6-2015

What It Is To Be Conscious: Exploring the Plasibility of Consciousness in Deep Learning Computers

Peter Davis
Union College - Schenectady, NY

What It Is To Be Conscious: Exploring the Plausibility of Consciousness in Deep Learning Computers

By

(Peter) Zachary Davis

****

Submitted in partial fulfillment
of the requirements for
Honors in the Departments of Philosophy and Computer Science

UNION COLLEGE
June, 2015

ABSTRACT

DAVIS, PETER Z. What It Is To Be Conscious: Exploring the Plausibility of Consciousness in Deep Learning Computers

ADVISORS: Kristina Striegnitz and David Barnett

As artificial intelligence and robotics progress further and faster every day, designing and building a conscious computer appears to be on the horizon. Recent technological advances have allowed engineers and computer scientists to create robots and computer programs that were previously impossible. The development of these highly sophisticated robots and AI programs has thus prompted the age-old question: can a computer be conscious? The answer relies on addressing two key sub-problems. The first is the nature of consciousness: what constitutes a system as conscious, or what properties does consciousness have? Secondly, does the physical make-up of the robot or computer matter? Is there a particular composition of the robot or computer that is necessary for consciousness, or is consciousness unaffected by differences in physical properties? My aim is to explore these issues with respect to deep-learning computer programs. These programs use artificial neural networks and learning algorithms to create highly sophisticated, seemingly intelligent computers that are comparable to, yet fundamentally different from, a human brain. Additionally, I will discuss the required actions we must take in order to come to a consensus on the consciousness of deep learning computers.

Contents

List of Figures

Chapter 1: Introduction

It seems that, from birth, you have come across an immense number of objects and people. In your earliest years, you have almost always had the luxury of being surrounded by others to guide and assist you. Though it is unclear exactly *how* you were guided in the development of concepts, you were most likely always exposed to help in categorizing and recognizing the various objects you came into contact with. Your parents or guardians may have shown you a ball while telling you the object was a ball. Over time, you developed an idea of what a 'ball' is. Using these newly developed rules about the world, you gained the ability to categorize objects by type, such as 'ball.' Regardless of exactly what role other people had in helping you learn concepts and distinguish between objects, it is true that you received help.

Now imagine a situation in which you are born alone and do not have access to such assistance. Think of yourself as a genius newborn, in the sense that you can reason well, but you know nothing. While you can think rationally, there is nobody to help you come to conclusions or correct you when your conclusions are wrong. It is up to you alone to decide what a 'ball' is, and which objects are balls. Surely, this task seems daunting and difficult.

In essence, this is what unsupervised, deep learning computers computer program that is capable of identifying aspects of an image, such as the face of a cat, and categorizing and grouping images of cats together (Le, et al., 2013). It is quite impressive that a computer is capable of accurately organizing images, but it is even more impressive

that it is able to organize the images without labels on the images, human interference or guidance, or even prior knowledge of cats.

Any system, be it a human or a computer, that can consistently complete a task that requires making judgments and applying rules seems to be, at some level, intelligent (though this may be a controversial claim). I do not want to focus on the debate over intelligence here, as I believe a far more controversial claim is that such a computer is conscious. After all, it seems like the computer *could* be conscious. It does what many conscious things do. The computer may even do things in a similar way; that is, it may learn from examples and essentially teach itself what a cat is (or at least what an image of a cat is). And maybe there is something about the computer that can explain how it can do all of that, and maybe that something is consciousness. But perhaps the computer is not conscious. Perhaps there is a perfectly acceptable and sufficient explanation for the computer's behavior that does not rely on the computer being conscious.

These ideas and questions are what I am concerned with. More specifically, I want to explore how three different theories of consciousness, functionalism (Levin, 2013) (including Daniel Dennett's Multiple Drafts Model (Dennett, 1991)), Ned Block's physicalist "hybrid" theory of consciousness (Block, 2002), and Giulio Tononi's modified functionalist Integrated Information Theory (Tononi, 2004), relate to deep learning, and how deep learning relates to those theories. Depending on which of these three theories of consciousness is accepted, we come to different conclusions, based on varying forms of support, about the consciousness of a deep learning computer. Yet, interestingly, there appears to be a weak, but definite, form of agreement between the theories on the topic of computer consciousness.

The question of computer consciousness also has varying implications on both philosophy and computer science. If a computer is considered conscious, then there are ethical concerns, such as turning off computers, that need to be deliberated on and addressed. Conscious computers would also change the study of the philosophy of mind, as we would need to include computers in the discussion of the mind in a way that regards them as conscious, not just a candidate for consciousness, so to speak. We would need to view consciousness in a way that includes both animals *and* computers, as well as other systems, not just animals. The field of computer science would also be affected significantly. What would it mean to build, program, and turn on a computer? Is it at all like having a child? Should we change our programming standards and practices in light of the computer being conscious? How will a conscious computer affect artificial intelligence development? While the focus of this paper will not be on these specific questions, but rather the plausibility of a conscious deep learning computer, their answers and considerations could surely be influenced by the conclusions I will make.

In the next section, I will examine deep learning computer algorithms, as well as machine learning, in general. Then, I will describe the three theories of consciousness mentioned above, as well as the potential validity of each theory, though I will not make any definite conclusions in that section. In the third chapter, I will synthesize the two preceding sections and discuss deep learning programs' ability to satisfy the conditions of consciousness, as outlined by the three theories. This will include any judgments on whether deep learning computers can be viewed as conscious beings. Next, I will illustrate the advantages that Integrated Information Theory has over the other two theories, ultimately suggesting that IIT should be accepted instead of the others. This will not be a

definitive defense of IIT, as IIT still has weaknesses, but minimally, I hope to show that the views of Block and Dennett are not attractive options for a theory of consciousness. I will also briefly investigate the "science of consciousness" and methods we may be able to use in order to advance the field of consciousness.

Chapter 2: Machine Learning and Deep Learning Algorithms

As the result of recent advancements in technology, the popularity of machine learning algorithms and techniques in computer science has increased. Scientists' ability to simulate and represent difficult logical operations, such as "exclusive or," has made machine learning applicable and fit for use in more interesting and challenging problems. However, in the midst of all the excitement, the true definition of *machine learning*, and a true understanding of what machine learning systems are used for, is occasionally overlooked, omitted from conversation, or otherwise skewed.

Machine Learning

When one hears the term 'machine learning' for the first time, it is most likely accompanied by ideas of futuristic, intelligent robots. And that response is normal. After all, the phrase does seem to imply that machines are *learning*, as humans do, and are therefore, advanced, thinking computers. And while that would be technologically impressive and interesting, that is not the reality of machine learning. Instead, machine learning would be best defined as an interdisciplinary field, involving computer science and mathematics, in which programs and algorithms are used to complete a specific task by way of "learning." Specifically, a computer that "learns" is meant to improve its ability to complete some task. One formal definition of machine learning that is generally accepted states that a "program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E* (Mitchell, 1997)." For example, imagine a particular

program was designed to play chess. The task *T* is playing chess, and the performance *P* is the program's ability to win games of chess, or the percentage of games won. If the program's win percentage increases as the result of playing games of chess (the computer's experience *E*), then the program can be said to be a learning program.

The "learning" feature of these sorts of computer systems is quite useful, and in some cases necessary, in the completion of many different problems. However, machine learning algorithms are not appropriate for any assignment. Normally, machine learning algorithms are only practical when the task to be completed is one that can be done by applying generalities derived from examples (Domingos, 2012). A computer's ability to generalize from specific examples is extremely beneficial. Just as humans are able to identify certain objects as chairs by viewing numerous instances of particular chairs and developing the general form of a chair, machine learning systems are capable of forming sets of general rules and analyzing information and data with those rules. Therefore, machine learning algorithms are widely applied to the fields of data mining and data analysis, where there is far too much information for humans to process manually.

There are two common types of machine learning algorithms: *supervised* learning and *unsupervised* learning. The basic difference between these two kinds of machine learning is the computer's access to the correct "answer." For example, if an algorithm is being applied to data about irises, and the task is to correctly classify each instance as one of three types of irises, then the type of learning would be supervised, since the data being used by the computer includes the class type for each instance. In essence, this allows the computer to "check" its answers as it works, thus allowing it to learn which attributes are important and which rules are good rules to follow. More generally, the purpose of

supervised learning is to identify and "learn the relationship between the input *x* and output *y* (Barber, 2012)." This makes supervised learning well-suited for problems of classification and association between features and classes. As a result, supervised learning is widely used for questions of prediction, or determining which class an input instance *x* will belong to, based on which classes similar instances belong to (Barber, 2012). The other form of machine learning is unsupervised learning. In cases of unsupervised learning, the input data is not accompanied by the known correct output. Instead, the computer is tasked with finding its own description of the input data (Barber, 2012). This means that the computer will not truly know what it is looking at, or looking for, but will instead identify patterns within the data so that it knows which instances are similar to other instances.

For the purpose of this paper, I will not explicitly focus on one type of learning over the other, as the supervision feature is contingent on the data being used. But it should be noted that unsupervised learning algorithms are more interesting, in the sense that they are capable of finding high-level conceptual patterns in data. The fact that task-specific information is not necessary means unsupervised learning is applicable to larger, more diverse types of data. Moreover, unsupervised learning algorithms parallel human and animal learning more closely. Humans and animals are still given guidance and feedback when learning concepts, though that is not always the case, especially in older organisms. That is, it is less common for an adult, rather than a child, to be shown and taught what a chair is through examples and training.

Computer learning has played a significant role in recent advancements in the fields of robots and artificial intelligence. The auto-learning nature of machine learning

programs has relieved artificial intelligence researchers of much of the burden of determining exactly how the human brain functions. Instead, they have implemented relatively simple machine learning algorithms[1] that complete a task without having to be told how to do it. And just like everything that initially seems simple and straightforward, there is much more to machine learning in AI than merely allowing the computer to learn what to do. While machine learning researchers do not have to know every detail and aspect of how the task should be completed, they do still have to identify what the important elements, or features, of the problem are. Different approaches, algorithms, and architectures have been created and used to tackle the various problems of artificial intelligence. One particular method of implementing a machine learning algorithm is the use of *artificial neural networks*.

Artificial Neural Networks

Artificial neural networks have greatly advanced the field of artificial intelligence. Neural networks have allowed researchers to develop computers and machines that can succeed in completing impressive tasks, such as recognizing handwriting, speech, and faces. These problems require processing large amounts of complicated data, which neural networks are well-suited for. Thus, neural networks are considered one of the more effective machine learning methods, and one of the most commonly used, as well (Russell & Norvig, 2009) (Mitchell, 1997). Due to their popularity and success, and the

[1] These algorithms are not necessarily objectively simple, but they are simple in comparison to the complexity of an algorithm that explicitly explained how to complete a task.

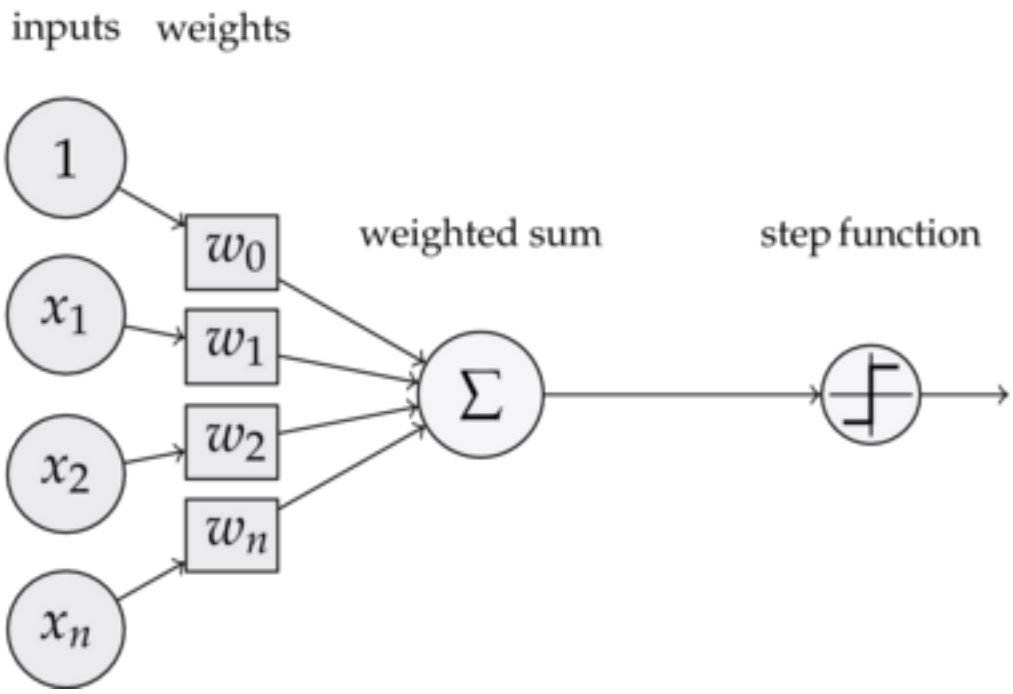
connections neural networks have to biology and deep learning, I will focus on them as the primary method of machine learning.

Basically, artificial neural networks are sets of individual units that are each capable of taking some input, processing it, and producing some new output (Mitchell, 1997) (Russell & Norvig, 2009). These separate units are analogous to the independent neurons of the biological animal brain. In fact, the entire concept of artificial neural networks was motivated by the structure and basic functionality of the human brain (Mitchell, 1997). The brain consists of an incredible number of special cells called *neurons*. These discrete neurons are interconnected to other neurons, creating a vast complex we know as 'the brain.' Yet, despite the overall intricacy of the brain, it is able to complete difficult tasks, while handling large amounts of data, at quick speeds. For instance, an average person can recognize another person on the order of 10^{-1} seconds (Mitchell, 1997); this is too fast for us to notice at all. This speed is even more surprising and amazing when we account for the fact that each neuron has a switching time, or the time it takes for it to be agitated and complete a process, of roughly 10^{-3} seconds (Mitchell, 1997). Therefore, it seems as though the actual process of analyzing all of the visual data pertaining to somebody's face and producing a judgment about it, i.e. you are looking at your mother, cannot be much more than a few hundred steps (Mitchell, 1997). Biologists and neuroscientists have taken this information to be evidence that general brain function is comprised of many parallel processes. A simple representation is that the information sent to the brain is distributed to various neurons and all of the individual neurons' outputs are collected and accumulated to form one coherent output. The creators of artificial neural networks have sought to mirror that theory. The basic principle of

allocating small pieces of the large set of data to individual processing units has been retained.

Perhaps the best way to begin understanding the basics of how artificial neural networks function is to look at the simplest version of a neural network: a single perceptron. A *perceptron* is an extremely basic neural network which consists of a single artificial neuron. This "neuron" is just a mathematical function designed to yield output based on input (Russell & Norvig, 2009). As is shown in Figure 1, the neuron takes in some set of inputs $i_1, i_2,...,i_n$, each carrying some weight, which can vary depending on the source of the input.

Figure 1: A representation of a single perceptron neuron.

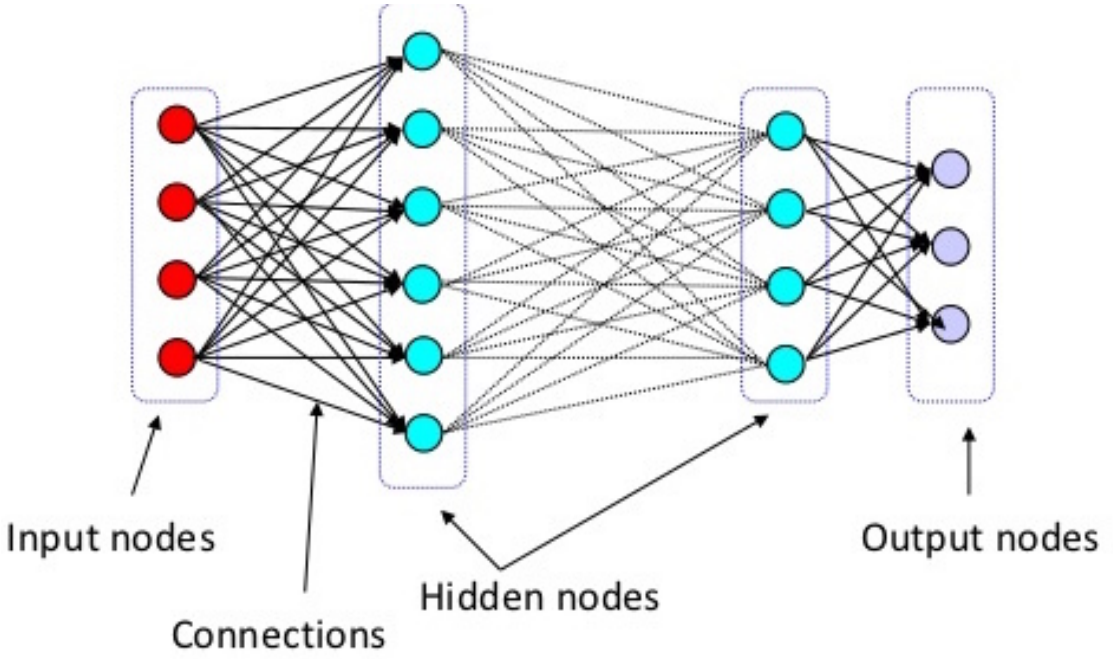


The weights of all the inputs are then summed. The neuron's processing function is activated, or triggered, when the inputs' total weight exceeds a set threshold, which, like the weights, can vary among neurons. If the threshold is exceeded, then the neuron handles the inputs' weight sum and emits an output (Russell & Norvig, 2009). The

perceptron model is the general model used for neurons in typical neural networks today. Modern neural networks connect sets of these neurons. However, there are different ways in which these neurons can be linked together. Two popular methods are *feed-forward* networks and *recurrent* networks. While both of these architectures are similar, in that they consist of layers of artificial neurons, there exist a few fundamental differences between them.

Feed-forward neural networks are comprised of links between neurons that all point in the same direction. This means that the initial raw input enters the preliminary layer of neurons, is handled in some way, and the output from those neurons is sent through to the next stage. In other words, information is always *fed*, or passed, in the same direction (Russell & Norvig, 2009). As Figure 2 illustrates, the links between groups of neurons never go backward, so a neuron does not encounter the same data more than once.

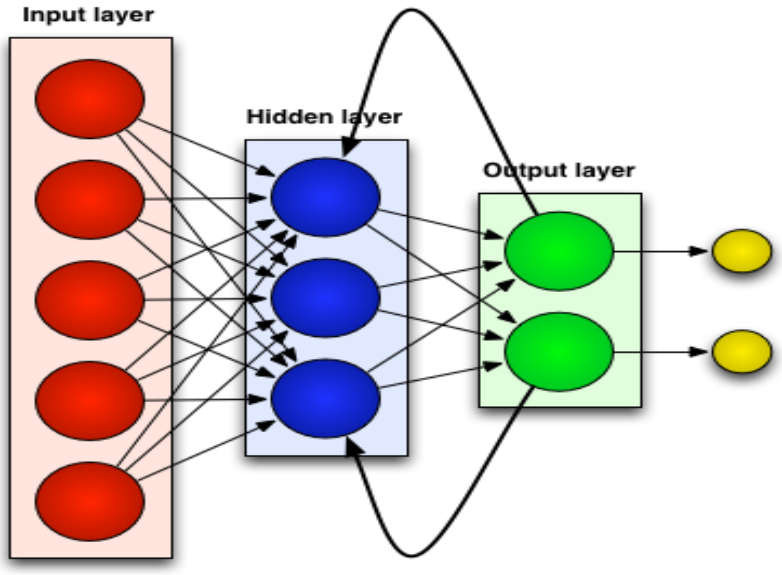
Figure 2: A common, basic feed-forward network. Notice that the links between the neural layers all point in the same direction. The neurons' output in each layer, except for the final layer, acts as input for the neurons in the next layer.



One could think of a feed-forward network as an assembly line in a car factory. The beginning of the line receives some material that does not resemble a car, but as it moves along the conveyor belt, parts are added and assembled until, at the end of the line, it is a complete car. During this process, however, the car can only move in one direction; since it cannot go back to a previous position, it is important that the individual workers or machines complete their task adequately, timely, and correctly. Failure to do so would affect the entire process. Similarly, each neuron in a feed-forward network needs to function correctly. Errors made at early stages will adversely affect the output of the network. In the case of matching information for the purposes of recognition, early errors will be compounded on as the information continues through the network, eventually resulting in mismatching and inaccuracies in recognition (Young, Scott, & Nasrabadi, 1997). Data *could* theoretically be processed twice by a single neuron, just as the seats could be reinstalled in a car if they were not finished the first time, but the information would have to complete a traversal of the entire network first, and then be fed into the beginning again, thus making it inefficient and costly.

On the other hand, and as the name suggests, recurrent neural networks link neurons in more than one direction. The network in Figure 3 is an example of a simple recurrent neural network with bidirectional neural links.

Figure 3: A simple recurrent neural network. In this network, there is a backward link between the final output layer and the hidden layer, allowing data to be fed back into the hidden layer, as well as out of the entire network.



The result of this design is that outputs are fed back into the network as inputs, creating a loop (Russell & Norvig, 2009). This means that unlike feed-forward networks, where the only internal state is that of the inputs' weights, the activation threshold of a neuron in a recurrent neural network is potentially dependent on previous inputs. Thus, recurrent network neurons are capable of, and in fact, require, support for short-term memory (Russell & Norvig, 2009). For example, imagine a single neuron responsible for simply calculating the sum of the weights of three inputs, i_1, i_2, i_3, where the neuron receives the inputs one at a time. If the neuron was in a feed-forward network, then it would receive input i_1, add its weight, e.g. 1, to the total (presumably 0), and output that total: 1. When inputs i_2 and i_3, with weights 2 and 3, respectively, are received, their weights are added to the total and outputted. However, the total would be "reset" each time and the overall output would be 3. However, a neuron in a recurrent network would be able to hold on to the necessary total weight. This means that the weights of inputs i_2 and i_3 would be added to totals 1 and 3, respectively, and the final output of the neuron would be 6. This

characteristic of recurrent networks, that is, the associative memory capabilities of the neurons, make recurrent networks more intriguing with respect to studying models of the brain. I will discuss recurrent networks further, later in this chapter.

There are more variations that can be made to neural networks than simply the direction the information flows. Take, for example, feed-forward networks[2]. Commonly, feed-forward neural networks are composed of layers of neurons, where a layer is simply a collection of neurons (a network) that represent a level of abstraction, or a step in the overall process. These layers are arranged in a manner such that the input into a neuron, or layer of neurons, comes only from the neuron, or layer of neurons, immediately preceding it in the direction of the data flow (Russell & Norvig, 2009).

Just as it is helpful to begin understanding the basics of neural networks by examining the single perceptron, grasping the concepts of feed-forward network layers starts with the simplest form: single layer networks. Until this point, I have been describing feed-forward networks primarily as those with multiple layers, or stages. That is, information is passed from one neuron to another. But that is not the only type of feed-forward network that exists. A simpler, earlier version was the single-layer network. These types of networks are immensely similar to the perceptron discussed above; in fact, the only genuine difference is that the system is made up of a number of artificial neurons, rather than just a single one. But a single-layer feed-forward network still produces output directly from the input, just as the single perceptron network does (Russell & Norvig,

[2] Although I mentioned that recurrent networks are generally more interesting in terms of brain modeling and studying intelligence, I am giving more attention to feed-forward networks due to their wide-spread use in computer learning and their relatively simpler structure. Many of the basic concepts that apply to feed-forward networks can be easily extrapolated to recurrent networks.

2009). In other words, input is fed into the network and output is yielded after just one iteration of neuronic activity. Intuitively, single-layer feed-forward networks do not seem entirely useful. As illustrated by the factorial problem above, a network that is only able to process data once, even if it can be distributed amongst a large number of neurons, will have limited applicability. In terms of Boolean functions, single-layer feed-forward networks, as well as single perceptron networks, are only able to complete conjunctive, disjunctive, and negative functions (Mitchell, 1997) (Russell & Norvig, 2009). In order for any one of these problems to be done, the data only needs to be examined once, and the correct answer can be given. More complex functions, such as exclusive disjunctive functions, which require at least two stages of basic Boolean functions, cannot be completed with just one iteration, regardless of the number of neurons. Imagine the car example again. If, instead of an assembly line, every phase of car assembly were done concurrently, the car would not be completed. Rather, you would have a number of partially assembled car pieces. Similarly, complex tasks cannot be done in single-layer feed-forward networks; the output would be incomplete. Luckily, the solution to this problem is fairly straightforward, at least conceptually.

By creating a neural network with more than one layer of neurons, more Boolean functions can be managed, even exclusive disjunction. This has been known, in the theoretically sense, for decades (Russell & Norvig, 2009). However, putting such a principle into practice proved to be a major difficulty in the field of machine learning. But eventually, researchers devised a way of creating appropriate feed-forward neural networks that passed output from one neuron as input to another neuron. The key component to multi-layer feed-forward networks are the hidden layers of neurons,

specifically, the way in which neurons link to the next layer. Hidden neuron layers are simply layers whose output is only available to other neurons and is not given as output of the network as a whole (Mitchell, 1997). The layers in these networks are typically structured in a manner such that the information direction is acyclic (Mitchell, 1997). The neurons of a particular hidden layer feed their respective outputs as inputs to the next layer of neurons. The aspect of these neural networks that is considered "learning" is the process of deciding how much weight, that is, how much importance each input has on the final output, to assign to each input and the links between neurons (Mitchell, 1997).

Before delving into general neural network learning algorithms and weighting rules, allow me to briefly return to discussing recurrent networks. Recall that recurrent neural networks share many of the same fundamental properties as feed-forward networks. They still consist of neurons with activation thresholds, and take in input and produce output. However, the output they produce does not necessarily have to go on to the next layer of neurons. Instead, a neuron's output can feed into a neuron that is "behind" it, or even back into the neuron that it came from. The Hopfield network is a well-understood version of a recurrent network, though it is not the only type that is used. In these sorts of systems, the direction in which information moves is both forwards and backwards, and the weights of the inputs are equal for each neuron (as I will explain below, this differs from feed-forward networks). Also, recall the neurons in recurrent networks support memory. That is, the current state of the neuron is affected by previous inputs and affects future outputs. When some input *i* enters a neuron, *i*'s weight is added to the activation threshold total. Regardless of whether or not the activation threshold was met, that total is "remembered" and the next input is processed in the context of the

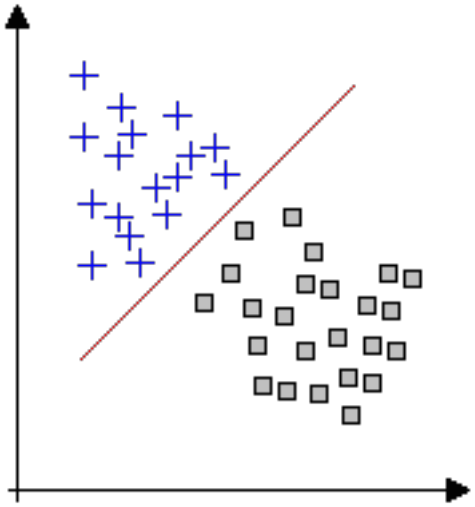
neuron's new state (with the activation threshold total at 1 rather than 0, for example) (Russell & Norvig, 2009). In terms of the functionality of Hopfield networks, this memory allows for the network to respond to new stimuli that resemble the stimuli used to train the network. For example, if photographs were used to train the network, and the new input was a section of one of the training photographs, then the neurons in the network would be activated in a sequence, or pattern, such that the original training photograph is recreated. Since different "parts" of the photograph are stored in each neuron, the correct "traversal" of the network will result in collecting all of the pieces of the photograph (Russell & Norvig, 2009). Additionally, multi-layer Hopfield networks have been used for tasks such as object recognition (Young, Scott, & Nasrabadi, 1997). Young et al. created a multi-layer Hopfield network, where each layer of the network was a single layer Hopfield network responsible for a particular task. The overall purpose of the network was to match pictures of objects with known training objects. The researchers used a method by which separate layers processed information concurrently and at different resolutions (either coarse or fine visual elements were analyzed). This method allowed the computer to recognize images in a manner that a single-layer network, or even multi-layer feed-forward network, would be incapable of with the same level of performance.

Hopefully, I have explained the basic and common structures of neural networks sufficiently enough to provide more details about *how* neural networks work. I briefly mentioned the idea of weights and thresholds above, but I did not describe where those values come from, or how they affect the system. To do so, I will return to the single perceptron network, since the algorithms used for perceptrons are pertinent to neural networks with many neurons and multiple layers (Mitchell, 1997). Single perceptron

networks are simple to understand, and the problems they are used for are normally not too complex. Therefore, determining the correct weights to assign to the inputs, such that the neuron functions correctly and the output is error-free, is a relatively simple problem to solve.

The most basic way to find satisfactory weights is to use a pseudo-guess-and-check method, i.e. choosing a random weight and modifying it when the perceptron makes a mistake on the training data. The weight is adjusted according to the *perceptron training rule*. Under this rule, the weight is changed in relation to the input, can be done using the following formula: $\Delta w = n(t - o)x$, where w is the weight, n is a constant (normally small), t is the target output, o is the actual output, and x is the input (Mitchell, 1997). While the perceptron training rule is generally successful in finding weights, it has its disadvantages. The primary defect of the perceptron training rule is that it struggles to find an adequate weight if the training data is not linearly separable (Mitchell, 1997). Linear separability simply refers to the ability to split the data, according to the output, on a graph. For example, imagine a soccer game with two outcomes, 'win' and 'loss.' And say Figure 4 is a plot of the results, with 'wins' being squares and 'losses' being crosses, with respect to goals scored on the x-axis and goals allowed on the y-axis.

Figure 4: An example of linearly separable data. The two outcomes can be divided such that all outcomes on either side of the line are alike.



The results are said to be linearly separable, since the results can be cleanly separated. In other words, data is not linearly separable just in case that there is no clear and ultimate dividing line between the data; when this is the case, there does not exists a line that can be drawn on the graph, such that there are not some number of all possible outputs on either side of the line. In order to solve this puzzle, another method, called the *delta rule* can be applied. The true motivation and underlying concept of the delta rule is *gradient descent*. And since the delta rule is, more or less, an application of gradient descent, and gradient descent is more widely used in neural network learning, particularly in the popular back-propagation algorithm, gradient descent is the true focus (Mitchell, 1997).

Before discussing gradient descent, it would be helpful to explain the process of back-propagation. The crucial aspect of the back-propagation algorithm is its ability to "go back into the network" and propagate the observed, total error over the neurons in the multiple layers. After the network is created, the training data is iterated over the network many times, using small, random weights, initially (Mitchell, 1997). After each iteration of the network, the entire network's output error, *E*, is calculated. Then, using a portion of

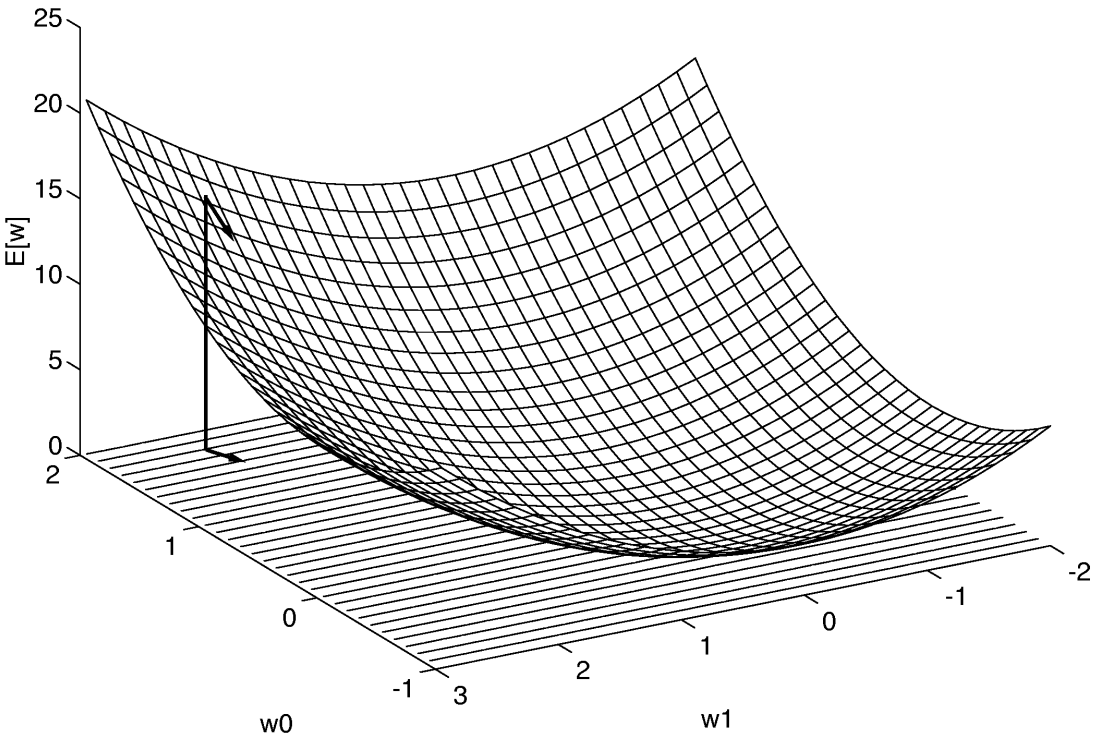
that error *E*, the error, *e*, of a specific neuron is calculated, and the weights are adjusted appropriately, starting from the output layer and working backwards to the first input layer (Russell & Norvig, 2009). More specifically, the algorithm finds the error for the whole network and derives the error for an individual (hidden) neuron, *h*, by summing the output error for every output influenced by neuron *h*. The weight is altered such that it reflects *h*'s responsibility for the network output error (Mitchell, 1997). Generally, this leads to small incremental weight changes, and so the training data may have to iterate over the network thousands of times (Mitchell, 1997). Despite this potential disadvantage, it is still an effective way for the neural networks to learn accurate weighting rules, for both feed-forward and recurrent neural networks.

In the case of recurrent networks being trained with back-propagation, it is easiest to think of each iteration through the network as being a separate copy of the network, where the output is fed into another copy rather than back into itself. By visualizing recurrent networks in this manner, they begin to take the shape of a feed-forward network. The back-propagation algorithm is then applicable to the recurrent network in the same way it is in the feed-forward network. The final weight adjustment of the entire network can be thought of as the mean weight adjustments from the copies of the network (Mitchell, 1997). Back-propagation is a bit more difficult to implement on recurrent networks than it is on feed-forward networks, but it is still possible, which is beneficial to the field of artificial intelligence.

Plainly put, the basis of gradient descent is error minimization. The use of gradient descent allows the algorithm to find input weights the lead to the minimum error in a set of possible outcomes. In order to do so, the algorithm takes the error in the output and

calculates small adjustments in the input weights until the error is reduced. Figure 5 shows a visualization of the "hypothesis space" for a particular problem. The two axes, w_0 and w_1 illustrate possible weights for some input, while the vertical axis, E, represents the training error (the error experienced in the output from the training data).

Figure 5: A visual representation of the hypothesis space for an arbitrary problem.



Notice the shape of the error surface, or the plotted error values experienced when certain weights, w_0 and w_1 are used. This visually shows the effects various weights have on the accuracy of the output. The key feature, though, is the minimum error, which will be achieved for some input weights. The gradient descent algorithm aims to find that global minimum of E by modifying the weight value on the appropriate axis such that it descends along the error surface until the minimum error is reached (Mitchell, 1997).

While the gradient descent algorithm is important in single perceptron networks, it is also useful as a basis for learning algorithms designed for large, multi-layer neural

networks. Typically, multi-layer networks, and networks with many neurons, use gradient descent, via the *back-propagation* algorithm, to determine the weights of inputs (though other methods do exist). But there is an obstacle that is faced when finding the appropriate weights in networks with a large number of neurons, and the obstacle is that of size; the hypothesis space is vast in these types of networks. Recall the error surface and hypothesis space for the single perceptron above. That was representative of one neuron with two input weights. It should be clear how searching through a hypothesis space and error surface that depicts all potential weights for all of the neurons in a multi-layer network is can be problematic. Another issue gradient descent encounters on multi-layer networks is that of local minima (Mitchell, 1997). The error surfaces for these networks can contain multiple local minimums, since there are multiple outputs (one for each neuron or layer of neurons) and thus, the error, E, is the sum of all of the individual outputs (Mitchell, 1997). It is plausible, then, that the algorithm moves towards one of these local minimums rather than the global minimum, as there is no way to decipher which minimum is being descended upon.

An additional difficulty one faces with a multi-layer network is determining the error in the hidden layers. In a single perceptron network, or even a single-layer network of any type, calculating the error is easy, since the input is known and the direct output is known. The mistakes that are observed are clearly the fault of a particular neuron. In a multi-layer neural network, however, there are hidden layers of neurons. The outputs of these layers are unknown. Furthermore, the training data does not provide the information necessary to assess the accuracy of these output, even if they could be retrieved (Russell & Norvig, 2009). Imagine a feed-forward network with just two layers. Although we

know what the inputs are and what the final outputs are, there is no manner by which to explicitly observe the error in the output from the first layer, which is used as input to the second layer of neurons. This is because that information is rarely, if ever, included the data; all that is known is the final output, not any intermediate outputs. So in terms of adjusting weights, it is not evident whether the first layer made an error, which skewed the final output, or the neurons in the second layer received correct information, but made a mistake that led to the ultimate error, or even if there were inaccuracies in both layers of neurons. Depending on which of these scenarios is the case, weights will be have to be modified differently. Yet, despite the complications I have just mentioned, back-propagation, using gradient descent, appears to be practically useful and accurate in determining input weights.

Deep Learning

To put it simply, deep learning is a subfield of machine learning. The term "deep learning" is just used to describe algorithms of a certain structure. It is natural, then, for deep learning to share many of the same principles and characteristics of general machine learning. However, the difference between a deep learning system and a general machine learning system lies in the architectural differences between the two. And naturally, these architectural dissimilarities have an effect on the approaches and applications of deep learning machines.

In principle, deep learning is not much different from other types of data mining and machine learning techniques. The overall goal and purpose of a deep learning program is to complete a task and improve performance as the result of some experience.

Just as with data mining tasks, the algorithms analyze input and identify patterns and correlations within the raw data. Then, the algorithms apply those learned relationships to new, novel input and produce some output that it believes to be accurate and correct. As I stated above, these deep learning algorithms are structured differently than more tradition algorithms, and therefore, can be utilized in different ways. Of course, deep learning machines can still tackle the same problems as other machine learning algorithms, and with similar levels of success (Bengio, 2009). Yet, a unique aspect of their structure allows them to be successful in other areas, such as vision and language processing in AI.

Deep learning systems are essentially larger implementations of artificial neural networks. More specifically, deep learning systems differ from most other types of neural network-based machine learning systems in that the depth of these networks, or the number of layers and levels of single networks (comprised of individual neurons), is great, thus yielding the name *deep* learning (Bengio, 2009).

As I explained in the previous section, the development of artificial neural networks has been a major breakthrough in the field of artificial intelligence. Deep neural networks make completing particularly difficult and complex problems in AI possible. The networks break down huge puzzles and large amounts of information and handle them through multiple levels of abstraction (Bengio, 2009). When dealing with high-level abstractions, such as face and hand-writing recognition, these networks are crucially important. Humans do not yet have a great enough understanding of how the brain works to make it possible to explicitly tell the computer which precise features are significant in each task (Bengio, 2009). In fact, it is unknown to researchers exactly how, and what, the machines are learning, even though individual aspects of the algorithms are understood.

That said, machine learning researchers normally have an understanding that is substantial enough to be able to point the program in the right direction. Yet, some researchers do not even provide the computer with that information. For example, in the case of recognizing and classifying handwritten digits and faces, Claus Neubauer simply used raw images (in some cases, the data was normalized by size and shade) as input and allowed the algorithm and convolutional network to find the necessary patterns and determine which particular aspects of the image were important (Neubauer, 1998). Of course, this is perhaps an extreme case of machine learning, but it illustrates the power of these networks.

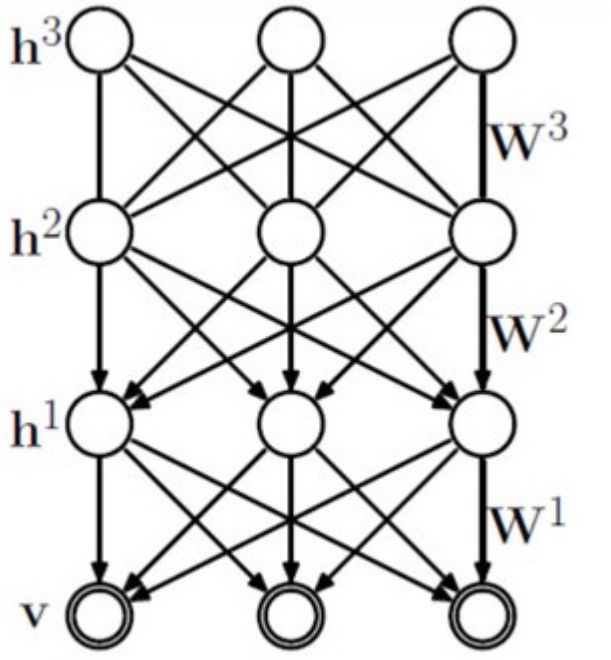
But why has deep learning become popular? What explicit advantages, other than what I have already alluded to, do researchers gain from using deep neural networks instead of the more traditional, shallow networks? In essence, shallow networks of only one or two neuron layers do not provide the modeling or representational power needed to solve complex, real-world problems. Deeper neural networks allow computers to successfully accomplish high-level tasks, such as linguistic and visual processing. Considering the biological inspiration for neural networks and the fact that deep learning is simply an application of neural networks, it is not surprising that the motivation for deep learning also came from biology and neuroscience. Many processes that human brains complete with relative ease, such as vision and speech, have been shown to be hierarchical in structure (Deng & Yu, 2014). Deep neural networks have been designed with the purpose of replicating this hierarchical structure and the behavior and capabilities of humans.

However, deep networks were not successful from their conception; the reproduction of human-like neural structures and abilities was not achieved by simply adding more hidden layers. As the networks grew larger, researchers were once again confronted by the issue of weighting. While the back-propagation was an acceptable and popular method for determining input weights in shallow networks, it proved to be inadequate for networks with many layers. Recall the issue of local minima in the shallow, two-layer network and the possibility of descending into a local minimum rather than the global minimum. As hidden layers are added to the network, the number of local minima increases, and the likelihood of becoming trapped in a local minimum that is a poor representation of the overall error also increases (Deng & Yu, 2014). This undesirable feature of using back-propagation and gradient descent on deep networks has had two effects. The first is a general avoidance of using deep neural networks for machine learning. But I will ignore this consequence, since my focus *is* on the use of deep neural networks. The second consequence of the weight learning problem has been positive, and has led to the development of other strategies and designs for learning. These new designs have been a recent breakthrough in artificial intelligence research, as they have allowed machines to accomplish the high-level tasks mentioned above. Two well-established and effective deep networks that have been used are Deep Belief Networks (DBNs), Convolutional Neural Networks (CNNs). These designs, as well as most others that have been successful for deep learning, have a shared driving principle: "guiding the training of intermediate levels of representation using unsupervised learning, which can be performed locally at each level (Bengio, 2009)." This means that the network is trained (the

appropriate weights are calculated and adjusted) one layer at a time, rather than all at once as is the case with back-propagation.

This type of weighting method was initially introduced in 2006 by Hinton, Osindero, and Teh (Hinton, Osindero, & Teh, 2006). The team developed a type of deep recurrent network called a *Deep Belief Network*, which solved the problems that back-propagation ran into with deep networks. As Figure 6 shows, DBNs are deep neural networks with layers of Restricted Boltzmann Machines (RBMs), which is another type of neural network

Figure 6: In this example of a simple DBN, layers 3 and 2 represent the RBM. The links between the layers are non-directed, as the neurons find suitable weights according to the weights in the other layer. The other layers 1 and v are trained based on the RBM.



This design choice means the problem of training a deep network is effectively reduced to the problem of training an RBM (Hinton, 2010). These Restricted Boltzmann Machines are capable of unsupervised learning (Bengio, 2009) by way of Gibbs sampling (Hinton, 2010). As a result, the correct output does not have to be included in the training data,

which increases the applicability of DBNs. However, RBMs are not always capable of flawlessly modeling the training data, that is, learning the ideal weights, so modifications must be made to the DBN to heighten the accuracy of the network. Hinton et al. employed a greedy algorithm[3] that trains the DBN one layer at a time (Hinton, 2010) (Bengio, 2009) (Arel, Rose, & Kanowski, 2010). This algorithm trains the "lower level" input-layer RBM in an unsupervised manner, which yields a set of weight values. The output from the initial layer is then used as input for the next layer, which is trained using an unsupervised learning algorithm. This is repeated until the entire DBN has been initialized with weight values (Bengio, 2009). At this point, back-propagation can be applied to the DBN in order to make the learning more accurate (Arel, Rose, & Kanowski, 2010). But in this case, back-propagation is more successful on deep networks because each layer is already weighted relatively correctly. The probability of descending a poor local minima is decreased because a particular weight is already close to optimal.

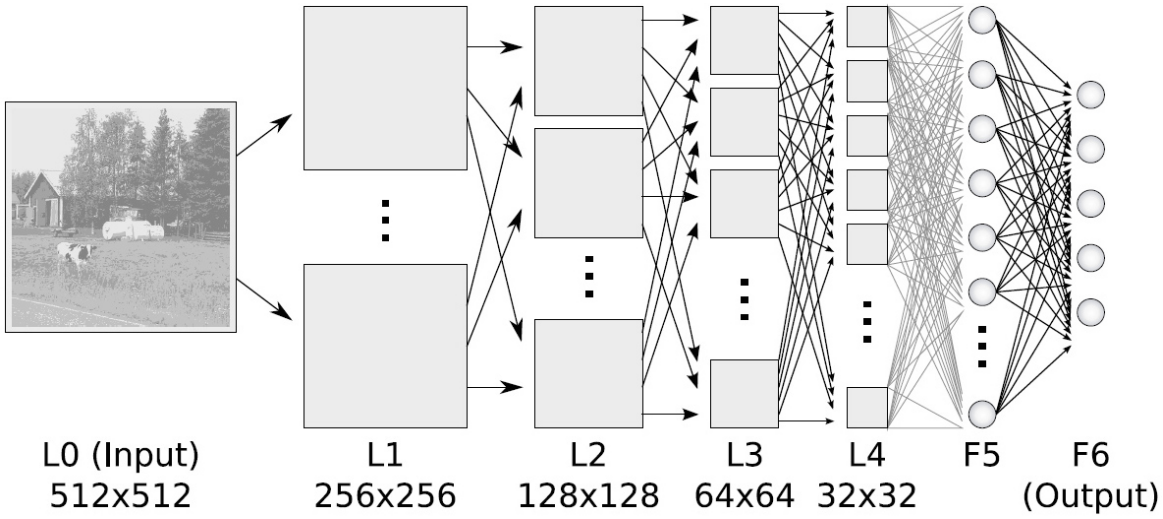
Not only does this method make learning on deep networks merely possible, but it improves efficiency, as well. Applying back-propagation on its own can require thousands of iterations over the network. By training each layer individually and using back-propagation *after* that initial training, rather than *as* the initial training, the network does not need to be traversed over as many times. In fact, training on a Deep Belief Network could, theoretically, be a one step process (Arel, Rose, & Kanowski, 2010).

Another important type of deep learning structure is the Convolutional Neural Network (CNN). Like DBNs, these networks are deep, multi-layer networks. But unlike

[3] A "greedy algorithm" is a type of algorithm that aims to solve the larger problem by solving smaller problems first. In other words, an algorithm of this sort finds the best solution for the entire task by finding the best solutions for each sub-task in the order that they are encountered.

DBNs, CNNs can be trained in a traditional manner, such as back-propagation (Bengio, 2009), since they are feed-forward networks. As is characteristic of neural networks, CNNs were motivated by biology, specifically, the human visual system (Bengio, 2009) and its hierarchical structure. It is natural, then, that CNNs have been applied to image and pattern recognition tasks with great success. Convolutional Neural Networks are commonly composed of five to eight connected layers of neurons, making training with random initial weights immensely difficult. Despite this difficulty, CNNs have been generally successful in vision and image related tasks.

Figure 7: This CNN pools the output of each layer such that the abstraction of the original input image is greater in each layer. Therefore, the number of neurons necessary to represent the image lessens as you traverse the network.



Convolutional Neural Networks rely on the principle of decreasing the number of learning parameters, via filtering, in each layer of the network (Arel, Rose, & Kanowski, 2010). This reduction in the quantity of parameters results in fewer connections between neurons, which improves the effectiveness of back-propagation training (Arel, Rose, & Kanowski, 2010). Generally, CNNs contain two types of neuron layers: convolutional layers and subsampling, or pooling, layers. The basic idea behind CNNs is that neurons in one layer

are associated with neurons from another layer. Also, the neurons in a particular layer are representative of a certain area of the image; the location of the neurons is significant (Bengio, 2009). Output from each layer is pooled before being fed to the next layer of neurons, where it is filtered, pooled and given to the next layer (Krizhevsky, Sutskever, & Hinton, 2012). The specific method of pooling typically changes depending on the project, as it is a feature designed by the researchers (Arel, Rose, & Kanowski, 2010). Figure 7 demonstrates the pooling principle of CNNs. The abstraction level is increased in each layer, so the number of neurons decreases.

And while it is not yet fully understood why CNNs are successful when more traditional deep networks are not, there are a few hypotheses. One reason that CNNs do so well is the relatively low number of inputs for each neuron. This would mean that the weight adjustments made during back-propagation would not diminish over so many layers of neurons (Bengio, 2009). Another hypothesis, which could work in cooperation with the low input hypothesis, is that the general hierarchical structure is well-suited for vision tasks. This is supported by research that has shown CNNs with random weights in the initial layers perform better on vision tasks than a fully-connected network that has been trained (Bengio, 2009). Further support for this hierarchical hypothesis is biology; CNNs are based on the human visual system's structure, so a replication of that will probably work well for visual tasks.

As I have explained different and important concepts and architectures in deep learning, I will examine the three theories of consciousness I mentioned earlier: the Multiple Drafts Model, physicalism, and Integrated Information Theory. Although I will

not explicit apply these theories to deep learning computers until Chapter 4, the particular conclusions may become apparent over the course of the next chapter.

Chapter 3: Theories of Consciousness

Sometimes, it can be frustrating to traditional scientists, as well as those new to philosophy, that topics in philosophy are not as explicitly empirically testable as topics in physics or chemistry may be. The answers to questions of testing new hypotheses and theories is often less clear-cut and defined in philosophy. And so it should be of no great surprise that there is notable disagreement about the answers to questions such as, "What is the nature of consciousness?" and, "What makes some thing conscious?"

As we have seen in the previous chapter, there is substantial variation within deep learning, primarily with respect to the architectures and structures of the networks. Deep learning, and machine learning in general, is comprised of distinctive algorithms, implementations, and motivations. In many ways, the philosophical subject of consciousness is similar to deep learning. However, one of the differences between the two lies in the testability of each field's respective solutions. When developing a deep learning algorithm or technique, and the subsequent deep learning program, one can assess the accuracy and correctness of the program by simply applying it to some task, such as identifying faces in images or categorizes handwritten letters. This can give the researchers immediate and definite feedback about whether or not that particular algorithm or program is a good one. Unfortunately, philosophers are not lucky enough to experience such immediate justification. When discussing the nature of consciousness and what the necessary properties of consciousness are, the process of checking the validity and truthfulness of the proposed answers and theories is more difficult.

It seems natural, then, that new theories of consciousness have been developed and presented since the topic's conception, just as different deep learning algorithms and architectures have been developed and presented. But while two philosophers may disagree about the plausibility or validity of some theory *T*, there have generally been no ways to resolve the disagreement and completely dismiss the *T* in the way one could dismiss a poor network design in the field of deep learning. Thus, theories of consciousness have continually been presented and modified without ever fully eliminating prior theories. And for that reason, deciding on which theories to use to examine deep learning can be an arduous task. That said, let us turn towards three influential theories of consciousness: a form of traditional functionalism (Levin, 2013) and Daniel Dennett's Multiple Drafts Model (Dennett, 1991), Ned Block's physicalist, "hybrid theory of consciousness," (Block, 2002) and Giulio Tononi's modified functionalist Integrated Information Theory (Tononi, 2004). These three theories all attempt to achieve the same aim of explaining consciousness, but they do so in interestingly contrasting manners. Furthermore, they each carry unique implications about what makes some thing conscious. For the interest of this paper, that thing is a deep learning computer.

Functionalism and the Multiple Drafts Model

It would be incorrect to say that functionalism is a single theory of consciousness. It would be more accurate to view functionalism as a category or school of theories of consciousness. There are many different philosophical theories that share important characteristics, and perhaps these characteristics could be attributed to some single view

called "functionalism," but, in reality, each theory differs in enough significant ways that it would be wrong to state that any given theory *is* functionalism, rather than stating that a given theory is a functionalist theory. One particular thesis of this sort is Daniel Dennett's Multiple Drafts Model (MDM), which he has developed over the course of his career and most explicitly presented in his book *Consciousness Explained* (Dennett, 1991). But in order to fully understand the Multiple Drafts Model, and how it is a functionalist theory of consciousness, an examination of traditional functionalism, as a doctrine, is necessary.

To put it plainly, functionalism is the view that mental states are just functional states (Block, 1981). Any mental state, such as being in pain or seeing red, is equivalent to the functions and causal relationships that the state has within a system (Block, 1996). That is, to say what it is to be in a certain mental state is just to say what that state's responses (outputs) to certain stimuli (inputs) are, in terms of the functional role of that state (Levin, 2013). A common and simple example is *pain*. Functionalism maintains that being in a state of pain just *is* being disposed to a state *P* such that *P*'s relationship with other states is functionally equivalent to the pain states that one would be in when in experiencing pain. Such "pain functions" may include producing states of anxiety, causing the person to stop and avoid the actions that are believed to cause the pain, or provoke a the pained person into saying 'ow.' Normally, mental states may differ in some ways, but are said to be of the same type when they share the same functional properties. For example, the mental state of feeling pain from stubbing my toe and the mental state of feeling pain from touching a hot pan may have different properties specific to each pain, such toe vs. hand, low severity vs. high severity, dull vs. burning, etc. Yet, there is *something* about both states that makes them both pain states. Functionalism posits that

the shared property between the two is a functional property (Block, 1981). Both states produce the same behavior and serve the same functional role within the system that they occur.

Due to this principle of functionalism, a single functional state, that is, a state that describes a mental state in non-mental terms, can be realized in more than one way (Block, 1996). The *realization* of a system is the manner by which the system is constructed. A system *S* that honks a horn when it is hit with a bat can be realized in various ways. *S* can be a series of pulleys and ropes, or it could be a series of hydraulic pistons and gears, or an electronic computer with an electronic G-force sensor, or it could even be a biological animal like a human. In any of these cases, if we were to describe the mental state of pain as that state whose functional role is honking a horn when hit, then, according to functionalism, all four of these different realizations of *S* can be said to be in pain. Each realization of *S* produces the same output (honking a horn) in response to the same input (being hit with a bat). Since functionalism characterizes the mental state of pain in terms of these inputs and outputs, they can all be said to have the state of pain (Block, 1996) (Block, 2002).

This is, of course, an incredibly simplistic and vague explanation of functionalism, which some functionalists may take issue with. But for the purpose of identifying the functional nature of Dennett's Multiple Drafts Model, I believe it to be sufficient. Even with such an unembellished functionalist view, one can begin to imagine how one might prefer functionalist theories of consciousness, especially in the face of two consequences of other views: *liberalism* and *chauvinism*. Functionalists commonly argue that liberalism

and chauvinism are undesirable ramifications of behaviorism and physicalism, respectively, that functionalist theories avoid (Block, 2002).

Liberalism is an objection functionalists generally have to behaviorism. Behaviorism can be seen as a simplified version of functionalism that attributes mental states, such as pain, to any system that emits a certain output when given a certain input (Block, 2002). This differs from the basic form of functionalism I outlined above in that the system does not have to be the same, internally, to be in the same state. The issue functionalists commonly have with behaviorists, and the point at which the two views diverge, is that behaviorism tends to ascribe mental states to objects and systems that do not, and cannot, have mental states in reality (Block, 2002). For example, imagine a great actor that can act in pain on cue. When he acts as if his leg has been broken, it is entirely believable to observers; there is no discernible difference between the actor acting like he broke his leg, and him actually breaking his leg. When tapped on his leg, the actor acts as if his leg has just been broken. In this case, a behaviorist would say the actor actually *is* in pain, and thus, has a mental pain state, since he behaves as though his leg is broken (which is considered painful by almost all people) after it is hit. A functionalist would disagree, however. Functionalism maintains that the actor can only be regarded as having a mental pain state when he both behaves as if he is in pain (produces cries and moans) *and* has the appropriate internal desires and functions. In other words, the actor's internal states would have to react appropriately, such as nerve firings and a true desire to avoid further contact on his leg. Without an internal desire to produce an output, a system does not have a certain mental property (Block, 2002).

Similar to how functionalism deviates from the behaviorism on the basis of liberalism, it rejects pure physicalism on the basis of chauvinism. Physicalism, in this sense, is the idea that all mental states, such as pain or seeing the color red, are identical and equal to the physical state a system is in at the time of experience (again, this is a highly simplified description) (Block, 2002). The issue many functionalists have with physicalism is that physicalism does not grant mental properties to systems that do, in the eyes of functionalism, have mental properties (and minds), in reality (Block, 2002). Since functionalism relies on internal desires and mental states to avoid the problem of liberalism, it makes sense that it would reject a theory that eliminates mental states and internal desires.

But there is a problem here. How can a functionalist avoid liberalism *and* chauvinism simultaneously and still offer a satisfying account of consciousness? It seems as though if a functionalist is willing to accept, and in fact, require, the existence of mental states in order to avoid liberalism, then in what capacity, and for what purpose, does behavior and function matter? If internal desires and mental states are the necessary factor in consciousness, then why not simply posit that? Likewise, if a functionalist is going to evade chauvinism by appealing to the fact that mental states do exist and one cannot ignore them, then it seems again that functionalism is ignoring function and behavior. In essence, functionalism seems to assert that a system has mental states, and consciousness, when the system has mental states, and therefore, consciousness. Anything else, or less, leads us to chauvinism or liberalism. So how this problem be reconciled? Daniel Dennett believes that his Multiple Drafts Model can, either directly or indirectly, solve this issue while retaining the key principles of functionalism.

Due to the great influence Rene Descartes' idea of Cartesian dualism has had in the way we think of the relationship between the mind and the body, we have a natural inclination to think of our perceptions as being serial. That is, the sequence of experiences we have is the same as the sequence of "arrival" of the stimuli. The order that stimuli enter the system (our brain) is the same order that we experience the stimuli. This idea relies on some central mark or boundary that stimuli must reach to be experienced. Daniel Dennett refers to this view of how humans and similar systems process information the "Cartesian Theater." This view of conscious experience remains persuasive and attractive simply because it seems to us as though experiences and events occur in this manner in our daily lives (Dennett, 1991). But when the brain is observed at the microscopic level, over an incredibly short period of time, the Cartesian Theater view begins to run into trouble. Mostly, empirical findings suggest that a single, chronologically-based point in the brain, through which all stimuli must pass in order to yield consciousness, is highly implausible, if not completely non-existent.

A famous example is the "phi phenomenon." In this case, two dots are flashed in quick succession and relatively close together in a patient's visual field. Rather than experiencing two dots in two separate locations on the screen, the patient experiences a single dot moving back and forth. And when the two dots are of different colors, say the first dot is red and second is green, the patient experiences something rather interesting. The patient initially experiences a red dot, which begins moving just as before. However, while in the middle of its [imaginary] transit, the dot turns from red to green. The common question to this result is, "How is it possible for the patient to experience a color change before the second dot has been flashed?" Under the Cartesian Theater view, this seems

impossible without some predictive capabilities in the brain. Therefore, the color change that is experienced must be experienced *after* the green dot has been flashed and the patient has processed the green dot. However, Dennett believes this is wholly wrong, and to think about a solution in such a way is incorrect. To do so, one would have to accept one of two theories about consciousness and information processing: Orwellian and Stalinesque revisions.

The Orwellian revision method alludes to the Ministry of Truth of George Orwell's novel *1984*. Similar to way the Ministry of Truth rewrote history and denied the truth of what actually happened, the Orwellian method posits that our brains rewrite our memory and erase the old version of an event, so that we are not aware that any revision has been made. In the phi phenomenon case, the patient would experience the red dot and then experience the green dot. Realizing that the two experiences do not cooperate well enough, the Orwellian mechanism would rewrite the experience to be that of a red dot moving and changing into a green dot, and erase any memory of experiencing two separate dots (Dennett, 1991).

In the case of a Stalinesque revision method, the brain simply creates and fills in an experience that makes sense. The method, which hints at Joseph Stalin's infamous show trials, advances the idea of a pre-conscious editor in the brain that receives the first experience of the red dot, then the second experience of the green dot. Given these two experiences, the editor creates and inserts an intermediary experience that connects the red and green dots, and sends the entire sequence of experience to the brain's stream of consciousness (Dennett, 1991). Despite the attractiveness or intuitiveness of either the Orwellian or Stalinesque revision methods, Dennett concludes that we have no good

reason to accept either of these theories of revision over the other. The problem with both methods is the appeal to a time before awareness for revision to occur and central point for consciousness for information to pass through. This means both revision methods are grounded in the idea of the Cartesian Theater.

The Multiple Drafts Model rejects the Cartesian Theater model and replaces it with a model in which the brain processes information and input in a parallel manner. Furthermore, the information is continually available for editing and revision (Dennett, 1991). This parallel processing works in tandem with the idea that stimuli are not immediately experienced; the interpretation of inputs takes some amount of time. During this time, that is, between the initial arrival of the stimuli and the conscious awareness of the experience, the information can be edited and changed. Another characteristic of the Multiple Drafts Model is that once some input is processed in a specific location in the brain, the information is available to the entire system and conscious stream. There is no central point the processed information must be sent to in order to enter the stream of consciousness (Dennett, 1991).

But where do we get our experience of consciousness from? How do all of these bits of processed input and information yield a stream of consciousness? Dennett concedes that these questions are still open, and MDM does not explicitly address these concerns. However, he does maintain that the collection and availability of these information bits yields something that is like a single stream of consciousness, and that the only reason it is not considered an actual single stream is its multiplicity; "at any point in time, there are multiple 'drafts' of narrative fragments at various stages of editing in various places in the brain" (Dennett, 1991).

Returning to the phi phenomenon, the Multiple Drafts Model does not appeal to any "retrospective, content creation." There is nothing to "fill in," nor is there a time or space to do the filling in. The brain does not construct new experiences in order to explain the real, "old" ones. Instead, experiential information is merely edited and the new edition is made available to the "stream" of consciousness, thus, effecting subsequent behavior. In fact, it would be a waste of time and power for the brain to intercept every experience solely for the purpose that it may have to be edited. The Multiple Drafts Model allows the brain to edit only the information that is necessary to make sense of some experiences. If no editing is necessary, no time or processing power is wasted (Dennett, 1991).

Recall that one of the objections I raised to functionalism was that its relationship to the problems of chauvinism and liberalism was peculiar. Dennett responds to this notion by arguing that the non-existence of qualia, the term normally used to denote phenomenal qualities, removes it entirely from the discussion of consciousness. Dennett takes 'qualia' to simply mean the way things look, smell, feel, etc. By this definition, qualia, in some sense, refers to our experiential mental states. But Dennett does not believe that qualia exist. In philosophy, the discussion of qualia is normally a discussion of epiphenomenal qualia. The problem is, the definition of 'epiphenomenal' is that of being an effect, but having no effect in the physical world (Dennett, 1991). Thus, there is no way of identifying epiphenomenal qualia, for identification would require some effect that is identifiable (additionally, identification would itself be an effect) (Dennett, 1991). When some person states that he has epiphenomenal qualia, that he is experiencing some phenomenal qualities, he is saying this in spite of qualia having no effect on him; there is no evidence, for himself or another person, that his experience has epiphenomenal qualia.

He would, and could, therefore, say he has qualia just in case that he does not have qualia. It follows that qualia, and thus, mental states and experience, cannot play a part in the discussion of consciousness because there is no evidence for, nor a way of determining, the existence of those states and properties. All that remains are *judgments* that one's experience "has" qualia, and any functional or behavioral effects that such judgments carry (Dennett, 1991).

In this way, Dennett's position seems to be closer to that of pseudo-behaviorism, in that he does not believe qualitative mental states play a role in consciousness (in fact, he denies their existence altogether). But he is not a true behaviorist, in the sense that he does not totally denounce mental states, or that their perception can have some functional or behavioral role in consciousness. For Dennett, it is fine for a person or system to have mental states and qualia, though Dennett would deny that a person or system could, for such mental states and qualia are neutral in the discussion of consciousness. Dennett is far more concerned with judgments of qualia, and the role such judgments have within a system.

It should be clear how Dennett's form of functionalism is exempt from the chauvinism vs. liberalism problem that traditional functionalism faces. However, his claim that mental states and qualia do not exist is not uncontroversial. One theory of consciousness that would reject Dennett's qualia neutrality is Ned Block's "Hybrid Theory of Consciousness," which actually insists that qualia, and phenomenal properties and experience, are an essential part of consciousness.

A Physicalist Theory of Consciousness

While functionalism may be attractive to some because of its more traditional and intuitive notions, such as reliance on judgments about qualitative experience (this seems to be similar to the idea that consciousness is "awareness"), there are some philosophers that believe this is not all that conscious experience is; there is something more to being conscious, namely, phenomenal experience, or qualia.

Take, for example, Ned Block's "hybrid" or "mongrel" theory of consciousness (as he has not officially named his theory, I derived this name from his description of consciousness for the purpose of this paper). This is a theory that generally seems to challenge functionalism in its premises, conclusions, and implications, though it is not clear that the two theories are mutually exclusive. Block's hybrid theory relies on his assertion that consciousness, as we normally speak and think of it, is actually a hybrid of two forms of consciousness: phenomenal consciousness (P-consciousness) and access consciousness (A-consciousness) (Block, 2002).

In the past, it had been suggested that consciousness is simply experiential. When some thing, *S*, is said to have consciousness, it means there is something that it is like to be *S* (Nagel, 2002). There is something it is like to be each of us. I can imagine what it would be like to be you, have your experiences, think your thoughts, etc. If there is no way to describe what it is like for *S* to be *S*, then we cannot say that *S* has conscious mental states. This excludes certain objects, such as robots and basic organisms, from being considered conscious, since we cannot imagine, or even attribute to them, a sense of what it is like to be them. But Block does not agree that consciousness is purely this idea of subjective experience and "what it is like-ness," so to speak. Instead, he accepts the

basic premise that there is something that it is like to be a conscious being, and calls that premise phenomenal consciousness (Block, 2002). In contrast to the Multiple Drafts Model, Block maintains that judgments and functions do not sufficiently explain phenomenal consciousness and experience.

As P-consciousness is roughly synonymous to phenomenal experience, it naturally follows that P-conscious states are those mental states that occur when we have experiences, such as seeing, hearing, tasting, and feeling sensations. Block wants to extend P-conscious states to also include those states that relate to thoughts, wants, and emotions (Block, 2002). Whether or not such states should be considered P-conscious states is arguable, but for the time being, I will grant Block their inclusion.

In the case of A-consciousness, forming a clear definition is not as easy and straightforward as it was for P-consciousness. Just as P-conscious states are representations of experiential properties, A-conscious states can be described as mental states that are available to a system's rational processes. However, this availability should not be confused with "reportability" (Block, 2002). That is, a state can be an A-conscious state even in cases in which the subject cannot report on any information. For example, a dog, which could be assumed to be conscious to some degree, should not be deprived of having A-consciousness simply on the basis that he cannot verbalize, or tell us about, the A-representation. While being able to report on an A-representation may be the most practical method of concluding whether or not a subject is A-conscious, it is not the only method. In the case of the dog, it seems clear that the dog does have A-conscious states when he is able to recognize his owners or navigate his way around the house. Presumably, there are P-conscious states that are representations of his experiences of his

owners' faces and the floor plan of the house (as well as numerous other experiences). When he sees his owner, or turns a corner, an A-representation is "accessed," thus providing the dog with the information he needs to perform the next appropriate action. A simplified manner of thinking about A-consciousness is that it is loosely synonymous with the way people, including Dennett, usually think of and use the term 'conscious.' It is rather common to hear people attribute consciousness to something or someone when that thing or person exhibits a sense of awareness. It is important not to confuse A-consciousness with the colloquial definition of the word conscious, but the two concepts are similar enough that comparing them can provide a better understanding of A-consciousness.

So fine, there are these two forms of consciousness, A-consciousness and P-consciousness. But what is the relevance of making such a distinction? Why is it insufficient to simply accept Nagel's notion of what consciousness is and chalk it all up to experience? The A/P distinction allows us to consider another problem related to consciousness: how the mind is structured or realized.

There are generally two sides to the debate about realizations of the mind. The *biological* view holds that the actualization of a system matters in terms of the mind and consciousness. John Searle, for example, could be seen as an advocate of the biological approach. He was of the opinion that an artificial thing could think and be conscious only if it shared key causal and organizational characteristics that are found in organic human brains (Searle, 1980). His view, which was a physicalist view, stated that consciousness and intentionality in thoughts and actions are derived from causal processes found in developed animal brains. It is only through replicating the structure and processes of these

(already assumed conscious) brains that one could theoretically create a conscious, thinking machine.

The other view, denoted as *computational*, is one we are already familiar with, as it can also be called behaviorism. To briefly reiterate, behaviorism regards consciousness in terms the production of output based on some input. Similarly, information processing can summarize the computational view of the mind, in that the mind can be characterized and explained by how information is handled. A computationalist, or behaviorist, does not believe the structure or make-up of some thing has an effect on the presence or absence of a conscious mind. Again, if system *P* performs and behaves in the same manner as a conscious, "mind-having" being does, then *P* is said to be conscious; it does not matter if *P* is a human with an organic brain or a robot with electric circuits. In relation to A-consciousness and P-consciousness, if A-conscious states are the same as P-conscious states, then the computational approach would be correct. This would mean that qualitative experience, P-consciousness, is nothing more, and is no different from, one's ability to use information in rational processes. However, if it is the case that A-consciousness and P-consciousness truly are distinct from each other, then it would follow that the realization of the thing in question does matter, for some realizations may not permit either A-consciousness or P-consciousness.

The natural question to ask at this point may be, "Are access consciousness and phenomenal consciousness distinct?" But I think it is equally important to question how we can determine if they are independent or not. Ultimately, there are two ways one could go about this task. The first is to plainly examine and identify any differences in the very nature and properties of A-consciousness and P-consciousness. The second, perhaps a bit

more difficult, is to establish and analyze instances of A-consciousness being present without P-consciousness, or P-consciousness being present without A-consciousness, in the real world.

One difference between the two forms of consciousness is that A-consciousness is functional, while P-consciousness is not. That is, A-conscious states are identified by their role, or function, within some system. If a mental state *S* is not used in reasoning or rational thought, then *S* is not said to be an A-conscious state. P-conscious states, on the other hand, do not rely on such contingent use. A mental state *S* is P-conscious just in case that *S* is a mental state with content "inside" of it. *S* does not need to be used in the way that it does in order to be A-conscious; *S* merely needs to exist at all (in fact, it is highly difficult to imagine a mental state without any content or information).

Another dissimilarity between access and phenomenal consciousness is that P-consciousness can be described as having a "type." There is something about an experience of pain, or hearing a violin, that every other experience of pain or hearing a violin has. It is this property that makes a particular experience an experience of pain or hearing a violin. This property can be labeled as a P-conscious type, and so, two P-conscious states are of the same type when they share some intrinsic property or properties.

A-consciousness, on the other hand, cannot be said to have such types, or at least, there is only one type of A-consciousness: A-consciousness itself. When a state or representation *X* is said to be A-conscious, it is a result of the fact that *X* is being used in rational thought and behavior at the time T^1 that *X* is being labeled A-conscious. It could be the case that five seconds later, at time T^2, *X* is no longer being accessed. At time T^2, *X*

is no longer A-conscious. Analogously, when one uses a hammer in construction, the hammer has the property of 'usefulness;' the hammer is in a "useful state." And when one uses a saw, that saw is also useful, and thus, in a "useful state." While both tools are useful, it would be wrong to claim that they are the same kind of useful (e.g. cutting things and hammering nails). Yet, they share the same property of 'usefulness.' A-conscious states, then, may be accessed differently, and be representationally different, but the fact that they are being accessed at all, in some way, makes them A-conscious states. Ultimately, it would be incorrect for one to say either that P-conscious states only come in one flavor, P-conscious, or that A-conscious states come in various forms (to illustrate the falsehood of such a claim, I cannot think of two different forms of A-consciousness to use as an example).

But maybe there are people that do not find these differences to be entirely satisfying. Perhaps, they are much more convinced by real, non-theoretical examples of the A/P distinction. Fortunately, such examples can be provided. Let us first consider A-consciousness without P-consciousness.

There is a phenomenon called blindsight that occurs when a subject has experienced brain damage that affects his visual processing. As a result, his visual field contains blind spots. However, the interesting aspect of blindsight is that patients are capable of accurately "guessing" what has been flashed in those blind spots, despite not being able to actually see anything. Now, this regular blindsight situation is not an example of A-consciousness without P-consciousness, since there are no states that can be said to be A-conscious. Whether or not there is some state that is a representation of the information in the blind spot, the patient is unaware of such a state and, as far as he is

concerned, he is simply making a guess about what is there. There is no state that is intentionally being accessed for the purpose of determining what is in the blind spot. Furthermore, there is no P-conscious state either. Since the patient cannot see in the blind spots, there is no experience of seeing something in the blind spots, and therefore, no phenomenal information. This absence of P-consciousness helps to explain the absence of any A-consciousness, since the patient's rational faculties clearly are not able to use information that does not exist; there is no information to be made available.

But now consider something similar to blindsight where the patient is somehow able to realize there is something in his blind spots that he is not seeing and therefore, makes a guess, with the same amount of accuracy, about what is there. Block calls such a situation *superblindsight*. A superblindsight patient would suddenly become aware that there is something in front of him that he is not seeing, and would proceed to make a guess about what it is, much in the same manner as a regular blindsight patient. This sudden awareness is not explainable in phenomenal terms, since there is still no visual experience, but there is now A-consciousness. The patient somehow knows that there is information he is not experiencing visually, and that perception of missing information is A-conscious, as it is available to, and used in, his rational decision to make a guess about what is in the blind spot.

Perhaps, in an attempt to deny any independence between A-consciousness and P-consciousness, a functionalist could object to this superblindsight scenario on the grounds that it does not exist, which Block admits. It could be true that the reason superblindsight does not exist in reality is due to the fact that it is not possible, and the impossibility is due to the fact that A-consciousness and P-consciousness are not actually distinct concepts,

but instead, forms of the same concept. It seems rather *ad hoc* for Block to admit that blindsight, which is a real phenomenon, does not constitute an example of A-consciousness without P-consciousness, and for us to present an example of A-consciousness without P-consciousness, we must imagine a situation that does not actually occur.

Block, and other advocates of the A/P distinction, may have a response to this objection that is similar to an argument for immaterialism. They could state that the mere fact that we can imagine superblindsight is that A-consciousness and P-consciousness are conceptually different; there is nothing that prohibits us from being able to imagine one occurring without the other. There is no logical inconsistency or impossibility that prevents us from thinking about one without the other, and so there is no logical necessity between the two. But another way of responding to the objection to superblindsight is to avoid the problem altogether and turn to an example of P-consciousness without A-consciousness that has empirical backing. Hopefully, an example of this type will be more convincing to those hesitant about accepting the A/P distinction, which would ultimately refute the idea that perceptual experience is vital in consciousness.

While Block believes it is helpful to imagine a scenario in which a conscious being's rational capabilities have been removed as the result of brain damage, there is actual empirical evidence of P-consciousness without A-consciousness. The first piece of evidence that can be presented comes from the observation of neural activity in the presence of visual stimuli. Experiments on patients show that stimuli of a short time duration or of weak enough strength, invoke neural activity without the patient's awareness of any such stimulation (Tononi & Edelman, 1998). In other words, there is

some sort of experience of the visual stimuli, as evidenced by the brain's response to it, but there is no knowledge or consciousness (in the colloquial sense) of it. Recall that P-consciousness is experience and A-consciousness is (roughly) awareness of experience or availability in rational processes. In the case just mentioned, there is a response to the visual stimuli in the brain, which would suggest P-consciousness[4]. The brain, though perhaps not the mind, is having an experience of seeing something, and there is a corresponding brain state in which the stimuli were seen. Remember, for a physicalist like Block, two physically identical brain states represent the same conscious state. However, the absence of any awareness of the experience suggests that the patient showed no A-conscious state, nor is there any chance of A-consciousness. If there was an A-conscious state, one that was available for rational thought and cognitive activity, then the patient would have been able to confirm that he had seen the stimuli. There are two potential explanations for this unawareness. The first is that there was a stimulus-representation state created, but this was inaccessible. The second is that no such representation ever existed. In either situation, there was no A-consciousness. And it appears as though it is true that P-conscious states are created independently of any A-conscious states, thus, P-consciousness is separate from A-consciousness and the A/P distinction holds.

This example is in some ways more satisfying than the blindsight and superblindsight examples. Primarily, it is based on tested experiments and empirical evidence rather than conceptually debatable thought experiments. It seems safe to conclude, based on the evidence above, that the A/P distinction is true. But where does

[4] The fact that the brain responds in the same manner in this case and cases when the person knowingly experiences the vision indicates that there is *some* type of phenomenal information being processed in the brain.

this leave us in terms of determining the most plausible theory of consciousness? As I mentioned above, the A/P distinction would imply functionalism to be conceptually false, so perhaps we can rule out functionalism to be a viable theory of consciousness. I will not make such definite conclusions at this point, but it should be noted that functionalism, in its simplest conception, might be untrue.

Still, all we have done up to this point is show that these ideas of access consciousness and phenomenal consciousness do not denote the same thing. How does this distinction form a theory of consciousness as a whole? Remember, Block believes consciousness is a mongrel concept; it is comprised of different forms of consciousness, the two primary ones being access and phenomenal consciousness. Something is generally only considered conscious when the various forms are presented together as a cluster. When one or more is missing from the cluster, especially P-consciousness, consciousness as a whole seems to be missing. For example, when A-consciousness is missing, as is the case with Freudian states and desires, we normally think of those states as unconscious states. And when P-consciousness is missing, in cases like blindsight (and superblindsight), it is not natural to attribute consciousness to those states either. Consciousness, as we think of it, typically arises in a subject when both A-consciousness and P-consciousness are present and when they interact with each other.

Must we accept this notion of consciousness merely because of the A/P distinction? Perhaps we do not. Perhaps the A/P distinction can be true regardless of the veracity of the idea that true consciousness arises when A-consciousness is accompanied by P-consciousness. Without elaborating excessively on such an objection to the Block's

ideas, one could argue that attention, that is, awareness and mental access, is in fact a phenomenon of consciousness, rather than an integral player (Smithies).

Attention and A-consciousness are roughly interchangeable terms. Attention is functional much in the same way that A-consciousness is functional. In cognitive activity, attention's function is to "make information accessible for use in the rational control of thought and action" (Smithies). The similarity to Block's concept of access consciousness is clear, as A-consciousness is responsible for making information available and accessible to rational processes (Block, 2002). However, while Block asserts that attention, or A-consciousness, is a part of the more general concept of consciousness, other philosophers, such as Declan Smithies and William James, believe that attention, or A-consciousness, is a phenomenon of consciousness, similar to other phenomenons such as pain and guilt. Phenomena like pain and guilt are known to conscious beings through experience, yet the concepts are difficult to define in terms that are anything more than examples. Similarly, conscious beings are aware of attention, but often have a tough time explaining exactly what it is (Block even spends a portion of his paper discussing the difficulty he has in explaining and defining A-consciousness). This struggle arises in part because of our misunderstanding of what attention truly is. Some believe attention is essentially a focalization or modification of our stream of consciousness (James, 1890) (Smithies). By this definition, attention is what makes information accessible in thought and cognitive activity, thus making attention a mode, or employment, of consciousness. In other words, attention, like pain, is a phenomenon of consciousness, not a part of consciousness. The implication is then that consciousness can exist with attention, but attention cannot exist without consciousness. In regards to Block's theory of

consciousness, it may be true that A-consciousness and P-consciousness are distinct, but it is not necessarily true that some thing is not conscious when missing one of those two forms of consciousness (in this case, A-consciousness). This objection may not be entirely sufficient in refuting the hybrid theory, and it does not do much in terms of restoring functionalism as a viable theory, as it seems to imply consciousness is more closely tied to experience than the hybrid theory does. But it at least raises questions about the assertion that multiple forms of consciousness are necessary for us to consider a subject as being conscious.

Regardless of the conclusiveness of this objection to Block's view of consciousness, let us focus on a fundamentally different theory. Integrated Information Theory is a theory of consciousness that starkly contrasts with Block's ideas and is more closely tied to functionalism.

Integrated Information Theory

While consciousness has been discussed primarily within philosophy of mind and psychology, recent scientific advances have expanded the scope of the conversation to include biology and neuroscience. The findings of many experiments in biology and neuroscience have shaped our understanding of the brain and its role in consciousness. Naturally, this scientific evidence has prompted the formation of new theories of consciousness. Additionally, the extension of the discussion of consciousness into more scientific and empirical fields has stimulated non-philosophers' interest in theories of consciousness. One example of a recent theory of consciousness, grounded primarily in neuroscience (as well as some mathematics), is Giulio Tononi's Integrated Information

Theory (IIT). This theory, which seems to rely on philosophical ideas akin to functionalism and materialism, explains consciousness by appealing to two measurable concepts: information, and a system's integration of that information.

At face value, Integrated Information Theory is simple and straightforward. Ultimately, IIT asserts that consciousness essentially *is* a system's ability to receive and process information, and the extent to which that information is integrated within the system (Tononi, 2004). The level of consciousness in a system, that is, the quality of the conscious experience, is provided and shaped by the relationships between informational complexes (a notion I will explain below) in the system (Tononi, 2008). And of course, the bases of this conscious experience are two crucial aspects mentioned above: information and integration.

When Tononi speaks of information, he does not mean just anything; in fact, his idea of "information" is quite specific. In everyday life, 'information' is used to denote data or details. The characteristics and properties of what people experience are referred to as information. Integrated Information Theory does not completely dismiss the traditional or colloquial meaning of 'information,' but it is concerned with a particular aspect of it. Tononi defines information as being the "reduction of uncertainty among a number of alternative outcomes when one of them occurs" (Tononi, 2004). In other words, information signifies and represents the number of representational alternatives are not the case. As the number of possible outcomes increases, so does the amount of information. For example, a simple coin toss will yield one bit of information, since there are two total outcomes of the coin toss, and one alternative outcome to the actual one. Thus, the amount of information in some event can be calculated mathematically using the entropy function,

or the logarithm of the number of alternatives. In the coin toss case, there are two possible outcomes, heads or tails, so there is $log_2 2$ bits of information, which, again, translates to 1 bit of information (Tononi, 2008). When it comes to our everyday experiences, the amount of information is immensely high.

Consider the apparently simple case of determining whether or not a screen is light or dark, where "light" means the screen is showing pure, white light, and "dark" means the screen is off. When the screen turns on, we, as humans, are easily able to determine that it is light. But while this may seem like an elementary task, the amount of information is actually quite high. This is due to the number of possible outcomes that we must discriminate against. The light screen that we do see is different from a light screen that is fully colored blue, or that is showing a picture of a tree. The screen can show an astonishingly high number of alternatives, all of which we are capable of distinguishing between, and all of which would prompt us to deem the screen as being lit up. Thus, the light screen carries with it a high amount of information.

Compare our situation to that of a photodiode, which can only distinguish between light and dark in a strictly binary sense. Regardless of whether or not the screen is showing blue or red or a picture of a tree, the photodiode is only capable of distinguishing between light and dark, or on and off. As far as the photodiode is concerned, there are two possible outcomes, light and dark. Therefore, the same screen has only 1 bit of information for the photodiode, but a large amount of information for the human. Integrated Information Theory seems to imply that the more specifically a system can discriminate one alternative from another, the more information that system is receiving. It is important to note that this ability to discriminate between outcomes is related to the

specificity of the system's representations of the outcomes. That is, if there exist a difference between a system's representation *A* of a tree, and representation *B* of the same tree without a leaf, then the system can discriminate between those two outcomes. When a system's representational power increases, its discriminatory power increases, and the level of (potential) conscious also increases (Tononi, 2008). However, there is another, perhaps more important, component of consciousness that must be taken into consideration: the integration of the information.

The information facet of Integrated Information Theory is not of much interest. Few people, if any, would argue that a photodiode, or other system that is capable of distinguishing between only two possible experiences, is in any way conscious. For IIT, the true origins of consciousness lay in a system's *integration* of information. A system experiences consciousness only when the system integrates information in a way such that subdivisions of the system are not able to integrate the same information independently of each other (Tononi, 2004). In other words, a system has conscious experience when the level of interdependency between parts of the system is high.

Take, for example, a camera with a sensor chip comprised of millions of photodiodes, each one similar to the photodiode in the above example. Since each photodiode can distinguish between two alternative states, light and dark, there are millions of separate states that the camera, as a whole, can distinguish between (increasing the number of photodiodes would exponentially increase the number of distinguishable states, hence, the more megapixels in a camera, the higher the specificity of the image). However, the camera is not considered conscious because each photodiode is causally independent of the other photodiodes in the sensor chip. Thus, we can consider the chip to

be a "collection of one-million photodiodes with a repertoire of two states each, rather than a single integrated system with a repertoire[5] of $2^{1,000,000}$ states (Tononi, 2004). Humans, on the other hand, cannot be deconstructed in such a fashion. The manner in which human brains are structured and integrate information does not allow us to redefine the brain's repertoire of states as the combination of individual brain parts' repertoire of states. That is, we can not say the brain's (a single integrated system) repertoire is equal to a collection of all of its individual neuron's repertoires (similar to each photodiode's repertoire of light and dark). While we can deconstruct a camera's image into pixels corresponding to individual photodiodes, we cannot deconstruct an experience of a tree into bits that correspond to individual neurons; there are causal processes between neurons that result with the entire experience of a tree (Tononi, 2008). Furthermore, we cannot deconstruct the entire experience of a tree to just the visual, audial, or olfactory characteristics. The entire experience is *comprised* of them, but cannot be reconstructed by any fewer than all of them.

It is worth noting that Integrated Information Theory is not suggesting that a physicalist, or a materialist, account of the mind is false. In fact, IIT does just the opposite. It supports, as well as relies on, a materialist view of the mind. Materialism is the idea that all mental states and processes, i.e. the mind, can be explained by the physical, or material, states and processes of the system, e.g. the human brain. It may appear as though IIT denies this claim about the mind when it states that human experience and information integration cannot be deconstructed down to the individual components of the brain, but

[5] Here, 'repertoire' refers to a system's collection of distinguishable outcomes. A photodiode, then, has a repertoire of two states, as it can distinguish on from off, while a human has an enormously large repertoire, as it can visually distinguish colors, shapes, shades, sizes, etc.

this statement should not be confused with opposition to materialism. All is meant by this claim is that the particular state that the system is currently in, or state being experienced, cannot be reduced to components and elements that do not depend on the other components and elements of the system (Oizumi, Albantakis, & Tononi, 2014).

Now that the concepts of information and integration have been defined, and the effects that each concept has on consciousness has been outlined, there are two potentially important issues that need to be addressed. The first is the problem of how information integration yields consciousness, rather than merely indicates it. And the second concern is how we determine which systems are conscious. That is, which systems integrate information in such a way that they are considered conscious systems.

Questioning Integrated Information Theory's ability to explain just how integrated information is consciousness is legitimate. Early versions of IIT avoid this problem by only claiming that consciousness only *corresponds* to integrated information (Tononi, 2004). But more recently developed and modified editions of Integrated Information Theory seem to suggest that consciousness *is*, in some way, the integration of information. These versions of the theory are the ones that could be targeted by the question at hand.

However, it appears as though there is no clear answer to this problem. The objection that there is an explanatory gap between integrated information and consciousness is equally true for many materialist theories of consciousness. Just as there is an explanatory gap between information integration and consciousness, there exists one between physical brain parts and consciousness[6]. But, IIT comes as close as it can to bridging that gap by asserting that experience is modal, and different modes of experience

[6] This is not necessarily true for all materialist and physicalist accounts of consciousness.

correspond to subsets of mechanisms and information integrating subsystems. Also, there are elementary experiences, like those of pure, primary colors, or pain, that cannot be explained in terms that are not merely synonyms or examples. This is not necessarily a flaw of IIT specifically, it is simply a challenge that many, if not all, similar theories of consciousness face. In IIT, such elementary experiences correspond to equally elementary subsets within the system being considered. In fact, it looks as though Integrated Information Theory may come a step closer to linking consciousness and physical properties than other, more general, materialist and physicalist theories of consciousness.

Amazingly, neuroscientists have been able to identify which neurons are responsible for particular experiences, such as feeling anxious. Unfortunately, they are still unable to offer a satisfying response to the questions of *why* or *how* that neuron, or group of neurons, gives rise to the experience that it does. Integrated Information Theory is at least able to state that that neuron, or group of neurons, integrates particular information in a manner such that [phenomenal] consciousness arises and the elimination of the neuron, or group of neurons, would destroy such consciousness. So while Integrated Information Theory's conclusions may not be wholly satisfying to any explanatory-gap-objectors, it can at least give them this new evidence and point them in the direction of the larger materialism debate, which I do not wish to address at length in this paper.

The second worry one may have concerns the way by which an Integrated Information Theorist determines whether systems are conscious or whether they are unconscious. Due to the fact that IIT describes consciousness in mathematical and quantitative terms (Oizumi, Albantakis, & Tononi, 2014), and those quantitative terms yield qualitative experience, the quality, or level, of consciousness in a system can

effectively be measured and calculated. This calculation of consciousness appeals to the measurement of information, as discussed above, as well as the measurement of a system's integration of that information.

As I outlined earlier, the amount of information represented by a particular state *S* is related to the entropy of *S.* But that is only a minimal description of information, and not one that is practical to use in the calculation of a system's consciousness. Instead, consciousness should be said to be impacted specifically by *effective* information. The amount of information in a system's state is not simply the number of possible states it can distinguish between, but instead, the number of states that are causally integrated within the system. In order to determine the amount of effective information in a particular system, one can examine the relationship between two partitions, *A* and *B*, of a system *X*, where $B = X - A$. By replacing *A* with *A*'s maximum entropy, that is, the number of individual states that are equal to the states represented by *A*, the entropy of *B* can be derived[7]. The effective information of *X* from *A* to *B* is equal to the amount of information shared between partitions *A* and *B*. The fact that *A* is nothing more than a collection of individual states means that any information *B* shares with *A* is the result of "causal effects of *A* on *B*" (Tononi, 2003), and not any effects of *B* on *A*. As the causal connection (integration) between *A* and *B* increases, so does the effective information from *A* to *B*; different outputs from *A* will produce differences in the processes and outputs of *B* (Tononi, 2004). Similarly, effective information from *A* to *B* will be low or zero if there is a minimal or non-existent causal connection between *A* and *B*. The effective information

[7] In the case of a one megapixel camera's sensor chip, imagine *A* as representing half of the sensor chip *X*. In order to calculate the entropy of *B*'s responses, *A* can be replaced with 500,000 individual photodiodes

for the system *X*, comprised of partitions *A* and *B*, is thus the sum of the effective information from *A* to *B* and from *B* to *A*, where *A* and *B* are the partitions that yield the lowest amount of effective information. This minimum amount of effective information, for a system *X*, is known as the Φ of *X*, and represents *X*'s capacity to integrate information (Tononi, 2004) (Tononi, 2008) (Tononi, 2003).

The assessment of Φ of a system *X* is as follows: if Φ of *X* is zero, then *X* is not capable of integrating information. Since there is no effective information in the bipartition of *A* and *B*, both *A* and *B* must be causally independent, non-integrated subsystems. This is evident in the case of the 1,000,000 photodiodes in the camera. Since the photodiodes are independent of each other, in that changing the states of one or more the photodiodes has no effect on the states of the other photodiodes, there is no effective information between partitions, and no information is integrated in the camera sensor (the system as a whole) (Tononi, 2004).

Recall that under Integrated Information Theory, consciousness is related to the information integration capabilities, or Φ, of a system's complexes. Basically, a complex of a system is a subset of that system with $\Phi > 0$ that is not part of another subset with a higher Φ (Tononi, 2004). The relationship between a system's complexes and the consciousness of a system is such that only complexes can integrate information. Elements of a system that are not contained within a complex cannot contribute to consciousness, even if they interact with elements in a complex. For example, while a human's stomach may interact and exchange information with a human's brain, the stomach does not contribute to a human's consciousness, since it is not part of a human's main complex (Tononi, 2004).

Surely, there are objections to Integrated Information Theory. One, which I already addressed, was the explanatory gap objection. On this objection, IIT does not sufficiently explain *why* consciousness corresponds to information integration, only that it *does* correspond with it. To reiterate, this no more of an objection to IIT as it is an objection to materialism, in general. Secondly, with such a strong reliance on the inputs and outputs of mechanisms, and the relationships between subsets and complexes of a system, it seems as though Integrated Information Theory may be susceptible to the same objections that have been presented against functionalist accounts of consciousness. Initially, it seems as though IIT is only concerned with how a system handles inputs and produces outputs. This is true, but not in the same way as it is true in traditional functional theories of consciousness. Integrated Information Theory is not concerned with the mere functional states of a system or subsystem. According to IIT, the functional role of a particular state is important, but not necessarily *because* of its functional role. True, functional states are essential to consciousness in a manner similar to functionalism, but they are perhaps more important in a different way. In IIT, the function of two states does not have to be identical for them both to be considered conscious. Instead, they only have to integrate information similarly. The functional role of a state may change between systems, but if they both integrate information to the same degree, they are both conscious.

By the definitions and key principles of Integrated Information Theory, an incredibly simple system, such as a single photodiode, could theoretically be conscious (Oizumi, Albantakis, & Tononi, 2014). However, any photodiode would not qualify for experiencing consciousness; the structure of the photodiode is essential to its

consciousness. A photodiode consisting only of a detector and output, the behavior of which is determined only by external inputs (light), is not considered conscious. The reason for the photodiode's unconsciousness is the fact that the detector's response to the external input is passed on as the system's output. The detector's response does not "come back into" the photodiode's system. Therefore, the detector and output do not have both causes and effects within the system, and thus, they do not constitute a complex and produce qualia (Oizumi, Albantakis, & Tononi, 2014).

On the other hand, a photodiode that functions in the same way, that is, produces the same output for the same input, but is structurally different *can* be conscious, although minimally so. Imagine a photodiode consisting of the same detector, but also includes some predictor, such that the detector responds to both external inputs and inputs from the predictor. As the detector receives the external output, it sends a signal to the predictor. The predictor receives input from the detector, then passes a response back to the detector such that the detector's response to the external input is causally affected by the predictor's information. When the photodiode is structured in this way, then the system is an informationally integrated complex with $\Phi > 0$, making it a minimally conscious photodiode (Oizumi, Albantakis, & Tononi, 2014). While traditional functionalism would classify both photodiodes together, thus making them both either conscious or unconscious, IIT allows for differences in the assessment of their levels of consciousness.

Now that I have outlined and examined three current theories of consciousness, as well as deep learning algorithms, we are in the position to investigate the primary goal of this paper: determining whether or not a deep learning computer is capable of being conscious. To do this, I will analyze deep learning systems through the lens of each theory

of consciousness, inferring each particular theory's position on the conscious of a computer. Then I will decide which theory is most plausible, either conceptually, empirically, or both, and the consequences of accepting one of the theories.

Chapter 4: Are Deep Learning Computers Conscious?

Based on the previous two chapters, it should be clear that answering the question of whether or not deep learning computers are conscious is not straightforward and simple. The reason for this uncertainty is due to the fact that accepting different theories of consciousness results in different conclusions about deep learning computers' potential consciousness. In order to make this inconsistency more explicit, I will infer each theory's verdict on the topic of consciousness in deep learning computers.

Applying Theories of Consciousness to Deep Learning Computers

Recall from Chapter 1, that I seemed to dismiss the idea of machines as thinking, intelligent things. I implied that deep learning computers do not learn in the same sense that people do. But is that so? Is it not the case that we are just following a set of rules, or traversing a decision tree at immense speeds and sometimes without our immediate, perceptual knowledge? Could it be that brains and computers are structured in a similar manner, in that they both function in this way, and thus, consciousness could also be shared between organic and artificially constructed things? Ultimately, the answers to these questions, as well as the principle question, "Are deep learning computers conscious?", vary depending on the theory of consciousness that one accepts; and such is the case with the theories mentioned above: physicalism with the A-consciousness and P-consciousness distinction, the Multiple Drafts Model, and the Integrated Information Theory. The general problem that these different theories aim to answer and solve is whether a computer, which has different physical structures and states from a conscious

being, but nonetheless functions in the same or similar manner, is conscious? In other words, is consciousness reliant on functionality or physicality? Let us examine the conclusions that proponents of each theory above would reach with respect to deep learning computers (for simplicity, I will first examine more elementary networks, such as the single perceptron model, then extrapolate to more complex neural network models).

When people are asked to define consciousness and state the requirements of a conscious system, a traditional and common answer is along the lines of 'awareness' or 'rationality.' And this seems to be a natural and intuitive idea. After all, people believe they are conscious and they possess both of those characteristics; it is what seems to separate us from other animals, so to speak. Dennett, more or less, agrees with this notion of consciousness. Recall that Dennett believes consciousness relies primarily on the judgments that a system makes about qualia and the behavior that the system exhibits as a result of that judgment (Dennett, 1991). To recapitulate his line of argument, qualia, in philosophy, are epiphenomenal in nature. This means, by the definition of 'epiphenomenal,' that qualia represent the effects of physical states, but do not produce effects on the physical world. As a consequence, any epiphenomenally qualitative mental state cannot, by definition, influence behavior (or consciousness). All that can be considered in the discussion of consciousness are judgments about qualia, and the functions generated by those judgments. There is no reason to believe that qualia, rather than judgments about qualia, play any significant role in consciousness (Dennett, 1991).

So how does this view translate to consciousness in deep learning computers? Well, simply, it means that deep learning computers are conscious, so long as they make accurate judgments and their states play the appropriate functional role in the system. For

Dennett, and others that are sympathetic to the Multiple Drafts Model, there is no significant difference, in the context of consciousness, between a deep learning computer and another healthy person (which is normally deemed conscious). There is no consequence for the fact that a computer represents an image using electricity and voltages, instead of chemical reactions, because qualia are not existent in either system.

In a deep learning computer, whether it is using a feed-forward or a recurrent network, the necessary core principles of the Multiple Drafts Model are present. First, a conscious computer must behave as if it is conscious. This means, in the context of some input, the computer must produce the appropriate output. For example, when a deep learning computer is given a picture of a cat as input, the behavior and judgments it makes should be of the sort that a person would have, such as stating, "That's a cat." This sort of judgment, presumably, would be equivalent in both the human and the robot due to the equal discriminatory abilities in both systems. The computer's judgment is evidence of a comparable level of power in differentiating between objects to that of a [conscious] human (Dennett, 1998).

The other feature of conscious beings is the continual availability and revisability of information. Deep learning computers satisfy this criterion as well. Remember, Dennett's notion of being able to revise information does not refer to *knowingly* or *willfully* revising it (though it could). In his illustrative phi phenomenon example, the brain revises information about the two dots in order to make sense of it, and this revision is done without the subject's knowledge. Deep learning computers are similar. In fact, continual revision is how deep neural networks work! When a computer learns with an artificial neural network, input weights are repeatedly changed and updated so that the

computer can become more accurate and productive in its task. In essence, the network is creating *multiple drafts* of representations of the data. For each different input weight in the neurons of the network, the input "looks" different, and so the overall response or output would change. In the phi phenomenon, the original draft represents two separate dots, and the second draft represents a line. As a result of the revision and multiplicity of the drafts, our judgment changes from judging that two dots were seen to judging that a line was seen. In a neural network, changing the weights, i.e. revising the drafts, changes the network's judgments about the input. It seems, then, that deep learning computers satisfy this criterion of having multiple, revisable drafts, as well.

Furthermore, information in the brain is shared and accessible between modules, and this is what leads to conscious experience. Likewise, information from one process is available to other processes in deep learning computers, so long as the connections are there to make that possible, which is a necessary component of information's availability in the brain, too. In deep learning networks, information is passed from neuron to neuron and layer to layer, much like information is shared between brain neurons and brain modules. Altogether, deep learning computers are working like conscious systems in order to act like conscious systems. And since qualia have been removed from the discussion of consciousness and replaced with judgments about qualia and functions resulting from those judgments, deep learning computers are conscious, and would have a stream of consciousness just as humans do, according to the Multiple Drafts Model.

But remember that the Multiple Drafts Model is only one type of consciousness theory. While Dennett denies that consciousness is truly anything more than functional states associated with qualitative judgments, physicalists, such as Ned Block, believe there

is something more to consciousness than simply functions. The key difference between a conscious and unconscious thing is the existence of qualia, or in Block's terms, phenomenal consciousness.

Physicalist accounts of consciousness, in general, state that the physical states of a system *are*, rather than are correlated with, conscious states (Stoljar, 2015). This means that a system is only considered conscious when the physical state of the system is identical to the physical state that another conscious being is in. Or in other words, every conscious state that a system is in at time *t* is identical to and explainable by the physical state that system is in at time *t*. Some physicalists, like Ned Block, believe these physical states, and therefore conscious states, include qualitative, phenomenal states.

Ned Block's physicalist account of consciousness, as explained above, asserts and relies on two principles of consciousness. The first is that conscious states, the states that make some thing conscious, are purely the physical states of that thing. The second characteristic of consciousness that Block presents is that there exists a distinction between two forms of consciousness: access consciousness and phenomenal consciousness, which both comprise consciousness as it is commonly thought of. As a result, Block, and other physicalists that share his views, would maintain that computers are not conscious in the way that other systems, like humans, are. The key component of consciousness that computers do not have, according to Block, is P-consciousness, or phenomenal experience. Block asserts that computers do not have the important part of consciousness, which is qualia (Block, 2002). When a deep learning computer performs some task, it receives input, processes it, and produces some output. This, in principle, is the same as how humans function. However, Block would argue that the type of input, as

well as the realization of the respective systems, is the dividing factor when it comes to consciousness.

For Block, when a person, or even another intelligent animal, produces "output" from "input," it is only the same as a computer producing "output" from "input" in a linguistic sense; the words used to describe the activity are the same, but the actual process is entirely different. The input into a human's processing system, in Block's view, consists of, and is understood in terms of, qualia, or P-conscious states. This means, when a person produces output, such as speaking, behaving, or having thoughts, it is the result of his phenomenal experience. The fact that the input is experience means there is something it is like to process that information; something is experienced. For example, when a person sees a color, such as red, there is information about the redness (among other properties) that enters the person's mind. When the brain processes that information, the result is the experience of seeing red, and thus, there is something it is like to see the color red, as opposed to seeing orange. A deep learning computer, on the other hand, has no such experience. When it receives its input, which contains no qualitative properties, but instead, mere combinations of electric currents, there is no mental state, or P-consciousness, of the computer. Sure, both the computer and the person have access consciousness, since both are capable of having and being in states of available information, but this alone is not enough to claim that both are conscious as a whole, according to Block. The access conscious states in the person seeing red are accessing *perceptual*, *experiential* information (Block, 2002), whereas the deep learning computer is simply accessing series of 1's and 0's.

Note that Block does not deny that computers are access conscious. He agrees on the idea that information being available to the processes of a deep learning computer means that they have access consciousness. When information is processed in a deep neural network, the information is representational and is used in the system's logical, information-processing operations. In the case of feed-forward convolutional networks, the information is passed between layers in the network. But while the information is representational of the input, it is not necessarily fully accessible; it is only accessible to the system after passing through the final output layer. But this is no different than information in a human brain; in fact, it is much the same. We would not think to call the low-level abstractions of seeing a chair that are produced by individual neurons to be A-conscious states, so there is no reason to do so in the convolutional network. The final output of the network, however, *is* an A-conscious state, just as the entire visual experience of seeing a chair is an A-conscious state. The same could be said for deep recurrent networks, such as deep belief networks. Therefore, deep learning computers are access conscious, which *is* a form of consciousness, after all (Block, 2002). But once again, access consciousness does not suffice for "complete" consciousness.

As described in Chapter 3, there are scenarios in which A-consciousness exists without P-consciousness and P-consciousness exists without A-consciousness. While P-consciousness without A-consciousness may be relevant in the discussion of consciousness as a whole, it does not apply to deep learning computers, as it is rare that the information in a program is inaccessible. A-consciousness without P-consciousness, then, is the scenario of interest, in the context of deep learning computers. Recall the example of this situation that Block presents: superblindsight. In the case of

superblindsight, the patient is able to guess what is in the blind spot by his own accord, rather than needing to be prompted to do so by another person, as is the case in normal blindsight. In regular blindsight, the fact that the patient has to be asked to guess means there is no A-conscious state; there is nothing available to rational processes. Conversely, the superblindsighter will suddenly have the knowledge that there is something in his blind spot, and at that point would make the accurate guess about what is in front of him.

Of course, this example of superblindsight may initially strike us as difficult to imagine, but that is precisely its purpose. Block wants us to conclude that the inability, or at least, difficulty, in conceptualizing what it is like to be a superblindsight patient is evidence that the patient has no P-consciousness. That is, the fact that it seems odd that the superblindsight patient is able to "just know" that they are missing something and should guess supports the idea that there is no experience, there is no way it is like to be a superblindsight patient. For Block, a computer is extremely similar, if not equivalent, to the superblindsight patient. The computer's realization is like the patient's visual system, in that it does not support phenomenal qualia that P-consciousness requires. The computer experiences nothing when it receives input that will later be accessible information, just as the superblindsight patient experiences nothing in his blind spot, even though he will later access that information. The deep learning computer, as well as all computers, just simply do not experience phenomena, and therefore, lack the important form of consciousness: phenomenal consciousness.

On the other hand, some people, like Tononi, believe that views like Block put too much stock in qualia, and that the existence of qualia is neither essential nor irrelevant to consciousness. In IIT, qualia is only important insofar as it is information. Integrated

Information Theory is only concerned with the amount of information in the system, and the level of integration of that information. For this reason, a proponent of IIT would come to a rather unique conclusion about deep learning computers. To put it simply, Tononi would say that computers *are* conscious, but not all computers are equally conscious. As the structure and design of the computer changes, the degree of information integration also changes, and thus, consciousness occurs in varying quantities. Note that IIT differs from the other two theories in that the others are views of consciousness that are binary: some thing is either [fully] conscious or not [fully] conscious. On Tononi's view, two separate computers, specifically, deep learning computers, can have different levels of consciousness. The reason for the differences in consciousness is the dissimilarity in structure, and thus, a variance in information integration.

In order to illustrate this point, return to the example of the simple, two-part photodiode from the previous chapter. This photodiode consists of two parts: a detector and a predictor. Recall that this photodiode, as simple as it is, relative to many other systems, it is actually conscious, though it is decidedly minimally conscious. According to IIT, the photodiode exhibits the necessary traits needed for consciousness. The internal states represent information and are maximally irreducible, meaning there are no partitions of the photodiode system that are capable of processing the same information (Oizumi, Albantakis, & Tononi, 2014). However, the level of consciousness in the photodiode is immensely low, compared to a human, and for this reason, it cannot be said that the photodiode's conscious experience is of much interest; rather than stating that the photodiode is experiencing "light," as it might seem to somebody observing it, it is more

accurate to state that the photodiode is merely experiencing this state *A* rather than that state *B* (Oizumi, Albantakis, & Tononi, 2014).

Now, return to the original photodiode comprised of only the detector and no predictor. This photodiode, while capable of the same function as the previous photodiode, is not conscious at all. The information in the system, that is, the state of the detector, is equal in both examples, but the one-part photodiode has no integration aspect. Since the information related to the detector only causes the output of the photodiode, rather than causing *and* being affected by it, the system does not qualify as a complex, and thus, has no information integration. The key difference between the two photodiodes is *feed-back* within the system (Oizumi, Albantakis, & Tononi, 2014). Additionally, the fact that these two photodiodes have different levels of consciousness rejects behaviorists' claim that consciousness is solely determined by the outputs produced from some inputs.

One may question whether or not feed-back is truly a necessary component of consciousness, and that concern is legitimate. I do not want to attend to this issue too greatly, so I will simply offer a few preliminary responses to the objection that feed-back is not important for consciousness. The first line of support comes from empirical findings. Feed-back and the reentry of information is present in human brains, and when people undergo anesthesia, certain neural systems and parts that utilize feed-back and reentry are not active (Oizumi, Albantakis, & Tononi, 2014) (Imas, Ropella, Ward, Wood, & Hudetz, 2005) (Boly, et al., 2012). And the reason that our brains have developed in this way can be argued on evolutionary, or more theoretical, grounds. From an evolutionary standpoint, a system that can accomplish a task by using an architecture with few parts that reenter information is more economical than one that uses many parts once.

Thus, a brain with neurons capable of feed-back has a greater cognitive capacity when compared to a brain with the same number of neurons, but is not integrated via informational feed-back. This principle is not evident in the photodiode example, as each is capable of the same function, despite one using more parts. But when more complex tasks are examined, a larger system is needed when there is no feed-back. Generally, this is one reason why a convolutional network may contain 6 or 7 or 8 layers of neurons, while a deep belief network used for the same task may have only 4 or 5. Additionally, a system without any sort of informational reentry is concerned solely and entirely with external input, whereas as a highly integrated system with feed-back is affected by internal states as well as external input, and so it is more likely to be autonomous (it will react specifically to the internal states) (Oizumi, Albantakis, & Tononi, 2014).

When we expand the systems in the discussion to include deep neural networks, many similarities to the photodiode example become apparent. From the perspective of an Integrated Information Theorist, convolutional neural networks are fundamentally the same as the one-part photodiode without the predictor, while recurrent networks, such as deep belief networks, are fundamentally equivalent to the two-part photodiode with the predictor.

When a convolutional neural network processes input, the input weights exceed a neuron's threshold and trigger the function of the neuron. After the neuron completes its function, its output is given to another neuron in another layer. This means that the output from a layer of neurons in a convolutional network does not necessarily have an effect on the outputs of another layer in a way that is unique to that neuron; the same effect on the overall output could be the result of significantly different weights in another neuron. In

other words, as far as the function of a neuron is concerned, there is no difference between receiving input *i* with weight *w* from an external source and receiving that same input *i* with weight *w* from another neuron in the network. The function of each neuron is independent of the functions of other neurons, in the sense that it will process inputs and weights in the same manner, regardless of whether or not they are external or internal. Therefore, there is no integration of the information. When there is no integration of information, then the network does not form a complex and has no consciousness. Any maximally reducible partition of a convolutional network will be capable of processing the same information as another partition. If a partition, with its current input weights, is not capable of processing information (regardless of the degree of integration) to the same extent as another partition, it will have a different Φ value, as well as lower representational and distinguishable power than another partition. If, for example, a convolutional network were split into two halves "length-wise," such that each layer of the network consisted of half of the neurons it previously did, the representational power would be half of what it was before, regardless of the fact that integration of the system would be the same (the causal relationships between neurons would remain intact). This imperfection is one of the reasons why convolutional networks are preferable over larger, more traditional feed-forward networks. Since neurons in convolutional neural networks share weights and their outputs are pooled between layers, convolutional networks are able to complete the same tasks as larger, more unsophisticated feed-forward neural can, but with few connections between neurons (Krizhevsky, Sutskever, & Hinton, 2012) (Neubauer, 1998). But even so, the neurons in feed-forward networks, in general, are not integrated in the way necessary for consciousness.

On the other hand, a deep recurrent network, such as a deep belief network, will be more conscious than a convolutional network. Just as the photodiode systems above are dissimilar, a convolutional network and a deep belief network differ in the presence of informational feed-back within the system's parts. When a deep belief network processes information, feed-back and the reentry of information *is* present. When deep belief networks, as well as other deep recurrent networks, process information, the input enters the first layer. In DBNs, this first layer is typically a Restricted Boltzmann Machine. Normally, these RBMs consist of a visible layer of neurons and a hidden layer of neurons. The input into the DBN initially enters one of the RBM's layers, and the weights are determined (i.e. the RBM learns on the data) by updating in relation to the states of the neurons in the other layer (Hinton, Osindero, & Teh, 2006). This means that the current internal states of the neurons in the RBM's layers have a cause-and-effect relationship with the neurons in the other neural layer. This relationship is analogous to the relationship between the detector and the predictor in the conscious photodiode; the network will not function properly or similarly if one of the layers is removed from the system. Thus, the RBM forms a complex, since the units that comprise it are integrated together and are dependent on each other. When a subsystem is removed, such as the hidden layer of neurons, it is equivalent to removing the predictor in the photodiode, and both the functionality and the information integration are lost. In terms of feed-back, the data in the DBN is reentered into each layer. That is, the neurons in both the hidden and visible neural layers of the RBM are not simply dealing with new external information each time, but instead, they are "looking at" the same information with regard to the outputs and states of other neurons in another layer. This evidence of integration means

that deep belief networks that use Restricted Boltzmann Machines, are conscious, at least to some extent. There is no general conclusion that one can make about the level of consciousness in a deep belief network, just as there is no single amount of consciousness experienced in a human[8]. It all depends on the particular deep learning computer being considered.

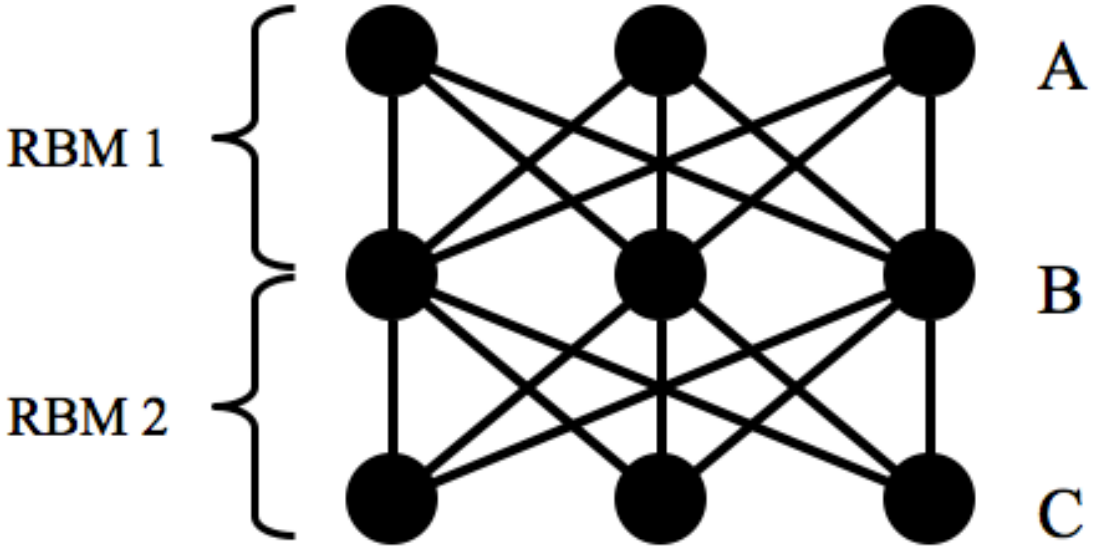
As was the case with the photodiode, the amount of information in each individual neuron is minimally low. Since the units in the DBNs are stochastic and binary, their states are of the kind such that they are either in state *A* or state *not-A*. But as the number of neurons increases, the representational power also increases. For example, a layer of neurons, such as one in a CNN, is capable of representing an image, rather than only a pixel, like a single neuron can. Similarly, a human brain's neurons are binary, in that they are in a state of being fired, or not being fired. But as the number of neurons increases, and continue to be integrated, they form a maximally reducible complex that contains a great number of informational bits (imagine the human visual system). Likewise, if the DBN (or RBM) is large enough, in terms of the number of integrated neurons, and its representational power is great enough, then the amount of information will be large, as well. Overall, the amount of information in the deep learning computer is contingent on the particular structure of the deep belief network being used.

In terms of the integration of the deep belief network, the degree is again determined by the architecture. For example, if the deep belief network were designed in such a way that the first two layers were layers in a Restricted Boltzmann Machine, but the subsequent layers were simply feed-forward, then the overall network would only

[8] Intuitively, it seems as though a newborn baby is not as conscious as a 25-year-old healthy adult, but this claim may be controversial.

have the same level of consciousness, or Φ, as the initial RBM, which would not be much, relative to humans and the like. However, if the deep belief network consisted of a Restricted Boltzmann Machine that fed into another Restricted Boltzmann Machine, such that the hidden layer in the first served as the visible layer in the second, then the two RBMs would be integrated together, forming a larger maximally reducible complex. While there is no difference in the number of bits of information between these two networks, the level of integration is higher in the network consisting of two RBMs, rather than one. This is due to the fact that the amount of [effective] information between two partitions, or subsystems, of the network is greater. Imagine there are three neural layers, *A*, *B*, and *C*, in the network, such that layers *A* and *B* form the first RBM, and layers *B* and *C* form the second RBM, as shown in Figure 8.

Figure 8: An example of connected and integrated Restricted Boltzmann Machines. Notice that the first RBM consists of layers A and B, while the second RBM consists of layers B and C. This integrates the two RBMs.



When the entire network is learning and the states of the neurons are updating, the two RBMs interact. The neurons in layer *A* update in reaction to the current states of the neurons in layer *B*, which update with respect to the states of the neurons in layer *C*. Under this design, the network cannot be divided in such a way that one of the layers, *A*, *B*, or *C*, is removed (making it equivalent to a single RBM) and the amount of effective information is not decreased to some extent. Thus, the second, two-RBM deep belief network is more integrated than the one-RBM deep belief network, and so the amount of consciousness in that network is greater.

The general rule for Integrated Information Theory's determination of consciousness is that of feed-back. If a system reenters information, and current internal states interact in some way with other internal states, rather than only with external input, then that system will experience consciousness. In deep learning computers, convolutional neural networks, as well as other strictly feed-forward networks, are not conscious, while recurrent networks, such as deep belief networks, *are* conscious, though to varying degrees.

Where Do We Go From Here?

In the previous section of this chapter, I examined two types of deep neural networks through the scopes of three different views about consciousness, and inferred the conclusions each theory would come to, and why. Unfortunately, but perhaps unsurprisingly, the theories do not fully agree on the matter. The Multiple Drafts Model states that deep learning computers are conscious, while Block's access consciousness and phenomenal consciousness distinction rules out the existence of the significant form of

consciousness, P-consciousness, in deep learning computers. And to complicate things further, Tononi's Integrated Information Theory states that some deep learning computers are not conscious, while others are, yet among those that are, the level of consciousness fluctuates. In some ways, one could argue, they *do* agree, since all theories attribute *some* consciousness to *some* deep learning computers, but it should be clear that this is hardly satisfying. In order to truly advance the topic of conscious computers towards some type of definitive answer, or at least, consensus, about whether or not deep learning computers are conscious, one has two options: engage in philosophical debate, or engage in empirical research about either the actual, necessary properties of consciousness, or a computer realization that satisfies all theories of consciousness.

Naturally, both of these routes have been taken in the past, and many people are still taking them today. Unfortunately, empirical study, definitive though it may be, has two drawbacks. One, which may seem trivial, is that it takes a long time. Neuroscientists and biologists are working towards finding the answers to questions about consciousness, but current technology can only accomplish so much, and the same can be said for neural network research. Philosophical debate, on the other hand, is able to step in where experimentation falls short. Contemplating theoretical principles is useful, but ultimately gains us no ground with regard to the consciousness of deep learning computers, since some principles simply cannot be tested, such as the phenomenal consciousness of a computer. This illustrates the second disadvantage of empiricism: some tests are extremely difficult. It is not clear exactly *how* to test for consciousness, primarily because it is unclear exactly what should be tested for. The existence of conflicting theories of consciousness makes it practically impossible to devise a test or method that is

unquestionably used. However, philosophy is able to help, once again. By utilizing argumentation for one theory over another, we can make small steps towards general acceptance on the nature of consciousness. In lieu of science being able to rule out certain theories with research, one can rule out certain theories based on other support, thus leading to the acceptance of just one.

In the context of the three theories of consciousness considered above, how might we go about the task of eliminating variation between conclusions? Empirically, we could institute a "consciousness science." That is, we can make empirical findings about consciousness, shedding light on fundamental and necessary properties of consciousness. This scientific approach has already begun and interesting findings, such as the role of awareness in consciousness, have allowed philosophers and scientists alike to rule out certain theories and properties (Block, et al., 2014). But while that may be generally helpful and eventually successful in coming to a single verdict about the consciousness of deep learning computers, it is not yet conclusive enough to be solely relied on. Analogously, people have not, and do not, wait until all questions regarding physics have been answered before applying theories and principles. In fact, they do the opposite. Historically, both the general public and scientists have accepted a notion, such as Newtonian physics, and revised their views when new findings suggested their current ideas were incorrect.

The other empirical route involves redesigning neural networks such that they are deemed either conscious or unconscious by all consciousness theorists. But this seems like the wrong option to take at this point in time. Just as is the case with consciousness science, neural network research takes time, and it should not be the case that we hold no

opinion until such time that a network architecture of that type is developed. Moreover, due to the fundamental differences between the views held be theorists like Block and Dennett, it is not clear that such a network *can* be designed. It seems, then, that when faced with multiple theories of consciousness, we should make an argument for one of the theories, particularly Integrated Information Theory, and use that theory as the basis of our conclusions and research. For the purpose of this paper, I do not wish to claim that the argument will be sufficient for accepting IIT over *all* other theories of consciousness. Instead, it should suffice to show that IIT is preferable to the other two theories, thus eliminating them from contention, and progressing us closer to a generally accepted theory.

A Case For Integrated Information Theory

When it comes to choosing one of the three theories of consciousness that have been examined in this paper, there are a couple of areas of interest to consider: the "hard problem of consciousness," and the objections from liberalism and chauvinism. Specifically, Dennett's Multiple Drafts Model and Block's physicalist theory of consciousness seem unattractive when viewed through these lenses. However, Integrated Information Theory is able to avoid the shortcomings that the other theories have, making it a more attractive, advantageous, and beneficial theory of consciousness. In order to illustrate IIT attractiveness, I will explain the problems that Dennett and Block face, and explain how IIT does not fall victim to the same objections.

As it was famously proposed by David Chalmers, the "hard problem of consciousness" is that of explaining subjective, phenomenal experience (Chalmers, 1995).

That is, why do qualitative, perceptual experiences exist, and how do we explain their existence? One of the criticisms of certain functional theories, like the Multiple Drafts Model, is that they are not concerned with this hard problem of consciousness, which leaves questions of our perceptions unanswered. In denying the existence of qualia as real, rather than effects of functions that only exist in our minds, Dennett avoids, not answers, the hard problem of consciousness. In fact, he believes that there is no question to answer; it is not possible to explain *why* qualitative experiences exist because they themselves do not exist (since qualia do not exist). But this response seems to be unsatisfying to many, primarily because it appears as though the hard problem is a real problem. To many people, qualia are real, as evidenced by our qualitative experiences, and so the hard problem is also real. To simply state that qualia are non-existent on the basis that they are epiphenomenal in nature (which is not an idea that every philosopher agrees with) is unattractive and puzzling. And even under the view that qualia are not the causes of our experiences, MDM does not explain why it *seems* we have experiences, which is all he claims is true.

Integrated Information Theory, however, is not a guilty of avoiding the hard problem in the way that the Multiple Drafts Model is. While IIT is a type of functionalist theory of consciousness, it is still concerned with the hard problem of consciousness, since it agrees that both qualia and experience exist. An Integrated Information Theorist would argue that qualia exist, but they are not special. Instead, qualia, and phenomenal properties, should merely be considered information. Under IIT, when one experiences a red chair, it is not the case that one is only experiencing the appearance of red (as MDM would claim). Instead, it is the case that the experience of red is one possible outcome,

with others being all other colors, i.e. green, blue, yellow, etc., that one is capable of distinguishing between. This then, means that experiencing the color red carries many bits of information, since the number of alternative possible outcomes is high. Additionally, IIT posits that perceptual experience *is* important in consciousness, insofar as a highly integrated system will have experiences. As information integration increases, autonomy increases (as the result of informational feed-back and reentry). Thus, a conscious system will have *some* sort of experience. A minimally conscious photodiode only experiences "this state rather than not this state," while a far more conscious system, like a human, will experience "light" (the quality of experience is related to the representational capabilities of the system). Therefore, Integrated Information Theory does address the hard problem of consciousness by providing an explanation for our subjective experiences: qualia is information and feed-back yields autonomy, or in other words, subjectivity. And IIT does this without removing qualia from the discussion like MDM does. This makes IIT a better functionalist theory than the Multiple Drafts Model.

But what about Block's physicalist view? How does Tononi compete with Block? Surely, it is not possible to reject Block's theory for the same eliminativist reason as MDM was rejected, since it is the case that Block's theory also addresses the hard problem. Unlike Dennett, Block does not dismiss qualia; in fact, qualia, and our ability perceive them, is what makes us conscious. Recall, P-consciousness, or perceptual experience, is significant in whether or not a being is conscious, according to Block. This appeal to phenomenal consciousness is satisfying in similar manner to how MDM is unsatisfying. The fact that we *are* phenomenally conscious makes it appear that phenomenal experience should be necessary for substantive consciousness. However, accepting that notion has

undesirable side-effects. What would a view like Block's, which claims that a necessary property of consciousness is P-consciousness, claim about the consciousness of a person in a deep sleep?

Imagine being in a truly deep sleep, in which you have no dreams. What is that state of sleep like? How would you describe it? You may find that to be a difficult task. When in dreamless sleep, it seems as if it never happened. You go to sleep and then you wake up hours later, unable to express, or even recall, what those hours of sleep felt like. This is similar to the superblindsight and blindsight scenarios mentioned earlier. The common feature between these cases is the difficulty in imagining what it is like to be in their respective states. And I think Block would have to agree that this lack of experience is due to a lack of P-consciousness and P-conscious states. Also, I think he would have to agree that a person in deep, dreamless sleep, is not phenomenally conscious[9]. But this would mean that a sleeper is not conscious, in the important and robust sense. While some people may be fine with this assertion, others, myself included, find it peculiar. On the one hand, deeming a sleeping person unconscious has ethical ramifications, such as being no more possessive of basic human rights than a rock or chair. But perhaps more interestingly, this idea of sleep-induced unconsciousness yields more questions. One could ask how it is that we regain consciousness when we wake up, or how we lose it in the first place when we go to sleep? What happens in sleep such that consciousness, which is physically realized, is removed and reinstated? These questions make Block's view more complex and complicated than is desirable. While Block's physicalist, P-conscious view of consciousness is subject to these difficult questions, IIT is not.

[9] If he disagrees that a sleeping person is not P-conscious, I would be curious to hear his description of what experiencing sleep is like.

Integrated Information Theory would agree that it is difficult to describe the experience of deep sleep. However, it would not conclude that a sleeping person is not conscious. Instead, one may just decline in one's level of consciousness when one enters a dreamless sleep. Since consciousness is determined by information integration, all that is necessary for a person, awake or asleep, to retain consciousness is to have integrated information in his brain. While the person may not have perceptual experience inputs and information, information most likely still exist in his brain. Brain scans have shown that there is brain activity even when one is asleep. This is evidence that the brain is functioning in some manner to some stimuli. Knowing how the brain functions when we are awake, we can infer that the brain is equally as integrated when asleep, since falling asleep does not change the structure of our brains or the methods by which they operate. And while we may not be able to determine the types of information that the brain is processing, we can know that it is processing *something* when we are asleep. This means that information is being integrated, and importantly, it is being integrated in a similar or equivalent manner to when we are not sleeping. Therefore, under Integrated Information Theory, sleeping people are still conscious, though perhaps at much lower levels then people that are awake. Furthermore, unlike Block's view, which generates new questions, one can explain why and how consciousness is increased and decreased with the entrance and exit of sleep. This, I believe, makes IIT more appealing than Block's requirement of phenomenal consciousness.

This dreamless sleep example is just one illustration of a larger problem that Block encounters, which is chauvinism. As I explained in Chapter 2, the chauvinist objection essentially states that physicalism does not attribute mental states, such as consciousness,

to beings that *do* have mental states, and consciousness. Block's view does not allow for a person in deep sleep to be considered robustly conscious, despite the fact that the sleeping person's brain is in the same physical state as when he is awake (and conscious). IIT, as I just explained, does not restrict which things are conscious and which things are not in this way. But simultaneously, it is not a liberal theory, either. While IIT may initially appear to be subject to the liberalism objection, which states that it assigns consciousness to unconscious systems, I find that it is not, or at least, it is not as poorly liberal. That is, while behaviorism is a liberal theory that would assign consciousness to both photodiodes (assuming it was agreed that the second one is conscious), IIT does not. It may be arguably liberal in assigning consciousness to the second, more complex photodiode, but it does so in a way that allows it to discriminate against other unconscious systems.

Imagine a zombie with no brain and no mental states, but that acts and functions exactly like a regular, conscious human. Despite the almost undeniable fact that the zombie is not conscious, since it does not have a brain, a behaviorist would label the zombie as conscious, since its behavior is the same as a conscious human. However, IIT does not fall into that trap. On the basis of information integration, IIT is able to deny that the zombie is conscious in any way. So while it can be said that IIT is a liberal theory of consciousness, it should be noted that its liberalism is a side-effect of its ability to justifiably discriminate against unconscious systems without being chauvinist, like a physicalist theory of consciousness is.

When contrasted with both the Multiple Drafts Model and phenomenal conscious physicalism, Integrated Information Theory is more attractive and versatile, at least in the context of the issues mentioned above. Again, it was not my intention to make an

exhaustive, definitive argument for IIT, but instead, show that it is the most preferable and appealing theory out of the three. In light of accepting IIT over MDM and Block's physicalism, we can revisit the question of redesigning deep neural networks in order to maximize computer consciousness. I understand that the

Empirical Advancements Revisited

As I mentioned briefly above, a deep neural network with more recurrent parts is a more conscious network. In my previous example, I explained how a deep belief network with two Restricted Boltzmann Machines, rather than one, is more integrated, thus yielding more consciousness. While this two-RBM deep belief network may be more conscious that a one-RBM network, it is still nowhere near the level of consciousness experienced in humans, or even less intelligent animals, for that matter. So how can we increase the amount of consciousness experienced by a deep learning computer? The simple answer is: add more recursive layers. By adding integrated neurons to the network, we are increasing the size of the main, maximally reducible complex of the system. As a result, the system is representationally more powerful, which means a particular state or representation contains more bits of information. At some time t_0, each neuron in the network will be in some internal state, namely, its functional threshold is either exceeded or not. If the input was modified ever so slightly, such that the state of one neuron in the initial input layer at time t_1 was also modified, the overall representation at time t_1 would be different than at time t_0, and so, the input at t_1 is an alternative outcome relative to the input at t_0. It is evident, then, that the amount of information in the network increases as

the number of neurons increases, since more subtle and minute alterations in the input will be represented, thus constituting an alternative possible outcome.

But simply adding neurons does not necessarily increase consciousness in deep learning computers. Adding neurons to the layers of a convolutional network does not make the network conscious, since those neurons are still not integrated at all. So in order to increase the consciousness of deep learning computers, the integration needs to increase with the amount of information. One method of doing so is to design deep neural networks with many recursive units and parts, such as Restricted Boltzmann Machines. In principle, the most integrated deep neural network would consist of a series of RBMs, situated such that the hidden layer of each RBM acted as the visible layer in the next RBM, much like an extension of the RBM model in Figure 8. In this type of deep belief network, there would be no partition that would constitute a complex with a Φ level as great as the entire network. Unfortunately, this would lead require immense amounts of computing and processing power. Researchers have applied this notion of using many RBMs, but even then, they were forced to use graphics processing units (GPUs), as only those were able to provide the processing power necessary to handle a four layer deep belief network (Raina, Madhavan, & Ng, 2009). In order to increase consciousness in deep belief networks to a level that is comparable to a human, an almost unimaginable amount of processing power is necessary. However, if the neural network were structured such that the hidden layer of every fourth RBM was used as the visible layer in another series of four RBMs, for example, then the integration would be higher than in the case of a purely feed-forward network, yet the requirements of the computer's hardware would be lesser, though still tremendous. In fact, this is closer to the way in which human brains are

structured. While the brain is integrated, overall, it is not the case that every neuron is connected to every neuron in a chain-like manner. That would be far too demanding on the system. Instead, there are smaller, highly integrated subsystems, such as those used for visual processing, toe-movement, etc., which are then integrated *as a complex* into the system as a whole. This makes information processing more parallel, and therefore, more economic. If a deep belief network were to model this structure, without sacrificing overall integration, it would lead to a more conscious deep learning computer, with more representational power, more feed-back and autonomy, a more rich experience, and more consciousness.

Chapter 5: Conclusion

Due to the fact that highly sophisticated computers and robots have been developed and utilized only relatively recently, many people are not familiar with the concerns that accompany the creation and operation of such computers. One of these concerns is certainly their consciousness. One's conclusion about the consciousness of a computer can have many effects, such as the ethics of using and turning off a computer. When designing computers and artificially intelligent machines, then, the issue of computer consciousness should earnestly be considered. However, due to the many different theories of consciousness and the many different types of computers in existence, the size of that task is great and frankly, intimidating. But perhaps the best way to begin is to examine one of the most advanced sorts of computers, and one that owes its inspiration to the most advanced animal: deep learning computers.

After examining how deep learning computers and two common types of deep neural networks, convolutional networks and deep belief networks, are structured and function, I explained three theories of consciousness that encapsulate common forms of materialist theories more broadly: Ned Block's qualitative physicalism, Daniel Dennett's Multiple Drafts Model, and Giulio Tononi's Integrated Information Theory. By extracting each of the three theory's judgments about the consciousness of deep learning computers using both convolutional neural networks and deep belief networks, I illustrated the fact that there is no one established opinion about whether or not these computers are conscious. Commonly, and perhaps as the result of the freshness of deep learning computers, ordinary people, as well as academics, are inclined to believe that computers

are not conscious. But simply because the idea of unconscious computers is traditional does not mean that the idea is true. Even though all three theories of consciousness discussed above were developed in order to explain consciousness in humans, or at most, in animals, deep learning computers can be examined in their lights. And when this examination process is done, a rather fascinating incident occurs.

Block, and other qualitative physicalists, assert that deep learning computers, regardless of the architecture of their networks, are not conscious in the way that humans are, since computers are not phenomenally conscious. However, Block's theory would posit that deep learning computers are access conscious, which, admittedly, is not a particularly robust form of consciousness, in terms of the quality of conscious experience, but is a form of consciousness, nonetheless. Under acceptance of Dennett's Multiple Drafts Model, we can see that deep learning computers are conscious much in the same manner as humans. The network structures allow for information to be processed in the way the human brain processes information, and thus, deep learning computers are functionally similar to humans, so they are conscious. And lastly, Tononi's Integrated Information Theory asserts that a system with no integration, such as a feed-forward convolutional neural network, experiences no consciousness, while a recurrent network with feed-back and integration does experience consciousness at different levels. The intriguing feature of these conclusions is that they all deem deep learning computers to be conscious in some sense or another. The individual conclusions differ, yet these three fundamentally contrasting theories do agree to some extent on the presence of consciousness in deep learning computers, which is a notion that is surprising and jarring to many people. And while Tononi's theory may be preferable to Block's and Dennett's

respective accounts of consciousness for its ability to address difficult problems related to the consciousness of other systems, it does not necessarily need to be the accepted theory of consciousness. However, regardless of the theory one is attracted to, one should be aware of the implications and extensions of that theory on systems that it did not originally attempt to explain, for the results may be unexpected.

And based on this conclusion that deep learning computers *are*, by some definition, consciousness, derived from interpretations of these theories, there are two paths we can choose between. The first would be to stay true to the traditional belief that computers are unconscious, and call for modifications to be made to these theories of consciousness so that they no longer judge deep learning computers to be conscious. But this seems incorrect to me, since these alterations would be made for purely reactionary and conservational reasons, rather than as the result of logical or metaphysical flaws. This would be equivalent to forever rejecting the notion of heliocentricity and remaining believers of geocentric models of the universe simply because we used to be heliocentricity-believers. And as there was no immediate acceptance of heliocentricity, perhaps there will be no immediate acceptance of computer consciousness. But we must be open-minded, and take the second path, which is to understand the consequences of current theories of consciousness, like functionalism and physicalism, and progress from that point. As technology advances and deep learning computers become more common and sophisticated, the manner in which we regard computers is likely to change. I am not claiming that we should blindly accept that deep learning computers are definitely conscious, or will necessarily become conscious, I am simply stating that we should adapt

and change our viewpoints to accommodate the ramifications that our current beliefs, theories of consciousness, and technologies produce.

References

Arel, I., Rose, D. C., & Kanowski, T. P. (2010). Deep Machine Learning: A New Frontier in Artificial Intelligence Research. *IEEE Computational Intelligence Magazine* .

Barber, D. (2012). *Bayesian Reasoning and Machine Learning.* Cambridge University Press.

Bengio, Y. (2009). Learning Deep Architectures in AI. *Foundations and Trends in Machine Learning* .

Block, N. (2002). Concepts of Consciousness. In D. Chalmers, *Philosophy of Mind: Classical and Contemporary Readings.* Oxford University Press.

Block, N. (2002). Troubles with Functionalism. In D. Chalmers, *Philosophy of Mind: Classical and Contemporary Readings.* Oxford University Press.

Block, N. (1981). What is Functionalism?

Block, N. (1996). What is Functionalism? In *The Encyclopedia of Philosophy Supplement.* Macmillan.

Block, N., Carmel, D., Fleming, S. M., Kentridge, R. W., Koch, C., Lamme, V. A., et al. (2014). Consciousness science: real progress and lingering misconceptions. *Trends in Cognitive Sciences , 18* (11), 556-557.

Boly, M., Moran, R., Murphy, M., Boveroux, P., Bruno, M., Noirhomme, Q., et al. (2012). Connectivity changes underlying spectral EEG changes during propofol-induced loss of consciousness. *The Journal of Neuroscience , 32*, 7082-7090.

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* .

Deng, L., & Yu, D. (2014). *Deep Learning: Methods and Application.* Now Publishers, Inc.

Dennett, D. (1991). *Consciousness Explained.* Little, Brown and Co.

Dennett, D. (1998). Instead of Qualia. In D. Dennett, *Brainchildren.* The MIT Press.

Domingos, P. (2012). A Few Useful Things To Know About Machine Learning. *Communications of the ACM* .

Hinton, G. E. (2010). A Practical Guide to Training Restricted Boltzmann Machines. *UTML TR 2010-003* .

Hinton, G. E., Osindero, S., & Teh, Y. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* .

Imas, O. A., Ropella, K. M., Ward, B. D., Wood, J. D., & Hudetz, A. G. (2005). Volatile anesthetics disrupt frontal-posterior recurrent information transfer at gamma frequencies in rat. *Neuroscience Letters , 387*, 145-150.

James, W. (1890). The Principles of Psychology.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25* .

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., et al. (2013). Building High-Level Features Using Large Scale Unsupervised Learning. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.*

Levin, J. (2013). *Functionalism*. (E. N. Zalta, Editor) Retrieved 2015 from The Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/archives/fall2013/entries/functionalism/

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Nagel, T. (2002). What Is It Like to Be a Bat? In D. Chalmers, *Philosophy of Mind: Classical and Contemporary Readings.* Oxford University Press.

Neubauer, C. (1998). Evaluation of convolutional neural networks for visual recognition. *Neural Networks, IEEE Transactions on , 9* (4), 685-696.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology* .

Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale Deep Unsupervised Learning using Graphics Processors. *Proceedings of the 26th International Conference on Machine Learning.*

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd Edition ed.). Prentice Hall.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences , 3*, 417-457.

Smithies, D. (n.d.). Attention is Rational-Access Consciousness.

Stoljar, D. (2015). *Physicalism*. (E. N. Zalta, Editor) Retrieved 2015 from The Stanford Encyclopedia of Philosophy: http://plato.stanford.edu/archives/spr2015/entries/physicalism/

Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience* .

Tononi, G. (2008). Consciousness as Integrated Information: a Provisional Manifesto. *The Biological Brain* .

Tononi, G. (2003). Measuring Information Integration. *BMC Neuroscience* .

Tononi, G., & Edelman, G. M. (1998). Consciousness and Complexity. *Science* .

Young, S. S., Scott, P. D., & Nasrabadi, N. M. (1997). Object Recognition Using Multiplayer Hopfield Neural Network. *IEEE Transactions on Image Processing* .