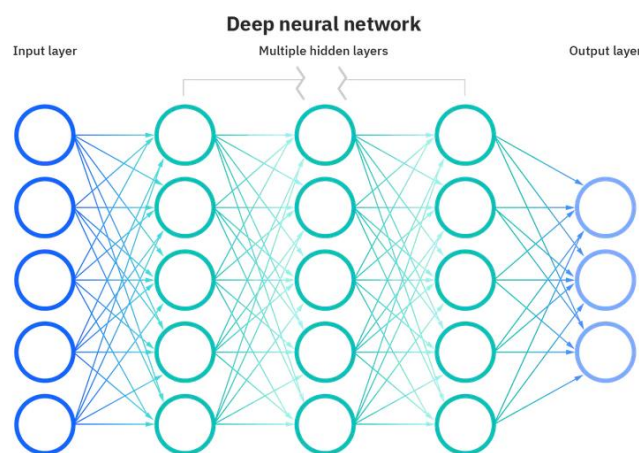
### 3. THE ARCHITECTURE OF DEEP LEARNING AI SYSTEMS

In the previous section, I introduced examples of the reasoning processes primarily involved in the use of X-Rays by human physicians. The low resolution of X-Ray images makes them quite difficult and tedious to read, even for skilled physicians. Therefore, one might assume that an AI diagnostic system could be more accurate and less prone to errors than a physician, particularly when analyzing this type of image. While this statement is mostly true, AI systems have their own limitations and can be more prone to errors in specific scenarios. This section consists of an explanation of relevant AI systems, including their development, their characteristics in general, and their specific functionality in modern medicine. Related to this, two scientific studies involving AI systems will be introduced to further evaluate AI's efficacy in various diagnostic situations, thus to reveal the strengths & weaknesses of these systems.

To fully understand the role AI can play in medicine, it is crucial to understand the basics of how AI systems work and their ability to “reason” to a particular conclusion. Recently, a new form of AI—Machine Learning AI, and especially, Deep Learning AI—has established itself as the most effective, practical, and flexible in various medical diagnosis situations. Indeed, the development of “Deep Neural Networks” has revolutionized AI application in a variety of domains, including autonomous vehicles, manufacturing robots, and financial investing. Moreover, various scientific studies of such networks in the medical setting have achieved results even superior to physicians in certain situations. How do convoluted neural networks function and what characteristics distinguish them from other kinds of AI?

The architecture of Deep Learning neural networks are inspired by the biological functions of neurons in brain tissue. But how do they work? Melanie Mitchell, a leading AI researcher and philosopher, has explained the inner workings of AI, in general and for such

neural networks in particular, in her book, *Artificial Intelligence: A Guide for Thinking Humans* (2020). She compares them to the neural pathways used for movement in humans. In the human system, the brain sends signals through neurons to muscles cells, which respond by contracting. In that case, the brain signal is the input, the neurons in-between form a hidden layer, and the muscle contraction is the output. Similarly, “Deep Neural Network” AI systems consist of an input layer, multiple hidden layers, and an output layer (in order). Inside each hidden layer are units, which are analogous to neurons. Mathematically, this can be described as follows: “Each unit multiplies each of its inputs by the weight on that input’s connection and then sums the result [...] Each unit uses its sum to compute a number between 0 and 1 that is called the unit’s ‘activation’. If the sum that a unit computes is low, the unit’s activation is close to 0; if the sum is high, the activation is close to 1. [This unit transfers its information to the next layer and this process continues until the final output layer.] The activation of an output unit can be thought of as the network’s confidence; the unit with the highest confidence [in the output layer] can be taken as the network’s answer” (Mitchell 37). It thus looks as follows:



(*What Are Neural Networks?*, 2021)

A deep learning algorithm has the ability to change its own characteristics, thus allowing it to respond to various kinds of situations and draw conclusions. How does the deep learning algorithm automatically auto-correct itself when it reasons to an incorrect conclusion initially? As Mitchell explains: “Artificial intelligence systems are composed of a model (representing the learned knowledge), a decision function (making it possible to answer to the problem when a new input is given) and an evaluation metric (to evaluate the quality of the answer provided by AI compared with the ground truth).” Deep neural networks are composed of layers of interconnected artificial neurons forming a “model.” The architecture of the network and the weights associated with each connection represent a “decision function.” From an input (e.g., a histopathological image), the neural network provides a prediction as an output (e.g., cancer or not cancer). In order to learn, the algorithm automatically optimizes its solution by calculating an evaluation metric function, which is basically the difference between the output proposed by the algorithm and the ground truth.

In Deep Neural Networks, the error computed by the evaluation metric is back-propagated through the layers of the network, and the algorithm modifies the weights of the connections between the neurons (cf. Pelaccia et al., 2019) This modification of weights eventually results in the AI system “learning” to match its output value to the ground truth value. Usually, the ground truth value derives from human analysis, in which a human looks at the same inputs as the AI system and reaches a conclusion. As a result, the AI system becomes increasingly more accurate in its conclusions compared to the human. This is analogous to reinforced learning in the training of animals. Oftentimes, the relevant training occurs by means of data sets specifically compiled for the AI system to learn. These data sets contain thousands of cases and are specifically compiled so as to yield the most variation in the learning process.

In a medical study conducted in 2020, entitled *Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents*, Joy Wu and colleagues analyze the efficacy of Anterior Posterior (front to back) Chest X-Ray conclusions generated by an AI algorithm in comparison to five radiology residents. (As mentioned in the human reasoning section above, X-Rays are the most difficult scans for physician doctors since they are lower resolution compared to more recent scans like CT scans.) In this study, the researchers created a new data set from a previously existing NIH CheXpert data set and an MIMIC CXR data set. This was done to create the most varied distribution of chest X-Rays, mimicking real world cases seen by professional radiologists. Three Board-Certified radiologists reached a triple consensus on the diagnosis of these chest x-rays to generate the ground truth value. Furthermore, these cases were categorized by type of case and given a label indicating the diagnosis, creating a lexicon of case studies as seen in Table 1.



Finding		Samples in modeling data set, No.	AUC of AI algorithm
Type	Label		
Anatomical	Not otherwise specified opacity (eg, pleural or parenchymal opacity)	81 013	0.736
Anatomical	Linear or patchy atelectasis	79 218	0.776
Anatomical	Pleural effusion or thickening	76 954	0.887
Anatomical	No anomalies	55 894	0.847
Anatomical	Enlarged cardiac silhouette	49 444	0.846
Anatomical	Pulmonary edema or hazy opacity	40 208	0.861
Anatomical	Consolidation	29 986	0.79
Anatomical	Not otherwise specified calcification	14 333	0.82
Anatomical	Pneumothorax	11 686	0.877
Anatomical	Lobar or segmental collapse	10 868	0.814
Anatomical	Fracture	9738	0.758
Anatomical	Mass or nodule (not otherwise specified)	8588	0.742
Anatomical	Hyperaeration	8197	0.905
Anatomical	Degenerative changes	7747	0.83
Anatomical	Vascular calcification	4481	0.873
Anatomical	Tortuous aorta	3947	0.814
Anatomical	Multiple masses or nodules	3453	0.754
Anatomical	Vascular redistribution	3436	0.705
Anatomical	Enlarged hilum	3106	0.734
Anatomical	Scoliosis	2968	0.815
Anatomical	Bone lesion	2879	0.762
Anatomical	Hernia	2792	0.828
Anatomical	Postsurgical changes	2526	0.834
Anatomical	Mediastinal displacement	1868	0.907
Anatomical	Increased reticular markings or ILD pattern	1828	0.891
Anatomical	Old fractures	1760	0.762
Anatomical	Subcutaneous air	1664	0.913
Anatomical	Elevated hemidiaphragm	1439	0.775
Anatomical	Superior mediastinal mass or enlargement	1345	0.709
Anatomical	Subdiaphragmatic air	1258	0.75
Anatomical	Pneumomediastinum	915	0.807
Anatomical	Cyst or Bullae	778	0.76
Anatomical	Hydropneumothorax	630	0.935
Anatomical	Spinal degenerative changes	454	0.818
Anatomical	Calcified nodule	439	0.736
Anatomical	Lymph node calcification	346	0.603
Anatomical	Bullet or foreign bodies	339	0.715
Anatomical	Other soft tissue abnormalities	334	0.652
Anatomical	Diffuse osseous irregularity	322	0.89
Anatomical	Dislocation	180	0.728
Anatomical	Dilated bowel	92	0.805
Anatomical	Osteotomy changes	76	0.942
Anatomical	New fractures	70	0.696
Anatomical	Shoulder osteoarthritis	70	0.698
Anatomical	Elevated humeral head	69	0.731
Anatomical	Azygous fissure (benign)	47	0.652
Anatomical	Contrast in the GI or GU tract	17	0.724
Device	Other internal postsurgical material	26 191	0.831
Device	Sternotomy wires	12 262	0.972
Device	Cardiac pacer and wires	12 109	0.985
Device	Musculoskeletal or spinal hardware	5481	0.848
Technical	Low lung volumes	25 546	0.877
Technical	Rotated	3809	0.803
Technical	Lungs otherwise not fully included	1440	0.717
Technical	Lungs obscured by overlying object or structure	653	0.68
Technical	Apical lordotic	620	0.716
Technical	Apical kyphotic	566	0.872
Technical	Nondiagnostic radiograph	316	0.858
Technical	Limited by motion	290	0.628
Technical	Limited by exposure or penetration	187	0.834
Technical	Apices not included	175	0.822
Technical	Costophrenic angle not included	62	0.807
Tubes and lines	Central intravascular lines	57 868	0.891
Tubes and lines	Tubes in the airway	32 718	0.96
Tubes and lines	Enteric tubes	27 998	0.939
Tubes and lines	Incorrect placement	11 619	0.827
Tubes and lines	Central intravascular lines: incorrectly positioned	4434	0.769
Tubes and lines	Enteric tubes: incorrectly positioned	4372	0.931
Tubes and lines	Coiled, kinked, or fractured	4325	0.857
Tubes and lines	Tubes in the airway: incorrectly positioned	1962	0.919

Table 1

The AI tested in this study is a Convoluted Neural Network (CNN), a specific kind of Deep Learning Algorithm which is useful in feature detection situations (for instance when

analyzing an image). A combination of previously existing CNNs, specifically VGGNet and ResNet, was used to generate the architecture of the CNN for this study. A full model of the CNN architecture is given in Figure 1 below.

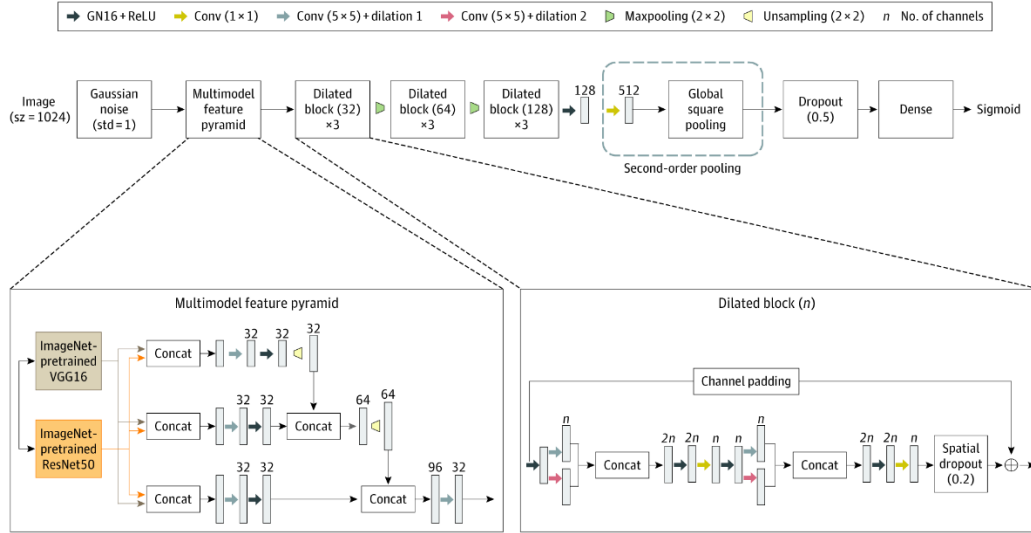


Figure 1

Finally, the AI was trained by using the aforementioned data set and results were compared to the three radiologists' conclusions. Results, statistically compiled and sorted based on case type, can be seen in Figure 2. The X and Y axis represent the specificity and sensitivity (respectively) produced from statistical analysis. The AUC (Area under Curve) value represents the accuracy of the AI algorithm in diagnosing these chest X-Rays. The **black dot** represents the average specificity & sensitivity value for the AI. The **blue square** represents the average specificity & sensitivity value for the radiologist. For reference, specificity and sensitivity are statistical parameters used to analyze the true positive and true negative rate, respectively. If the black dot is higher than the blue square (on both axes), the radiologist is more effective for those specific type of diagnosis for the chest X-Ray. Vice versa for the Neural Network AI system.

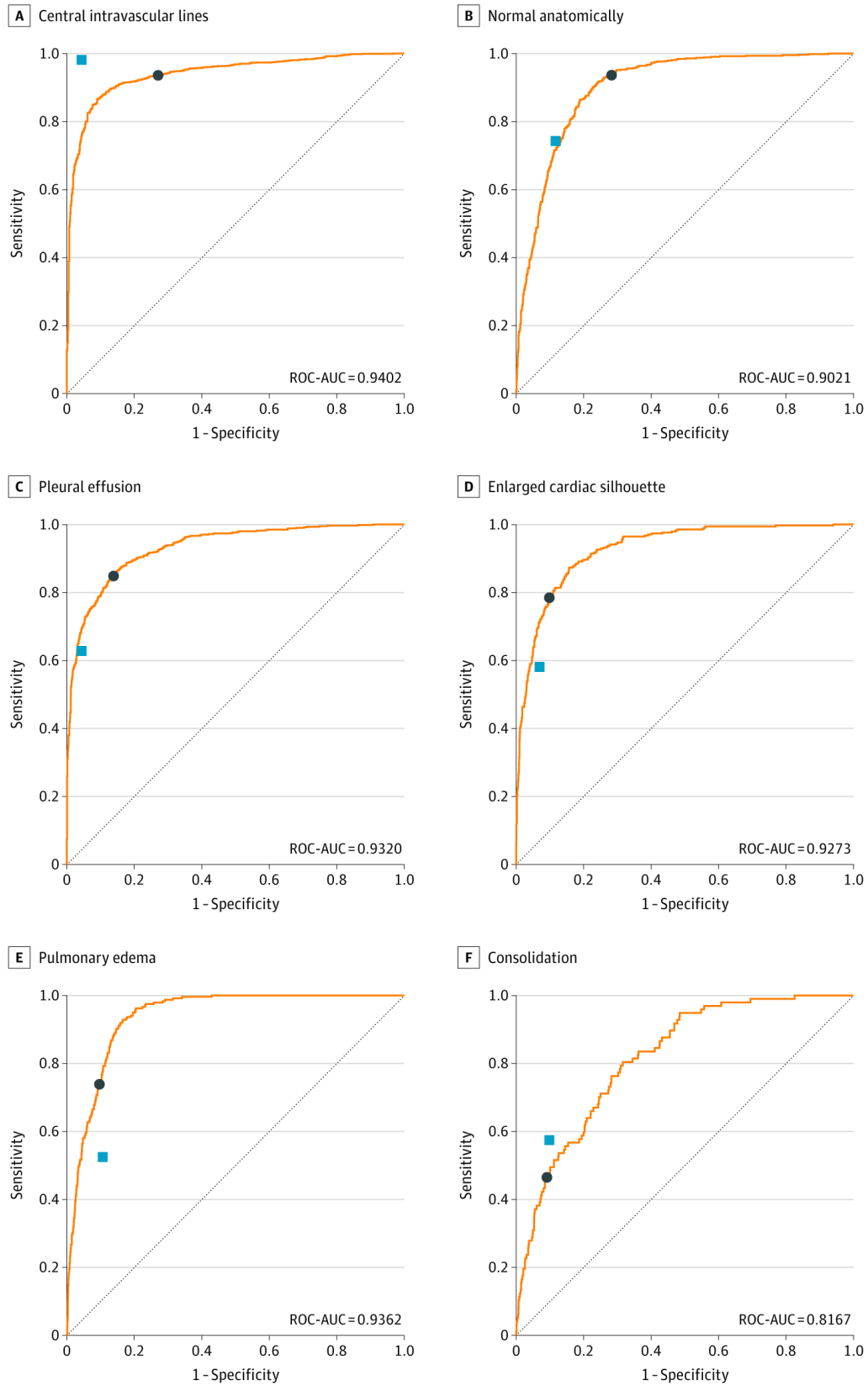


Figure 2

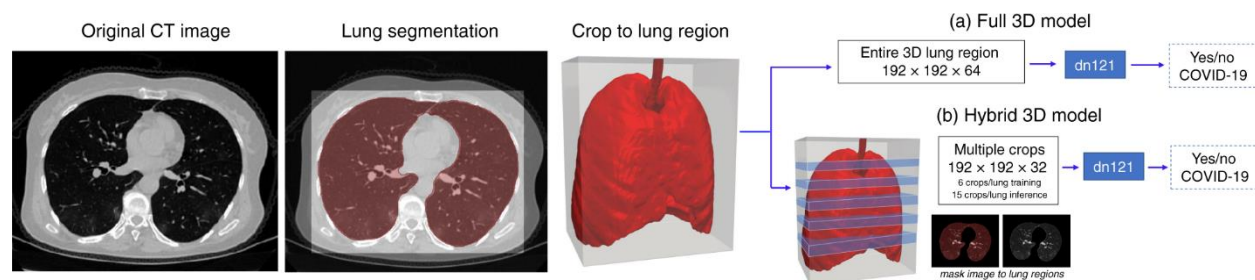
The study reveals interesting results: “Overall, the AI algorithm performed similarly to residents for tubes and lines and non-anomalous reads, and generally outperformed for high-prevalence labels, such as cardiomegaly, pulmonary edema, subcutaneous air, and hyperaeration. Conversely, the AI algorithm generally performed worse for lower-prevalence findings that also had a higher level of difficulty of interpretation, such as masses or nodules and enlarged hilum.” This can be generalized: Current AI algorithms are more accurate overall in their diagnosis in comparison to physicians in the majority of case studies.

Still, AI systems have their own strengths and weaknesses. In general, they struggle in cases that require high sensitivity. On the other hand, they excel in scenarios that involve less ambiguity, which are often cases requiring high specificity. Intuitively, one might expect AI systems to be more accurate than humans in cases involving smaller, ambiguous abnormalities. The actual result is counterintuitive since the AI algorithm should theoretically “see” all the details of the X-Ray far better than the physician’s eye. The study does not explain why AI systems have these strengths and weaknesses. Perhaps a philosophical analysis of the reasoning involved can help us to understand why AI systems arrive at their conclusions and how AI can be further improved? I will explore this idea further in a later section.

To further illustrate the differences between AI and physician reasoning, it is beneficial to look at another kind of medical scan commonly used in medical practice: the CT scan. The intrinsic nature of the CT scan makes it more costly and time consuming, especially when compared to the X-Ray. However, CT scans generate images with much greater resolution, including the ability to view the images in a cross-sectional manner. This particular scan is conducted to see clearer, discrete details in the chest, which are often too vague or miniscule to be identified in a normal X-Ray scan.

With the relatively fast spread and prevalence of Covid-19 from the virus *SARS-CoV-2*, many healthcare organizations have felt overburdened in diagnosing the large quantity of potential Covid-19 patients. As a result, medical research institutions have been studying the efficacy of AI systems in the proper diagnosis of Covid-19 to alleviate the strain in the medical system. One such study, published in the highly reputable journal *Nature Communications*, is entitled *Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets*. It analyzes the efficacy of a Deep Learning system to diagnose Covid-19 in patients through the use of CT scans. The study also tested the AI's ability to distinguish between Covid-19 and similar (non-Covid) pulmonary diseases using the CT-scan data.

The conclusion of this study is that AI algorithms “trained in a diverse multinational cohort of 1280 patients to localize parietal pleura/lung parenchyma followed by classification of COVID-19 pneumonia, can achieve up to 90.8% accuracy, with 84% sensitivity and 93% specificity, as evaluated in an independent test set (not included in training and validation) of 1337 patients. Normal controls included chest CTs from oncology, emergency, and pneumonia-related indications.” (Harmon et al., 2020)



Interestingly, the scientists in the study used two distinct types of CT scans: Full 3D Model & Hybrid Model. Both types were used independently to discover how the AI system responded to each scan. The result was this: “Training converged at highest validation accuracy of 92.4% and

91.7% for hybrid 3D and full 3D classification models, respectively, for the task determining COVID-19 vs. other conditions. The highest test set accuracy was observed with the 3D classification model (90.8%), with resultant probability of COVID-19 disease demonstrating 0.949 AUC” (Harmon et al., 2020) Overall, the AI system performed better with the Full 3D Model as opposed to the Hybrid Model. This can be attributed to the AI’s specific ability to “see”. Generally speaking, AI systems perform better on tasks that involve more 3D manipulation of the data. Still, pixel density of the data seems to affect the ability of the AI systems to draw accurate conclusions, another surprising result revealed by the study.

In addition, the study found variations in false positives based on the CT-Scan input data. Basically, not infrequently the AI system concluded that COVID-19 was present when in reality other disease processes were responsible for the symptoms. However, the false positive percentage varied based on the disease process: “Misclassification rates in control patients was lowest in patients undergoing CT for oncologic staging and workup (ranging 3.8–5.5% in SUNY, LIDC, NIH datasets) compared with patients with laboratory confirmed pneumonias (10%) and general population of patients undergoing CT as part of clinical care, ranging 2.7–27.3% for general evaluation to acute/trauma-related care” (Harmon et al., 2020). In essence, the more comorbidities were involved in the patient, the less accurate the AI diagnoses were.

From these studies, some differences between human reasoning and AI reasoning start to emerge. There are many factors that could explain why AI systems are quite accurate in some situations but inferior to humans in others. For example: “A physician solves most clinical problems in an intuitive and deductive way, whereas AI problem-solving depends on access to and analytical and inductive processing of large quantities of data that relate to the case” (Pelaccia et al., 2019). This might explain some weaknesses in the AI systems. Perhaps the large

information processing required for AI systems leads to a sort of indecisiveness, akin to “analysis paralysis” in human decision making. Melanie Mitchell describes the unique intricacies of ConvNets (CNNs) in chapter 5 of her book in another way. According to Mitchell, “the kinds of errors made by ConvNets are different. While they also get confused by images containing multiple objects, unlike humans they tend to miss objects that are small in the image, objects that have been distorted by color or contrast filters the photographer applied to the image, and ‘abstract representations’ of objects such as a painting or statue of a dog, or a stuffed toy dog” (Mitchell 2020, 93). Hence, perhaps the saturation or contrast of the chest X-Ray and the CT Scan makes the AI struggle to recognize smaller features on it. In addition, Deep Learning AI systems don’t “classify” the objects they “see” as humans unconsciously do. In that case, integrating a classification mechanism into AI systems might prove beneficial.

Overall, Melanie Mitchell articulates the difference between human and AI reasoning succinctly as follows: The human brain is a symbolic system, whereas the AI is a sub-symbolic system. “Sub symbolic systems seem much better suited to perceptual or motor tasks for which humans can’t easily define rules. You can’t easily write down rules for identifying handwritten digits, catching a baseball, or recognizing your mother’s voice [or recognizing a nodule in a chest X-Ray]; you just seem to do it automatically without conscious thought. As the philosopher Andy Clark put it, the nature of sub-symbolic systems is to be ‘bad at logic, good at Frisbee’” (Mitchell 2020, 41). The ability to see a nodule in a chest X-Ray is a specific skill developed by the physician and honed over time through years of repetition. It is possible that the ability to extrapolate from ambiguous data or images relies on the intuition of the physician in an almost unconscious way, which might be difficult to replicate in an AI system.

#### 4. TOOLS FROM LOGIC AND THE PHILOSOPHY OF SCIENCE

When considering diagnostic reasoning in medicine, it is clear that the methodology underlying diagnosis is similar to scientific methodology. Scientists investigate phenomenon in the natural world and use relevant scientific terminology to explain it. Many of these investigations involve discovering and describing a pattern that emerges from the study. Similarly, when providing medical care, doctors “see” patterns in the imaging results (or symptoms) of their patients. In that connection, the doctors use their medical knowledge and experience to arrive at a diagnostic conclusion concerning the patient’s health problem.

Yet, as much as scientists and physicians would behoove admitting it, these “patterns of discovery” (and their resulting conclusions) are inherently influenced by human bias, scientific background, and explanatory worldviews of the individuals involved. This is a major finding in the philosophy of science literature from the past century, and it applies to medical research as well. Our goals, our background assumptions, and our language implicitly affect scientific theories that are heralded as true. One well-known philosopher of science from the mid-twentieth century who has elaborated on this point is Norwood Russell Hanson.

According to Hanson, the observation of patterns (in science) is inherently based on the background theories and corresponding interpretations of scientists involved. He argues that several individuals can arrive at different conclusions about a phenomenon despite observing the same phenomenon. In such a case, it is their interpretive frameworks that differ, leading to differences in explanations of those phenomena. Hanson gives the astronomical theories of Kepler and Tycho Brahe as an illustration. Both Kepler and Tycho saw the same path of the sun across the horizon. Yet, they arrived at drastically different conclusions about the nature of



planetary motion (heliocentric for the former and geocentric for the latter). Hanson analyzes this fact by distinguishing between two kinds of “seeing”: “seeing as” and “seeing that”.

For Hanson, the logic of “seeing as” applies to the general perceptual case. In another example used by him: “Consider [some] footprint in the sand. Here, all the organizational features of seeing as stand out clearly, in the absence of an 'object'.” In such a case, we can ask: What is that? We don’t know yet, since it is not clear how to characterize and contextualize the situation further. This contrasts with “seeing that” in science, at least when further context and background knowledge is implicated. In Hanson’s words: “[S]eeing that' threads knowledge into our seeing; it saves us from re-identifying everything that meets our eye; it allows physicists to observe new data as physicists, and not as cameras. We do not ask 'What's that?' of every passing bicycle. The knowledge is there in the seeing and not an adjunct of it.” (Hanson 1958, 12-14) One consequence is that, while two scientists may see the same phenomenon (seeing as), they may disagree on the explanation of said phenomenon (seeing that).

Crucially, this disagreement is tied to the language, the contextualization, and the evaluation of the evidence used by the two scientists in their deliberation. This is quite similar in medical diagnosis. For instance, two highly-trained physicians may reach different, even incompatible diagnoses despite looking at the same X-Ray. In this scenario, the one physician may say that the lesions in the X-Ray are indicative of cancer, whereas the other physician may conclude that those lesions indicate a tuberculosis infection. Although both physicians are examining the same X-Ray (and assuming both have equal evidence to support their position), it may be unclear which diagnosis should be seen as more accurate. Hanson highlights such disagreement by comparing it with famous optical illusions, for example when one person sees a rabbit and another a duck in a certain drawing. In that case, most people can switch back and

forth between the two options, while in the scientific case that is often much harder.

Connecting Hanson's work to AI technology reveals a stark difference between AI systems' and human physicians' reasoning processes. While physicians have the ability to categorize, contextualize, and deliberate their diagnostic conclusions ("seeing that"), current AI systems do not have that capacity. For that reason, incorporating a classification system might improve the functionality of these AI systems. However, doing so has proven quite difficult due to the lack of logical coherence between images and sentences. As Hanson notes, "early logical constructionists were inattentive to the difficulties in fitting visual sense-data to basic sentences. Had they heeded the differences between pictures and maps, they might have detected greater differences still between pictures and language. [...] The picture shows x; the statement refers to and describes x" (17). To interpret an X-Ray more effectively, the AI system would have to be able to do the latter; but that involves understanding the effects of a lesion, nodule, etc.

A scientific perspective on AI systems, as used in earlier sections of this essay, employs distinctive parameters and scientific terminology to calculate the efficacy of these AI systems (controlled experiments, statistics, etc.). While such a data-driven methodology is undoubtedly useful, it fails to fully encapsulate the broader scope and function of these AI systems. In this section we already went further to some degree, by bringing in ideas from the philosophy of science. In particular, humans are capable of "seeing that", while it is not clear how AI systems can do that. (We will come back to this point later in the essay.) Another way to extend the usual scientific analysis of AI systems, including with respect to medicine, is to bring in ideas from logic, especially the distinction between several basic kinds of reasoning.

Beginning with the American logician Charles S. Peirce, logicians have distinguished three broad types of reasoning processes: deductive reasoning, inductive reasoning, and

abductive reasoning. It is the interplay between all three that gives humans true flexibility in their thinking and decision making. In most of daily life, humans rely heavily on inductive and abductive reasoning. In contrast, computer programs rely on purely deductive processes. Before unpacking the nuances of AI reasoning and the differences to human reasoning along such lines, it will help to briefly explain first deductive, then inductive, and finally abductive reasoning.

Deductive reasoning plays a fundamental role in all reasoning. It may be best to explain deductive reasoning by using an example. “Deductive reasoning allows you to make statements that are necessitated by facts that you know. For example, you are given two facts: 1. It rains every Saturday. 2. Today is Saturday. Deductive reasoning allows you to determine the true statement that it is going to rain today if today is Saturday. In deductive reasoning, one infers a proposition  $q$ , which is logically sensical from a premise  $p$ .” (Panesar 60). This is the kind of reasoning covered in a typical introduction to logic class (like Phil 008 at UCR). This particular kind of reasoning entails necessity. Essentially, the truth of the premises necessitates the truth of the conclusion. The rules pertaining to deductive logic are often so implicit in human reasoning processes that they are rarely explicitly expressed; but it is possible to do so.

Inductive reasoning is a bit more complex. “Inductive inferences form a somewhat heterogeneous class, but for present purposes they may be characterized as those inferences that are based purely on statistical data, such as observed frequencies of occurrences of a particular feature in a given population. An example of such an inference would be this: 96 percent of the Flemish college students speak both Dutch and French. Louise is a Flemish college student. Hence, Louise speaks both Dutch and French” (Douven, 2021). Notice how in inductive reasoning the truth of the premises does not guarantee the truth of the conclusion. It may be true that 96% of Flemish college students speak both Dutch and French, but that does not necessitate

that Louise speaks Dutch and French. Inductive reasoning is based on assumptions and implications made on the basis of statistical data and facts.

Abductive reasoning is the most common and human-like form of reasoning, although it is not often considered and investigated as such. Essentially, abductive reasoning consists in “inference to the best possible explanation” for a given situation or group of facts. Since abductive and inductive reasoning share many similarities, philosophers disagree about the exact differences between them. The *Stanford Encyclopedia of Philosophy* suggests that “the best way to distinguish between induction and abduction is this: both are ampliative, meaning that the conclusion goes beyond what is (logically) contained in the premises (which is why they are non-necessary inferences), but in abduction there is an implicit or explicit appeal to explanatory considerations, whereas in induction there is not; in induction, there is only an appeal to observed frequencies or statistics” (Douven, 2021).

To further explain the differences between inductive and abductive reasoning, assume that event A occurs. A strict application of inductive reasoning (using all the facts available and relevant to event A) might leave one with the conclusion that event A, while it had a certain probability, has an unknown explanation. In contrast, a strict interpretation of abductive reasoning (using all the facts available and relevant to event A) should lead to the emergence of multiple possible explanations for this event. For that reason, abductive reasoning is much more useful in a practical sense; and it is therefore used widely in almost any aspect requiring human reasoning. On the other hand, the nature of abductive reasoning, especially the creation and comparison of plausible hypotheses in it, makes its implementation into AI difficult.

Despite AI having emerged from computer technology (and indeed having a deductive backbone), AI systems operate in what one way call a pseudo-deductive-abductive manner,

especially Deep Learning systems. The back-propagation algorithm mentioned in the previous section, commonly used in Deep Learning AI, has proven to be very effective in application. In essence, a back-propagation algorithm can be described as a brute-force trial-and-error algorithm with variable parameters. The back-propagation aspect allows these variable parameters to be manipulated in an effort to become more efficient in the designated task. Once the AI system has achieved a certain level of competence, these variable parameters become fixed, eventually resulting in a machine that appears to solve problems using abduction. Yet this pseudo-problem-solving method points to a key difference between human and AI reasoning. While humans use abductive reasoning directly, Deep Learning technology uses deductive reasoning to replicate or simulate abductive reasoning. This is clear from the substructure of the AI neural network.

To understand this point more fully, let us consider an AI system that learns to identify handwritten numbers from 0 to 9. Below is a possible deep learning neural network for this task.

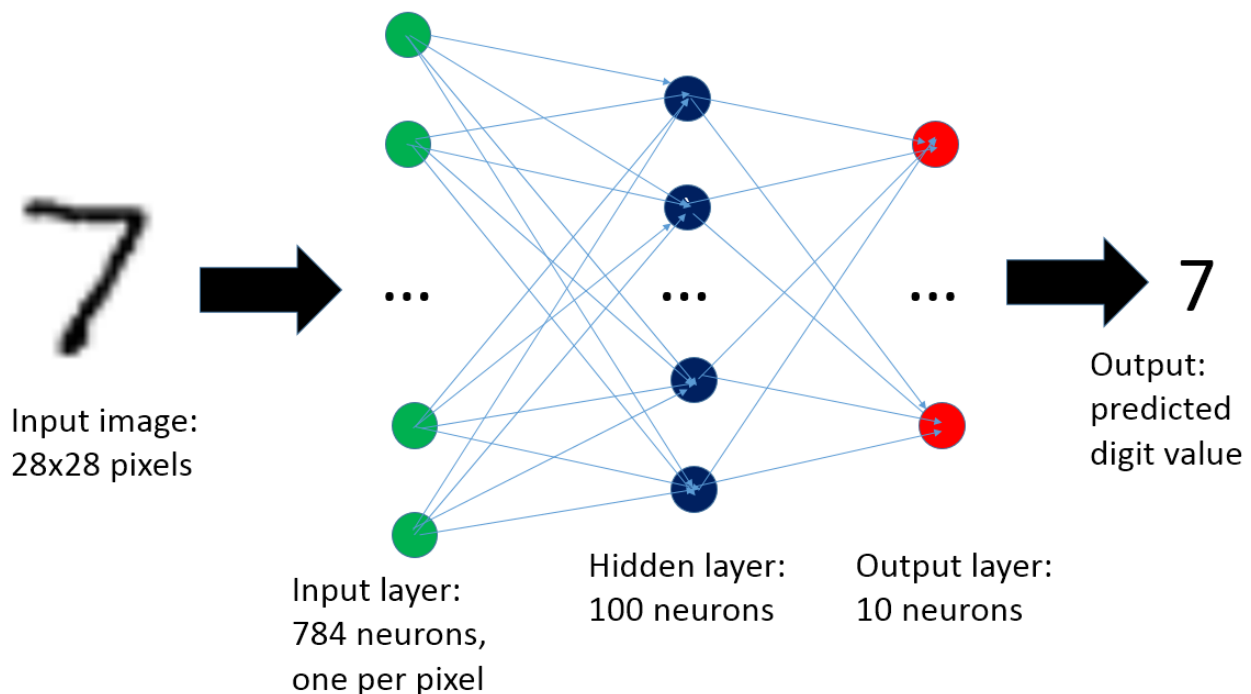


Figure 1

Melanie Mitchell describes this kind of scenario in her book. Similar to what we saw earlier, she explains how “each [unit/neuron] here multiplies each of its inputs by the weight on that input’s connection and then sums the result. [...] Each unit uses its sum to compute a number between 0 and 1 that is called the unit’s ‘activation’. If the sum that a unit computes is low, the unit’s activation is close to 0 [and the signal will not continue to the next layer]; if the sum is high, the activation is close to 1 [and the signal will continue to the next layer]” (Mitchell 2020, 37). The last output layer functions in a similar manner. For the sake of simplicity, assume that each of ten neurons in that layer corresponds to the numbers 0-9 respectively.

Given this information, let us conduct a thought experiment. Assume that an AI system is attempting to identify the handwritten number 9 on a piece of paper. The data from the image moves through the neural network and arrives at the output layer. As mentioned above, each output layer unit/neuron corresponds to the numbers 0-9. The information from the image that passes through the neural network arrives at the output state where two outputs are initially activated, say the numbers 4 & 9. Basically, the AI system is having difficulty distinguishing between two probable outputs. After all the data from the image has passed through the system, the unit activation for the number 9 is higher than the unit activation for the number 4. So, the deep learning algorithm concludes that the handwritten number is 9.

In this hypothetical scenario, the AI appears to reason abductively. The numbers 4 and 9 are both plausible numbers for the input image that the AI “sees”. Upon receiving further information through the neural network, the AI appears to make a final decision in favor of output 9. This shows that (to some extent) AI uses “inferences to the best explanation” when making decisions. This is akin to a human deliberating between option A and option B to explain an event. However, a human reasoner considers relevant factors related to each option before

committing to a decision, as is typical for genuine abduction. The significance of this difference, and of the abductive reasoning capability humans have, cannot be overstated. At the same time, we can see how abductive reasoning can be emulated using a purely deductive AI system.

Before continuing to the next section, a final clarification should be added. The intention of this section (and of the next section) is not to downplay the role or function of deductive logic in human reasoning. Deductive reasoning does play a vital role in almost every field of study imaginable. Instead, the goal is to determine the limits of the deductive reasoning processes used in current AI systems and, perhaps, to develop a better simulation of abductive reasoning based on deductive reasoning. Basically, the goal of the next section is to provide a deductive framework that inductive & abductive reasoning can emerge from. For that purpose, the logic and philosophy of science literature can provide especially useful insight. By combining our three reasoning processes, it may be possible to develop a distinctive Deep Learning AI system that has the potential to be more effective in its application, especially in clinical medicine.

## 5. USING PHILOSOPHY FOR IMPROVING DEEP LEARNING AI SYSTEMS

In an age of ever-increasing medical terminology, the study of medicine has become a vast cornucopia of scientific and methodological research. But while there is plenty of effort involved in the clinical study of disease, few scholars study the reasoning processes intrinsic to a diagnosis in medicine. In this understudied area, philosophy can provide unique models for characterizing and explaining these processes. Most of these models involve a combination of deductive, inductive, and abductive reasoning, as explained in the previous section. The goal of this section is to further explore the deeply philosophical nature of human reasoning and to further reveal the implicit reasoning processes prevalent in medical diagnosis. A related analysis will elucidate the intricacies of medical diagnostic AI technology and make explicit the reasoning processes used in generating a conclusion by it. On the basis of both, human-like reasoning can perhaps be implemented into AI systems to improve their reasoning capabilities in the medical setting.

Lorenzo Magnani, an Italian philosopher of science and cognitive scientist, has contributed extensively to the study of human reasoning, including in the context of medicine. As an expert in computational philosophy and philosophy of mind, he has provided a unique perspective on the mechanisms which underlie the diagnosis reasoning processes used by physicians. In his 2001 book, *Abduction, Reason, and Science: Processes of Discovery and Explanation*, Magnani discusses the different types of abduction used by AI technology and by humans. In the later chapters of the book, Magnani considers clinical reasoning in particular, as well as the process of integrating diagnostic AI into medicine. Although his book contains a large amount of useful knowledge regarding AI reasoning in relation to diagnostic reasoning, it was published well before the creation of Deep Learning AI systems.



As noted in previous sections, Deep Learning AI systems can be significantly more accurate, effective, and efficient in clinical diagnosis compared to physicians in some contexts. The aim of this section is to explore and to extrapolate some of Magnani's ideas with respect to Deep Learning AI systems, thus further explaining their limits in clinical medicine. With these limits fully explicit, the goal will be to find ways to improve the scope and skills of AI technology so that it may be better integrated into the modern clinical setting. By also more fully understanding the physician's multi-faceted approach to diagnosis and treatment, the latter half of this section will develop a proposal that provides AI systems with a new sub-logical system that can be embedded into their current architecture and improve them.

Before explaining medical diagnosis, Magnani characterizes various types of abduction used in science. Among them, he focuses on "selective abduction". This is the main type of abduction used in clinical medicine, as he argues later. As Magnani explains: "Selective abduction is the making of preliminary guesses that introduce a set of plausible diagnostic hypotheses, followed by deduction to explore their consequences, and by induction to test them with available patient data, (1) to increase the likelihood of a hypothesis by noting evidence explained by that one, rather than by competing hypotheses, or (2) to refute all but one" (23).

After that, Magnani presents his Select and Test Model (ST-Model) which uses a particular combination of abduction, induction, and deduction to characterize the diagnosis processes used by physicians. With respect to the notion of abduction, Magnani explains that "there are two main epistemological meanings of the word abduction: 1) abduction that only generates 'plausible' hypotheses and 2) abduction considered as *inference to the best explanation*, which also evaluates hypotheses" (2001, 19). His ST-Model utilizes both the creation and evaluation of specific abductions, and its cyclic nature entails that conclusions are

re-evaluated based on new information or data (see Figure 1).











Figure 1

To understand this process more fully, imagine a physician diagnosing a patient. First, the physician conducts a physical examination of the patient and ask questions about the symptoms. If the diagnosis is obvious (no other possibilities reveal themselves), the physician may make a generalization from the initial data to a specific conclusion (diagnosis). This step would be considered inductive since no other competing explanations are involved. If the symptoms are more complex, the physician may generate multiple hypotheses based on the patient's symptoms and rank them based on the initial data, power of explanation, and/or simplicity (ex. Occam's Razor). The physician may then order more tests to further evaluate the highest-ranked hypothesis, thereby justifying or refuting the initial abductive conclusions.

If the physician is trained properly, he or she will know what to expect from these tests assuming the current diagnosis is accurate. As many of the relevant tests will come out either

true or false (normal or abnormal, respectively), this step in the process can be considered a deduction. If after all this, the test results do not correspond to expectations, the physician will ask for more information about the patient's symptoms, medical history, or other factors relevant for a diagnosis. With that new information in hand, the cycle repeats itself. This cyclic nature of the ST-Model ensures that the system draws defeasible conclusions from incomplete information. As such, this is a form of what is called "non-monotonic logic" in philosophy and in computer science.

This ST-Model seems to accurately describe the reasoning processes used by physicians when diagnosing. For instance, consider one of the clinical case studies involving the use of X-rays from the first section in this essay.

<p>8 This is a middle-aged woman with a cough and fever for 3 days.</p>		8
 		
1. On A  and B  , there is a 3-cm area of consolidation in the _____ lobe.		1. left lower
2. There is blunting of the _____ costophrenic angle. a. left b. right c. both d. neither		2. a. left
<p>Note that the blunting is against the small ribs (less magnified), therefore on the left. See Figs. C  and D  to the right.</p>  		
3. The history and x-ray findings support a diagnosis of		3. pneumonia. The x-ray is also compatible with a cancer, but the 3 day history suggests pneumonia

As a reminder, in this case study a patient arrives at the hospital complaining about cough and a fever for three days. For any issue related to cough, the most reasonable step is to conduct a chest X-Ray. Upon analyzing the X-Ray, the physician notices blunting of the ribs at left costophrenic angle, which the sharp angle of left-most, bottom part of the lung. (This can be seen in Figures C and D.) After a first analysis, the physician is left with two options: either cancer or pneumonia. The physician takes the patient history into consideration before making a decision. As we described the case earlier, the physician reasons that it is extremely unlikely that this illness is a

cancer diagnosis because of the timeframe of the patient complaint. However, if the patient had been experiencing these symptoms for over three weeks, it would have given the doctor a stronger reason to conduct a cancer test before making a final diagnosis.

This scenario can be considered primarily abductive since there are two competing explanations for the physician's diagnosis: pneumonia or cancer. Although the physician did initially consider cancer as an option, it was ruled out easily and a cancer biopsy was not prescribed. The plausible "cancer diagnosis" did not warrant enough concern to be analyzed, so that a diagnostic test wasn't necessary. Here, the physician uses a heuristic to inductively eliminate the "cancer diagnosis". But had the "cancer" diagnosis seemed more plausible, the physician would have prescribed a tissue biopsy to further examine the "cancer" diagnosis. After this, the physician would have waited for the results of the tissue biopsy before making a final decision. Since tissue biopsies predict either cancerous tissue or normal tissue, this part of the procedure can be considered fully deductive.

If the cancer option is ruled out by the test, the physician may conclusively diagnose the patients with pneumonia. On the other hand, if the results of the biopsy reveal new information about the patient that was previously unknown, the physician may restart the cycle again by considering new plausible diagnoses. At that stage, the physician's reasoning can again be classified as fully abductive in the sense that it generates hypotheses, which are then tested deductively and inductively to further confirm or refute the physician's conclusions. Overall, this shows that the physician's reasoning is intrinsically non-monotonic.

In humans, this non-monotonic reasoning is grounded in a process that logicians and psychologists call "belief revision". Magnani explains that "belief revision is a dynamic notion dealing with the current stage of reasoning. At each stage of reasoning, if it is correct, a belief is

held on the basis that the reasoning is justified, even if subsequent stages dictate its retraction” (24-25). Such belief revision intrinsically allows a physician to “backtrack” when a diagnosis is potentially incorrect or invalid. This psychological mechanism establishes itself phenotypically when it allows humans to deliberate on and re-evaluate seemingly well-established beliefs in various areas of human life.

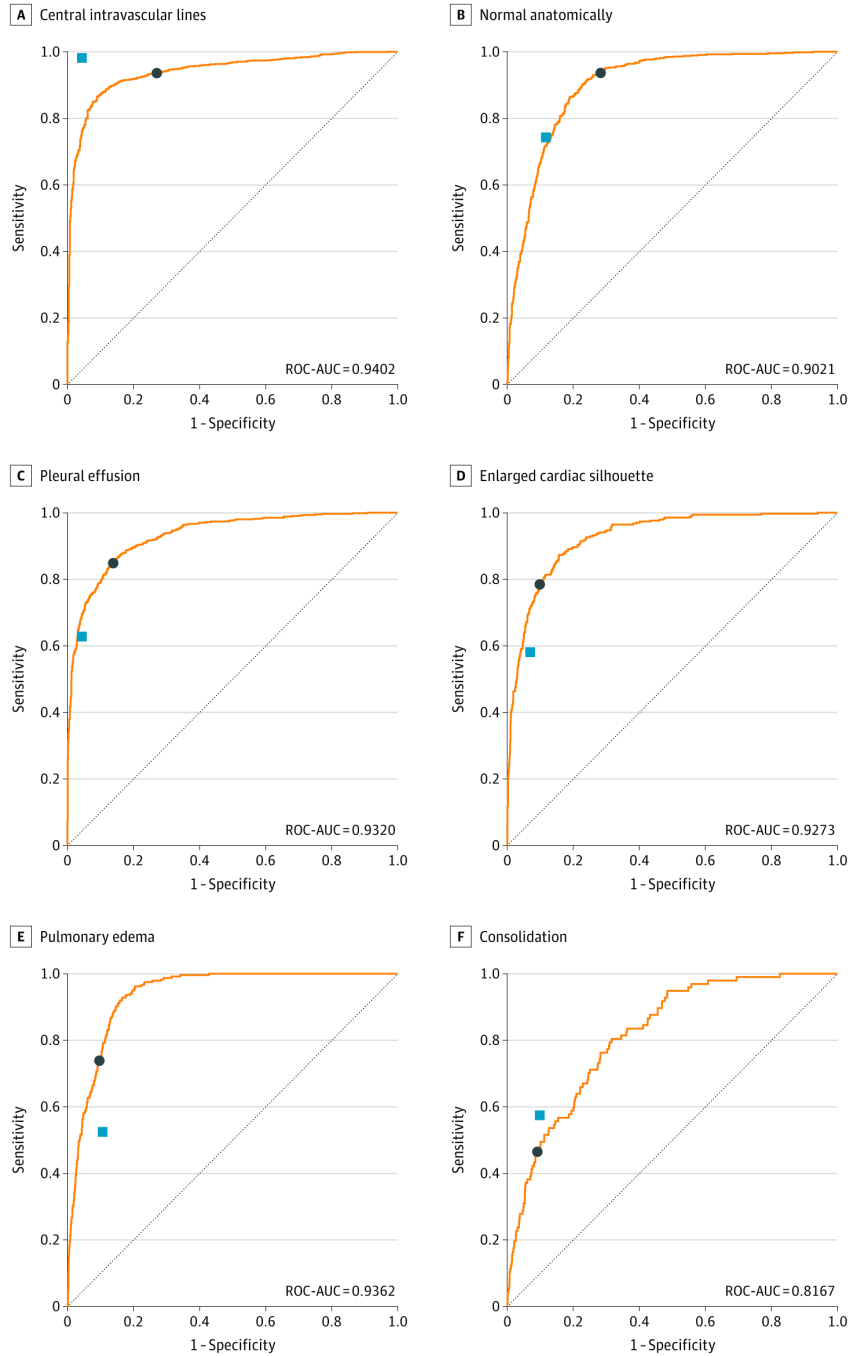
This constant re-evaluation of assumptions is a concept also highly emphasized in medical education, and is a vital, enduring trait of good physicians. In that connection, the best physicians often develop a “differential” (also known as a back-up option) in the scenario in which their original diagnosis reveals itself to be false. With the differential in mind, the physician already has a potential avenue for discovering the true diagnosis if the current one turns out to be false. When student doctors are trained in medical school, such “differentials” are implicitly encouraged to broaden the student’s consideration in the diagnostic process. Having a “differential” in mind also tends to constrain the physician’s ego in diagnosis, in the sense that it may restrain him or her from committing too soon to an unjustified diagnosis.

Based on this analysis, further distinctions between human and AI diagnosis processes reveal themselves. As we saw above, it may seem initially as if AI systems reason abductively in the same manner that physicians reason. More specifically, Deep Learning neural nets, combined with a back-propagation algorithm, may seem to arrive at answers through abduction. Yet, as Magnani has explained the non-monotonic nature of his ST-Model, the distinction between AI and human reasoning becomes clearer. As he adds, the ascending part of his ST-Model incorporates “non-monotonic behavior of limited rationality of commonsense reasoning [which] allows [for the] discharge and abandonment of old hypothesis to make possible the tentative adoption of new ones. Notice that this adoption is not merely tentative but rationally tentative, in

the sense that, just as abduction, it is based on a reasoned selection of knowledge and on some preference criteria which avoid the combinatorial explosion of hypothesis generation” (24).

In humans, including human doctors, the selection criterion used for abduction is based on reasoning and heuristic-based problem-solving strategies. In contrast, in Deep Learning AI technology, corresponding “selection criteria” arise from a constant tweaking of weights and biases of the system with the goal of achieving accuracy in a particular setting. These weights and biases are only held constant after the system is deemed truly accurate for its purpose. Implicitly this involves a “combinatorial explosion” of plausible hypotheses, each “competing” with each other to have greater weight and influence, and each except one then ruled out mechanically. In essence, the AI system arrives at its weights and biases through immense hypothesis generation and brute force correction, in stark contrast to humans. As noted above in passing, the problem-solving method of Deep Learning AI systems constitutes a new style of reasoning: pseudo-abductive reasoning. Such pseudo-abductive reasoning leads to interesting intricacies that can be confirmed by re-considering some of our Deep Learning AI studies.

Recall our X-Ray cases from above. In the 2020 study entitled *Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents*, Joy Wu and colleagues analyze the efficacy of Anterior-Posterior (front to back) Chest X-Ray conclusions generated by an AI algorithm in comparison to five radiology residents. They gauged two statistical parameters to compare between AI accuracy and physician accuracy: sensitivity and specificity. Sensitivity is the true positive rate, while Specificity is the true negative rate. The **black dot** represents the average specificity & sensitivity value for the AI. The **blue square** represents the average specificity & sensitivity value for the radiologist. (See the Figure Below, repeated from earlier in the essay.)



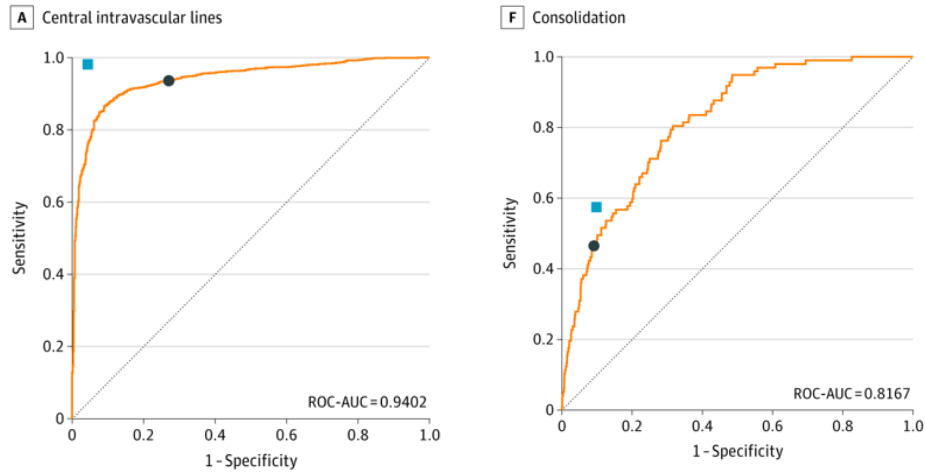
While Magnani does not fully articulate this aspect, since he does not consider Deep Learning AI system, their reasoning should be classified as the pseudo-abductive reasoning described in the previous section. As they arrive at their conclusions by brute-force trial-and-error, Deep Learning neural networks cannot be considered genuinely abductive. Empirical



evidence for this claim can be found in the statistically-relevant deviations between the AI and radiologist average specificity and sensitivity values in the study. As the average values are statistically-different, the AI systems and radiologists must have used different reasoning processes to arrive at their diagnoses. Of course, Deep Learning AI systems can generate accurate conclusions (in certain cases) through the use of this brute-force problem-solving method. Furthermore, both AI systems and physicians can arrive at the same diagnosis despite different problem-solving methods. This is analogous to mathematical problem solving where various mathematical tools and methods can be used to arrive at the same answer.

The pseudo-abductive nature of Deep Learning AI is simultaneously the source of its super-human ability to diagnose correctly in some contexts and its greatest limitation in application to the medical field more generally. Results from the study by Wu and colleagues again demonstrate this fact. In their study, the AI algorithm performed similarly to residents for tubes and lines and non-anomalous reads, and it generally outperformed them for high-prevalence labels, such as cardiomegaly, pulmonary edema, subcutaneous air, and hyperaeration. At the same time, the AI algorithm generally performed worse for lower-prevalence findings that also had a higher level of difficulty of interpretation, such as masses or nodules and enlarged hilum. In general, such AI systems are more accurate in cases that involve high specificity values, while they struggle with specific cases that involve more ambiguity etc..

The benefits and drawbacks of Deep Learning AI systems can also be seen clearly when comparing two classifications of X-Ray diagnosis used in the study: Central Vascular Lines and Consolidation. Recall here the following results:



Central intravascular lines are special IV lines directly connected to the heart. Once placed, an X-Ray is often prescribed to ensure proper placement and function. In this scenario, the diagnosis of the human physician showed high sensitivity and low specificity. Practically speaking, it is relatively easy to distinguish between incorrectly placed and correctly placed central intravascular lines. Unusually, the Deep Learning AI systems deviated significantly from this seemingly easy diagnostic scenario. While they are still reasonably accurate in comparison to physicians, they appear to trade off some sensitivity to increase the specificity in certain types of X-Rays such as Central IV line placement. Such a reciprocal exchange appears to be a unique mechanism used by Deep Learning AI systems to improve their overall accuracy in diagnosis. Actually, this might be a useful methodology for physicians to mimic in their practice.

On the other hand, Deep Learning AI systems are not as accurate in specific scenarios compared to physicians. Once again, Pulmonary Consolidation is a medical condition where specific regions of the lung have filled with liquid, resulting in spotty patches (nodules) in a chest X-Ray. Our study results show that AI systems performed slightly worse in comparison to human physicians for consolidation. In this scenario, the specific limitations of the AI algorithms

emerge. As they utilize an incredibly large data set for training, they have a propensity to over-analyze data sets, leading to inaccurate or misguided diagnoses in real-world application.

## 6. A FUNCTIONAL SUB-LOGICAL SYSTEM FOR IMPROVING DEEP LEARNING AI

Having now discussed the complicated nature of Deep Learning AI systems theoretically in some detail, this last section of the capstone involves considering very practical aspects of current AI technology and suggesting the implementation of a new sub-logical systems into Deep Learning AI systems with the goal of increasing its diagnostic capabilities.

As we saw, human beings have a belief revision system that is vital in medical diagnosis. Oftentimes, proper diagnosis involves retracting or modifying assumptions made by the physician initially. This type of revision is achieved implicitly by means of the non-monotonic nature of physician diagnostic reasoning. In contrast, current Deep Learning AI systems do not have a non-monotonic reasoning component that allows them to “revise” conclusions. Without such a “revision” system, which encapsulates the ascending portion of Magnani’s ST-Model, AI systems struggle to diagnose in situations where they do not “know” the answer quickly. Indeed, in some scenarios (after the training phase), the AI system may be unsure of any diagnosis but may provide one even if it is not “certain” of its accuracy. To circumvent this limitation, my suggestion is the following: We should develop and then incorporate a “confidence parameter”.

During the training phase, AI researchers will be able to find weaknesses or specific situations which may be conducive to errors in the AI system relatively easily. This should allow them to determine an appropriate “confidence parameter”, i.e., a threshold for when we can be relatively sure of the diagnosis that is provided. After the training phase, this parameter can then be interpreted functionally as the AI system “asking for help”. If it falls below a certain value, this triggers an alert to the effect that a physician in charge considers the case him- or herself. In effect, such a “confidence parameter” can serve as an epistemological foundation for the AI’s reasoning, providing a useful framework for assessing the descending portion of the ST-Model.

This “confidence parameter” might be functionally implemented into the neural network pathway of the AI system by using tools from another part of logic, namely what philosophers call the “logic of justification”. Consider the “justification rule”, commonly also referred to as *application*, in such a logic. This logical rule is analogous to multiplication.

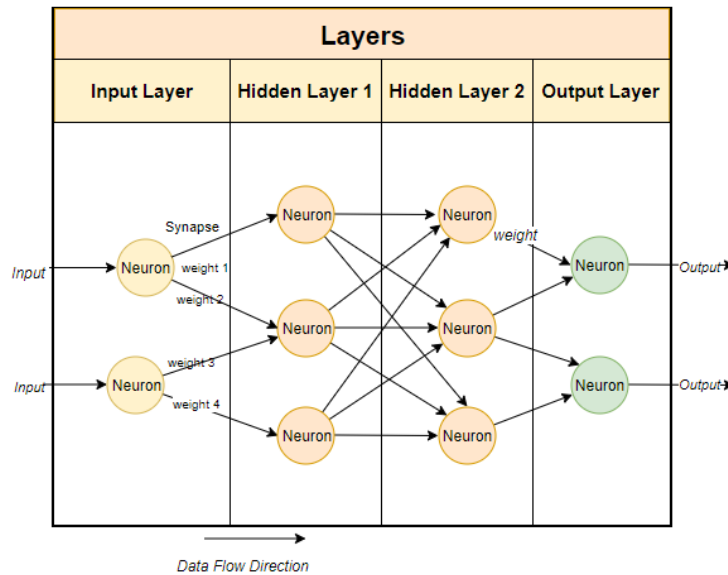
If  $s$  is a justification for  $A \rightarrow B$  and  $t$  is a justification for  $A$ , then  $[s \cdot t]$  is a justification for  $B$

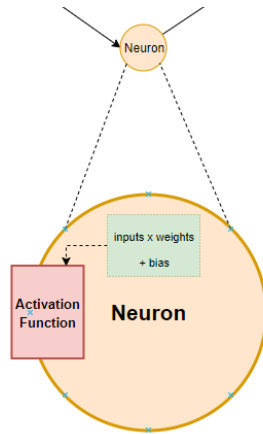
That is, the validity of the following is generally assumed:

$$s:(A \rightarrow B) \rightarrow (t:A \rightarrow [s \cdot t]:B)$$

[Stanford Encyclopedia of Philosophy: Justification Logic]

The suggestion is to use this *application* rule, or something analogous to it, to numerically evaluate the “confidence” of competing hypotheses in a neural network. As a refresher on the computational structure of neural networks, recall the image below:





In the computational architecture of such a neural network, each neuron in hidden layer 1 is connected individually to each neuron in hidden layer 2. The connection between two neurons in adjacent layers (visually represented by the arrow) is numerically represented by the parameter “weight”. A higher weight is analogous to a stronger connection between two neurons in adjacent layers. Based on this, the activation function of a neuron unit is a representation of the sum total of the  $[(\text{inputs} \times \text{weights}) + \text{Bias}]$  of all connecting neurons in the previous layer. The signal that arises from the combination of these computations results in complete propagation through the neural network, eventually arriving at the final output layer. Finally, the output neuron with the highest deemed activation value is interpreted as the most “accurate” conclusion by the Deep Learning AI system.

While such an activation function does convey some epistemic weight of the conclusion, it fails to consider the statistical difference between activation values when generating conclusions. Evidence for this claim can be found in the thought experiment in the previous section. Even if the output neuron for conclusion A is only marginally better than the output neuron for conclusion B, the Deep Learning AI system will robotically portray conclusion A as the most plausible explanation; and it will do so without providing any justification for it (also without mentioning other possibilities). For that reason, some justification parameters is

necessary to provide an better foundation for conclusions reached by Deep Learning AI systems.

Consider the Final Hidden Layer and the Output Layer of the AI neural network described above. Now, consider adding the aforementioned *application rule*. The same kind of rule could be used to create a “confidence value” for the neural net’s diagnostic conclusions.

If  $s$  is a justification for  $A \rightarrow B$  and  $t$  is a justification for  $A$ , then  $[s \cdot t]$  is a justification for  $B$

That is, the validity of the following is generally assumed:

$$s:(A \rightarrow B) \rightarrow (t:A \rightarrow [s \cdot t]:B)$$

Where,

A=Final Hidden Layer Neuron Unit

B=Output Layer Neuron Unit

$s$  = Weight of Connection between Final Hidden Layer Neuron and Output Layer Neuron

$t$ = Activation Value of Final Hidden Layer Neuron Unit

$\therefore$  Justification for the Output Layer Neuron = Weight  $\cdot$  Activation Value of Hidden Layer Neuron Unit

However, since the output layer neuron receives weights and activation values from all peripheral neurons in the previous hidden layer, the total justification for the output layer neuron would be a net sum of the individual weights multiplied by the activation value of the individual hidden layer neurons.

$$\therefore \text{Total Justification for Output Layer Neuron} = \sum (\text{Weight}_n \cdot \text{Activation Values}_n) + \text{Bias}_n$$

Where  $n=1, 2, 3, \Rightarrow \infty$  representing each Final Hidden Layer Neuron Unit

Establishing this justification for an output layer neuron provides a value that can convey the confidence of each output neuron. Such values can then be compared with each other to establish

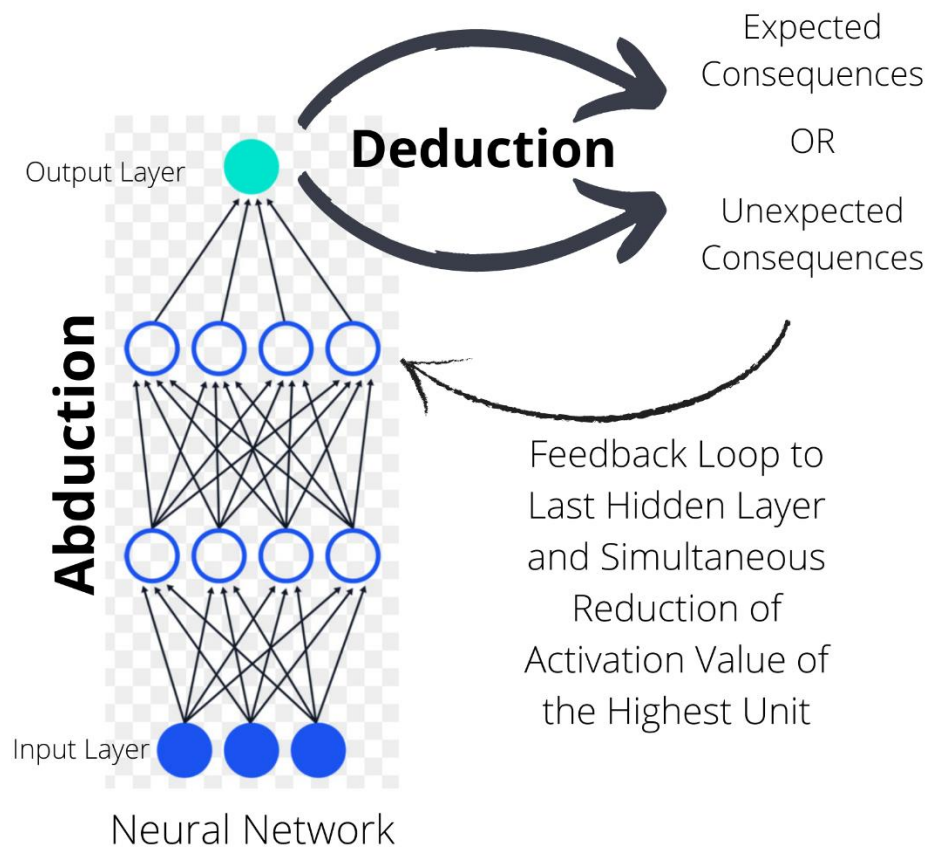
statistically different conclusions at which the AI arrives. For instance, if the difference between the total justification values of two distinct output layers is not statistically significant, the AI system does not “have enough confidence” to accurately deliberate between two almost equally-plausible conclusions. In such a scenario, it is best to have a human physician intervene and consider the case independently. When AI researchers conduct research on improving neural network accuracy, this certainty value can, in addition, do two other things: 1) provide an accurate gauge of the certainty of the AI system’s claims, thus drastically alleviating the black-box problem that makes AI research difficult; 2) allow scientists to introduce varied data sets designed to maximize the epistemological justification value of the AI system during the training phase. Overall, this could significantly improve the effectiveness, accuracy, and versatility of Deep Learning AI systems in the medical setting.

Coming back to our earlier discussion, current Deep Learning AI technology utilizes the ascending portion of Magnani’s ST-Model, but it fails to incorporate its descending portion. This can be attributed to a lack of functionality of Deep Learning AI systems. In addition, physicians have the capacity to deduce from their abductions conclusions tied to medical testing for long-term care and prognosis. As Deep Learning AI systems are still in an early stage of development in the medical setting, they currently only have the functionality to draw initial conclusions based on medical images, patient data, or medical facts. Incorporating an additional deductive structure that proceeds the pseudo-abduction phase, as described above, may not only help to increase the accuracy of diagnosis, but also give AI systems the ability to perform long-term prognosis. As this process is primarily deductive in nature (even in physicians), it should be easy to incorporate into the Deep Learning architecture (possibly after the pseudo-abduction neural network). Below is a proposed mechanism for incorporating this cyclical reasoning.



Concerning this last point, in current Deep Learning AI systems the output neuron with the highest activation value is considered the most plausible explanation for the patient's symptoms. This plausible explanation is tied deductively to expected medical test results and lab values. If the AI system came to the wrong conclusion, the data from the deductive testing in the descending phase of the ST-Model will not match the conclusion of the AI system. Now, a feedback loop can be inserted from the deductive data results directly to the most-plausible output neuron generated by the AI system. As this conclusion is (computationally) the output neuron with the highest activation value, the feedback loop can be designed to significantly reduce the activation value of that output neuron if the conclusion does not correlate to the deductive test results. In this case, the second-best explanation (the output neuron with the second highest activation value) will become the highest activation value explanation. At this point, the descending portion of the ST-Model can be tested for its coherence with the now-highest activation value conclusion. This cycle can be repeated until the testing conclusions match the expected results of the AI diagnosis with the highest activation value.

Such a mechanism would be useful after the training phase of the AI system and will ensure that it continually re-evaluates its conclusions, leading to a better justified, accurate diagnostic conclusion. In conclusion, through these kinds of mechanism Deep Learning AI systems can incorporate non-monotonic logic, giving them additional functionality in the medical setting. (For a visual representation of the whole process, see the diagram below.)



## REFERENCES

- Bilbrey, J. A., Ramirez, E. F., Brandi-Lozano, J., Sivaraman, C., Short, J., Lewis, I. D., Barnes, B. D., & Zirkle, L. G. (2020). Improving radiograph analysis throughput through transfer learning and object detection. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://doi.org/10.21037/jmai-20-2>
- Douven, I. (2021). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2021/entries/abduction/>
- Elkins, A., Freitas, F. F., & Sanz, V. (2020). Developing an app to interpret chest X-rays to support the diagnosis of respiratory pathology with artificial intelligence. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://doi.org/10.21037/jmai.2019.12.01>
- Goodman M.D. FACR, L. (Ed.). (2021). *Felson's Principles of Chest Roentgenology, A Programmed Text—5th Edition* (5th ed.). Elsevier. <https://www.elsevier.com/books/felsons-principles-of-chest-roentgenology-a-programmed-text/goodman/978-0-323-62567-8>
- Hanson, N. (1965). *Patterns of Discovery*. Cambridge University Press.
- Harmon, S. A., Sanford, T. H., Xu, S., Turkbey, E. B., Roth, H., Xu, Z., Yang, D., Myronenko, A., Anderson, V., Amalou, A., Blain, M., Kassin, M., Long, D., Varble, N., Walker, S. M., Bagci, U., Ierardi, A. M., Stellato, E., Plensich, G. G., ... Turkbey, B. (2020). Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nature Communications*, 11(1), 4080. <https://doi.org/10.1038/s41467-020-17971-2>

- Howard, J. P., Tan, J., Shun-Shin, M. J., Mahdi, D., Nowbar, A. N., Arnold, A. D., Ahmad, Y., McCartney, P., Zolgharni, M., Linton, N. W. F., Sutaria, N., Rana, B., Mayet, J., Rueckert, D., Cole, G. D., & Francis, D. P. (2020). Improving ultrasound video classification: An evaluation of novel deep learning methods in echocardiography. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://jmai.amegroups.com/article/view/5205>
- Jorstad, K. T. (2020). Intersection of artificial intelligence and medicine: Tort liability in the technological age. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://doi.org/10.21037/jmai-20-57>
- Martin-Carreras, T. T., Li, H., & Chen, P.-H. (2020). Interpretative applications of artificial intelligence in musculoskeletal imaging: Concepts, current practice, and future directions. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://doi.org/10.21037/jmai-20-30>
- Mitchell, Melanie (2020): *Artificial Intelligence: A Guide for Thinking Humans*, Picador  
*Neural Networks Bias And Weights. Understanding The Two Most Important... | by Farhad Malik | FinTechExplained | Medium.* (n.d.). Retrieved December 27, 2021, from <https://medium.com/fintechexplained/neural-networks-bias-and-weights-10b53e6285da>
- Pelaccia, T., Forestier, G., & Wemmert, C. (2019). Deconstructing the diagnostic reasoning of human versus artificial intelligence. *CMAJ*, 191(48), E1332–E1335. <https://doi.org/10.1503/cmaj.190506>
- Riihimaa, P. (2020). Impact of machine learning and feature selection on type 2 diabetes risk prediction. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://doi.org/10.21037/jmai-20-4>

- Sathyakumar, K., Munoz, M., Bansod, S., Singh, J., & Babu, B. A. (2020). Physician-assist automated AI lung cancer detection: A narrative review. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://doi.org/10.21037/jmai-20-24>
- Schmidt, W., Regan, M., Fahey, M., & Paplinski, A. (2019). General movement assessment by machine learning: Why is it so difficult? *Journal of Medical Artificial Intelligence*, 2(0), Article 0. <https://doi.org/10.21037/jmai.2019.06.02>
- Sooknanan, J., & Seemungal, T. (2019). Not so elementary – the reasoning behind a medical diagnosis. *MedEdPublish*, 8. <https://doi.org/10.15694/mep.2019.000234.1>
- What are Neural Networks?* (2021, August 3). <https://www.ibm.com/cloud/learn/neural-networks>
- Wilder, B. (2020). Artificial intelligence in medicine: Is the genie out of the bottle? *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://doi.org/10.21037/jmai-20-36>
- Wu, J. T., Wong, K. C. L., Gur, Y., Ansari, N., Karargyris, A., Sharma, A., Morris, M., Saboury, B., Ahmad, H., Boyko, O., Syed, A., Jadhav, A., Wang, H., Pillai, A., Kashyap, S., Moradi, M., & Syeda-Mahmood, T. (2020). Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Network Open*, 3(10), e2022779. <https://doi.org/10.1001/jamanetworkopen.2020.22779>
- Xu, Q., Wang, L., & Sansgiry, S. S. (2020). A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning. *Journal of Medical Artificial Intelligence*, 3(0), Article 0. <https://jmai.amegroups.com/article/view/5197>