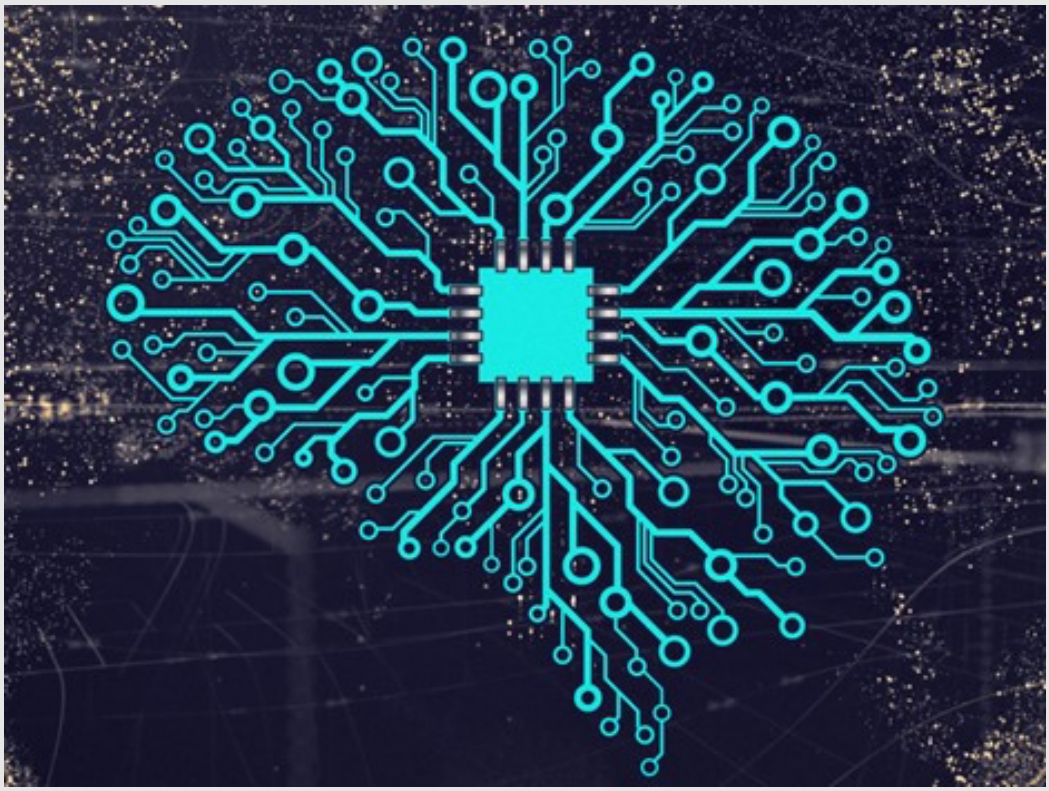
EPQ Dissertation

TO WHAT EXTENT DO THE RISKS OF NEURAL NETWORKS OUTWEIGH THE BENEFITS?



Vivian Lopez

Table of Contents

1 - Abstract

Neural Networks are a rapidly advancing field because of both the improvements in hardware technology to facilitate the complexities of neural networks, as well as the realisation and utilization of their capabilities with notable examples including self-driving cars produced on large scales as well as companies including google and Facebook taking advantage of neural networks to learn about their users.

With this increasing development in the field, we cannot assume that all the results are positive and conducive to improving peoples live as, with all improvements to tech, the potential for misuse increases.

History shows that cybersecurity threats evolve along with new technological advances. Relational databases brought SQL injection attacks, IoT devices brought new ways to create botnets, and the internet in general opened a Pandora's box of digital threats. Social media created new ways to manipulate people through targeted content delivery and made it easier to gather information for phishing attacks. And bitcoin enabled the delivery of crypto-ransomware attacks.

Clearly, new technology entails new security threats that were previously unimaginable. And in many cases, we have had to adapt to such threats.

Recently, deep learning and neural networks have become very prominent in shaping the technology that is becoming prevalent in various industries. From content recommendation to disease diagnosis and treatment and self-driving vehicles, deep learning is playing an increasingly important role in making critical decisions.

What could this lead to in the future? This is something I will explore through various examples in society today and how they could be indicative of what the future may hold.

2 - Introduction

The original goal of the neural network approach was to create a computational system that could solve problems like a human brain. However, over time, researchers shifted their focus to using neural networks to match specific tasks, leading to deviations from a strictly biological approach. Since then, neural networks have supported diverse tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games, and medical diagnosis. [1]

Now, they are known to be suitable to solve different problems they are extremely useful, but also for certain negative uses, they can be infamous. These negative uses include deepfakes, malware i.e., viruses and spyware, among others. The example of deepfakes is an intimidating idea, if you have come across the infamous Obama deepfake video, you would realise how realistic and terrifying they can be, this not only highlights the power of neural networks but also the potential they have for malice. I will go onto explain exactly how threats like these are created and where they have been taken advantage of.

Often, the possibility of being affected by these are overlooked and this leads to vulnerabilities in companies as well as individuals, hence I have decided to write a dissertation on the topic and explore the extent to which these negative applications pose as risks and to what extent they limit the positives.

2.1 - Research Review

To answer the question: 'To what extent do the risks of neural networks outweigh the benefits', I first had to decompose it into different stages that would allow me to build up both points of the question, benefits, and risks, and to do this, I began by building my knowledge on neural networks themselves.

My starting point for the project began with an online course on Codecademy titled 'Build Deep Learning Models with TensorFlow'. Despite the programming aspect of the course being the centre of focus, the basic knowledge that it developed on neural networks gave me a feel for what they were, how they functioned and where they were used as well as how they were used. Beginning with the introduction to perceptrons (A simple model of an artificial neuron) and then going on to building deep neural networks, the course highlighted the core aspects of neural networks and the kinds of problems they were used to tackle in addition to some possible areas of misuse, for example, the issue of the AI black box was posed when considering neural networks, this concerned the idea that only people who created the network know what it does and how it works and if the model is being widely used, the creator has power over what the outcomes are. This highlighted an issue related to neural networks however, not many other examples and areas where they pose risks were given in the course so after completing it and having a grasp of the concepts behind neural networks, I moved onto to further research on risks specifically as this was something I needed to find out more on, directly related to my research question.

Alongside taking the course, I watched videos and read the Oxford short Introduction to Artificial Intelligence. The videos included those on the 3Blue1Brown YouTube channel

which gave clear and informative descriptions of neural networks through the example of character recognition, and I found this to be extremely helpful in understanding the inner workings of neural networks. The short introduction gave me a general picture of Artificial Intelligence and how neural networks fit into AI, even though it wasn't fully targeted to my research question, it was helpful nonetheless and explored concepts in an interesting way so the book was useful, and it provided me with more information that could help in writing the dissertation.

Despite having already completed the course and read the short introduction, I still wanted to see further into neural networks and to do this I looked to a book titled 'Neural networks and deep learning' by Michael Nielsen. The book went into much greater depth, due to this I was able to get a feel for ideas such as the mathematics behind neural networks which was highly complex, and this highlighted to me the complex nature of the networks as well as solidifying the idea of them being a black box to those who were only involved in using them rather than making them.

Since risks aren't considered greatly, I found it difficult to locate books relating to risks directly, most on the topic of neural networks covered in-depth explanations of how they function or how to code/apply them to problems. So, for the majority of my sources relating to risks, I searched for and used articles on the internet and there were some which were particularly useful in highlighting the dangers, for example a Tech Talks article on the security threats of neural networks introduced me to a range of different threats which I had not previously known about including adversarial attacks where mistakes that neural networks are prone to are leveraged, for example a certain type of glasses caused a network to be unable to recognise a person through a camera and this clearly poses a threat as it could allow numerous security measures to be bypassed.

However, rather than going to the internet and finding relevant articles, it may have been worth looking into more of other types of sources such as a greater range of videos but given that the course which I took which gave me the background knowledge I needed, I felt that this wasn't fully necessary and the best way to find out about my research question specifically, was to directly find information research online. Though other sources such as some books did provide information which was useful and informed parts of the dissertation such as the brief history of neural networks.

The articles that I found were very useful in directly giving risks and going in depth around them however, since their purpose was to describe risks and give disadvantages of neural networks, this affected their reliability as they would have had a tendency to paint out neural networks as having greater risks than they actually have, hence I decided to search and make use of scientific papers, as, although the process of peer review may not be perfect, it exists as some form of moderation and this will contribute to a more balanced point of view when writing the main section of my dissertation.

Overall, my research included articles, scientific papers, a course, and videos to both give me an understanding of the area of neural networks as well as examples of applications that are relevant to my question.

2.2 - Defining Neural Networks & key terms

In this section I will introduce a range of terminology that will be important for the rest of the dissertation since it is a highly technical topic.

AI is the area of research where neural networks originated. It means intelligence demonstrated by machines. Face recognition and chatbots are good examples as they highlight situations where a machine can carry out human-like tasks i.e., recognising faces and chatting. Machine leaning is a specific type of AI; it is a branch of AI which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy [2].

One further field down is deep learning, a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data. This oxford dictionary definition may seem abstract so I will take the example of digit recognition. Say we wanted to give a computer an image of a handwritten digit and ask it to recognize the digit for us. Traditional programming would struggle with this greatly as each image is different and hardcoding rules for what a certain digit is, doesn't work particularly well. Instead, we can use deep learning with multiple layers of a 'neural network', with each layer recognising higher and more complex features. So, for example if the digit was a 1, the first layer could be detecting a straight line somewhere in the image, the next layer could be detecting an angled line joined to the larger line at the top and subsequent layers would detect more detailed features. (In reality, we don't know for sure that the network is doing this, but it will be doing something similar). Deep learning allows for complex and detailed things to be learned which is what makes it so unique and why it is becoming increasingly popular. [3] [4]

Going a further layer down into machine learning, we get to neural networks. Neural networks, also known as artificial neural networks (ANNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

ANNs are comprised of node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. One of the most well-known neural networks is Google's search algorithm.

Due to the risk aspect of neural networks related to technology, cybersecurity is an important part of addressing them. As defined by IBM, it is the practice of protecting critical

systems and sensitive information from digital attacks. Also known as information technology (IT) security, cybersecurity measures are designed to combat threats against networked systems and applications, whether those threats originate from inside or outside of an organization. [5]

2.3 - How do Neural Networks Work?

Each node (neuron) has its own linear regression model, composed of input data, weights, a bias (or threshold), and an output.

Once an input layer is determined, weights are assigned. These weights help determine the importance of any given variable, with larger ones contributing more significantly to the output compared to other inputs. All inputs are then multiplied by their respective weights and then summed. Afterward, the output is passed through an activation function, which determines the output. If that output exceeds a given threshold, it "fires" (or activates) the node, passing data to the next layer in the network. This results in the output of one node becoming the input of the next node. This process of passing data from one layer to the next defines this neural network as a feedforward network.

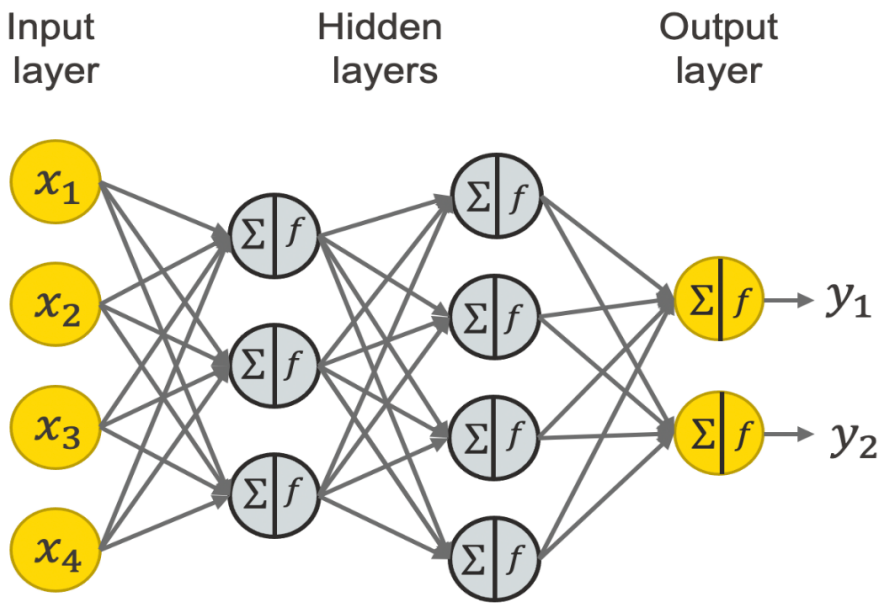


Figure 1 Feed Forward Neural Network

As we train the model, we'll want to evaluate its accuracy using a cost (or loss) function. Ultimately, the goal is to minimize our cost function to ensure correctness of fit for any given observation. As the model adjusts its weights and bias, it uses the cost function and reinforcement learning to reach the point of convergence, or the local minimum. The process in which the algorithm adjusts its weights is through gradient descent, allowing the model to determine the direction to take to reduce errors (or minimize the cost function). With each training example, the parameters of the model adjust to gradually converge at the minimum. [6]

2.4 - Brief History of Neural Networks with Perceptron Algorithm & Backpropagation

Frank Rosenblatt's Perceptron:

Frank Rosenblatt was a psychologist who built on the work of Warren McCulloch and Walter Pitts. He created a simplified mathematical model of how neurons in our brain operate, taking a binary input, multiplying it by weights (synapse strength to each neuron) and thresholding the value (checking if it is above a certain threshold, so determining whether the neuron will fire or not). The significance of his work on the perceptron was that it was able to carry out basic logic operations: AND, OR and NOT. This gave people a reason to believe that AI could think, as if it could carry out these logical operations, it could think. [7]

The main drawback of the model created by McCulloch and Pitts was that it could not learn, this was what made the perceptron excel. Rosenblatt, was able to make perceptron's learn, given the work of Donald Hebb who stated: "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." When translated to perceptrons, this gave the following algorithm:

1. Start off with a perceptron having random weights
2. For the inputs in an example in the training set, compute the output of the perceptron
3. If the perceptron output does not match the known correct output, then:
    a. If the output was 0 and should have been 1, increase the weights that had an output of 1
    b. If the output was 1 and should have been 0, decrease the weights that had an output of 1
4. Go to the next example in the training step and repeat steps 2-4 until the perceptron makes no more mistakes.

This algorithm allows the perceptron model to learn linearly separable data. However, this posed an issue, problems that weren't linearly separable, such as an XOR logic gate were not possible. This is something that was addressed later with a different type of neuron.

Things started moving quickly for neural networks around this time and in 1959 at Stanford, Bernard Widrow and Marcian Hoff developed the first neural network successfully applied to a real-world problem. These systems were named ADALINE and MADALINE after their use of Multiple ADAptive LINEar Elements, the latter of which was specifically designed to eliminate noise in phone lines and remains in use today. These artificial neurons however were different from perceptrons in what they returned as output, which in this case was the weighted input, rather than a binary output.

One of the problems, as well as not being able to learn XOR gates, that rose was with the exceedingly long runtimes required for running these networks.

All this came to an end in 1969 with the publication of a book “Perceptrons” by Marvin Minsky, founder of the MIT AI Lab, and Seymour Papert, director of the lab. The book argued that the Rosenblatt’s single perception approach to neural networks could not be translated effectively into multi-layered neural networks.

To evaluate the correct relative values of the weights of the neurons spread across layers based on the final output would take at a minimum, several iterations all the way to an infinite number.

Minsky in his text laid out these and other problems with Neural Nets and effectively led the larger scientific community and most importantly the funding establishments to the conclusion that further research in this direction wouldn’t lead to anywhere or produce any fruitful results.

The effect of this text was powerful and dried up funding to an extent that, for the next 10–12 years, no-one would take on any project that had that involved neural networks. The next period was known as the AI Winter however, after some time they did resurface.

Backpropagation, a method devised by researchers since the 60’s and continuously developed well into the AI winter, was an intuition-based method. The first person to see the potential for neural network was Paul Werbos who developed the multilayer perceptron. However, this work wasn’t noticed by anyone in the community until Parker published a report on his work at M.I.T. in 1985. It was only after being re-re-discovered by Rumelhart, Hinton, and Williams and republished in a clear and detailed framework that the technique took over the community by storm. The same authors also addressed the specifics drawbacks laid out by Minsky in his 1969 publication in a later text.

Backpropagation along with Gradient Descent forms the backbone and powerhouse of neural networks. While Gradient Descent constantly updates and moves the weights and bias towards the minimum of the cost function, backpropagation evaluates the gradient of the cost of weights and biases, the magnitude and direction of which is used by gradient descent to evaluate the size and direction of the corrections to weights and bias parameters. [8]

And by the 1990’s, Neural networks were back, with new and powerful techniques that would open the doors to many problems yet to be solved.

Now, looking at the current state of neural networks, they have been developed greatly, thanks to the increase in computational power. There are a wide range of specific types of networks with different purposes and different networks are better suited to different problems, for example the multilayer perceptron network [9], can approximate problems which can be given as continuous functions such as stock analysis, image identification and spam detection, in fact this type of network is quite diverse in the problems we can solve. In the next section, I’ll be looking at generative adversarial networks (GANs) which is a specific type of network that doesn’t consist of just one set of layers of nodes but two different sets, one called a generator, and one called a discriminator, and both work together to produce surprisingly accurate results.

3 - Neural Networks and their Uses

3.1 - Generative Adversarial Networks

Generative adversarial networks (GANs) are architectures that use two neural networks, faced against the other, hence “adversarial”, to generate new, synthetic instances of data that can pass for real data. They are used widely in image generation, video generation and voice generation.

GANs’ potential for both good and evil is huge, because they can learn to mimic any distribution of data. That is, GANs can be taught to create worlds eerily like our own in any domain: images, music, speech, and text. They are robot artists in a sense, and their output is impressive. But they can also be used to generate fake media content, and are the technology underpinning Deepfakes, an example which I will return to.

The two networks used involve different types of algorithms, discriminative and generative algorithms, work against each other to give a final high-quality output. To Describe the generator, it might be helpful to first talk about the discriminator, taking an example, a discriminator would tell the difference between an email which is spam and one which isn’t. This is in a sense the opposite to what I have explained previously about perceptrons, taking input about certain features, and giving output based on the weights associated to the features. With a discriminator, the features are the things that are trying to be predicted i.e., if an email is spam or not, given certain other information such as the types of words that are included in the email. A generator tries to answer the question, assuming something given is true (i.e., an email is spam), how likely are the features?

Within GANs, discriminators and generators have specific jobs, generators create new data instances and discriminators evaluate them for authenticity.

Let’s say we are trying to do something simpler than mimic paintings. We are going to generate hand-written numerals like those found in the MNIST dataset. MNIST is a database. The acronym stands for “Modified National Institute of Standards and Technology” and it contains handwritten digits (0 through 9) [10]. It is often used to test an architecture for a neural network. The goal of the discriminator, when shown an instance from the true MNIST dataset, is to recognize those that are authentic.

Meanwhile, the generator is creating new, synthetic images that it passes to the discriminator. It does so, hoping that they, too, will be deemed authentic, despite being fake. The goal of the generator is to generate passable hand-written digits: to lie without being caught. The goal of the discriminator is to identify images coming from the generator as fake. [11]

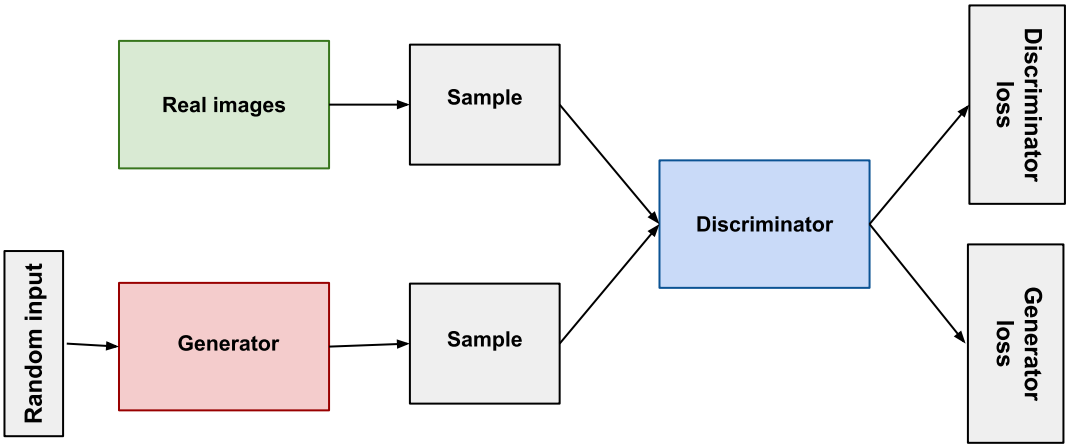


Figure 2 Structure of Generative Adversarial Network

[12]

Since GANs are designed specifically to create new data, they can be helpful in areas requiring new data to be generated which passes as authentic. One example of such a situation is Image Super-Resolution. This is the ability to generate high-resolution versions of input images. [13]

A paper written on this topic of super resolution published in 2016 highlighted certain GAN models such as SRGAN which performed exceedingly well leading to the higher resolution image generated being indistinguishable from the original high-resolution image after using the lower resolution version to generate the new image. This means that GANs are effective in increase the quality of images which can be a useful tool.

One area in which super-resolution is useful is in healthcare, in fact they are becoming increasingly popular in this area. The reason for the high demand for GANs in healthcare is that images should fit particular requirements and be high-quality. High image quality can be difficult to obtain under certain measurement protocols, for example, there is a strong need to decrease the effect of radiation on patients when using low dose scanning in CT or MRI. It has the effect of complicating efforts to obtain good quality pictures because of the poor-quality scans.

Super-resolution improves the captured images and can remove the noise quite well, however adoption of GANs in the medical area is quite slow as many experiments and trials have to be made due to safety concerns. When dealing with healthcare, it is mandatory to involve a number of domain experts to evaluate the models and ensure the denoising does not distort the actual content of the image in some way that could lead to an incorrect diagnosis.

GANs can also be used as a service, instead of finding a specific niche application for the models, some companies offer access to GANs and all the infrastructure and interfaces to handle the data, train the models, and obtain the final results.

Runway AI is one of such companies, positioning itself as a platform for Machine Learning and enabling novel content creation techniques. This kind of service that allows clients to leverage the use of GANs easily proves to be highly useful for content creators, allowing them easily and effectively create content with less effort. This service brings GANs to the masses without an interface that would prove inconvenient to non-programmers. [14]

As shown, GANs can be helpful in areas where data creation/augmentation is required. I have given some examples of such areas however there are even more. Some further uses include colorising black and white images, and even generating data for neural networks to be trained on where there is lack of data.

3.2 - Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are everywhere. It is arguably the most popular deep learning architecture. The recent surge of interest in deep learning is due to the immense popularity and effectiveness of these networks.

CNNs are now the go-to model on every image related problem. In terms of accuracy, they blow competition out of the water. The latest network is called LeNet-5 which a 5-layer CNN that reaches 99.2 % accuracy on character recognition. [15] It is also successfully applied to recommender systems, natural language processing and more. The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision. For example, given many pictures of cats and dogs it learns distinctive features for each class by itself. So, we don't need to tell it what different parts of the image should like, it finds out on its own.

CNNs are also computationally efficient. They use special operations which cut down the parameters, essentially reducing the problem then solving it. This enables CNN models to run on any device, making them universally attractive. In fact, they can now do image classification better than humans.

CNNs can be considered to be a more complex version of the standard perceptron model. This is because of the fact that they automatically extract features from input images. They consist of a set of layers.

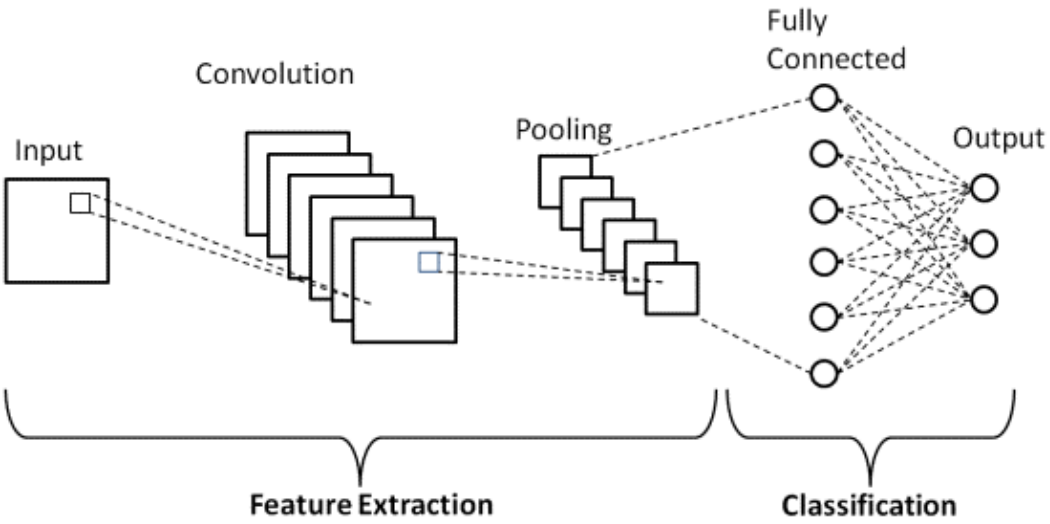
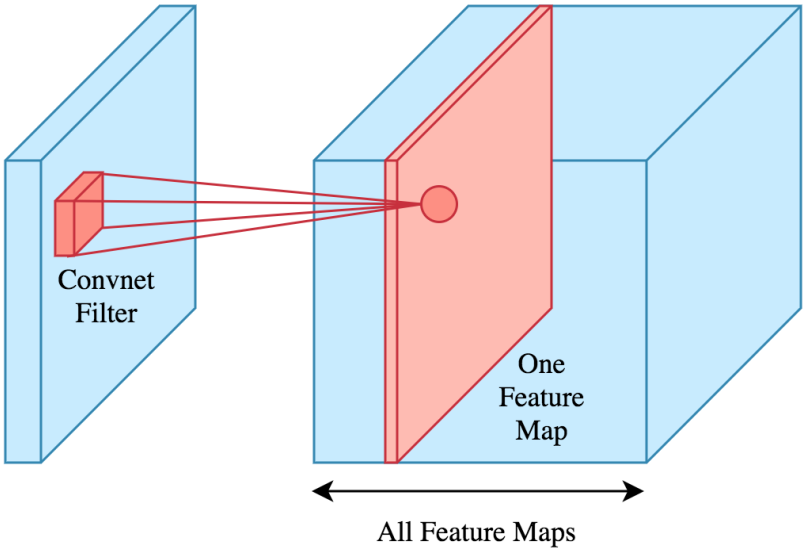


Figure 3 Convolutional Neural Network Architecture

[16]

The first is a standard layer in neural networks, the input layer. This is where the image is fed into the model.

The second layer is the convolutional layer. Convolution is one of the main building blocks of a CNN. The term convolution refers to the mathematical combination of two functions to produce a third function. It merges two sets of information. With a CNN, the two pieces of information that are combined are called a filter and the input image. The result of combining the image along with a filter, results in a feature map. To execute convolution, the filter is effectively slid over the image. Multiple feature maps are generated.



Filter used to create feature map from image

[17]

The third layer is the pooling layer in which the information from the feature maps is condensed, often Max Pooling is used which takes the maximum value in in each window from feature maps. The function of pooling is to continuously reduce the dimensionality to reduce the number of parameters and computation in the network. This shortens the training time and controls overfitting (which is where the model becomes specifically trained for the input data). If a model becomes highly accurate with training dataset, this doesn't mean it works well in the real world, i.e., it is not generalized which is prevents the model performing well in real-life scenarios.

The final layer is essentially the same as the multi-layer perceptron model. This is where backpropagation is used and results in a final model which can classify images.

Due to the proven high accuracies, they are more reliable in areas such as medicine, and arguable can be used in even more useful ways than GANs in this field. Specifically in radiology, with methods such as CT and PET scans, implementing neural networks into the process can allow for instant classification of abnormalities. An example of this I came across through a scientific paper was that of detecting lung cancer.

Lung nodules (small abnormal areas that are sometimes found during a CT scan of the chest) are able to be classified into being benign or cancerous by CNNs to very high accuracies.

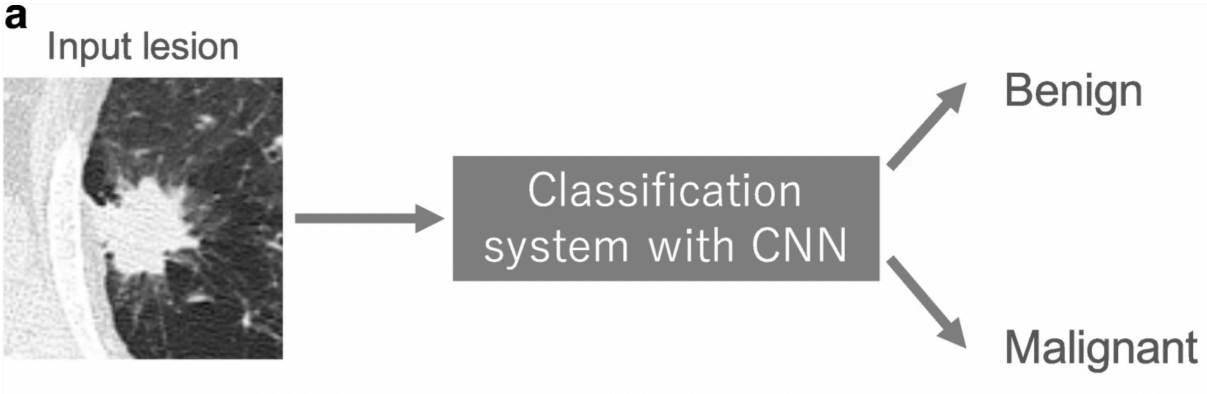


Figure 4 Process of CNN Classification of Lung Nodule

[18]

The advantage of this is clear, it allows doctors to improve the analysis in reports and aid in diagnosis which improves the process of diagnosis and gives doctors another resource which they can refer to if they require it.

Another example which I wanted to take a look at specifically was face recognition since it is such a widely used technology.

With CNNs, the development of facial recognition capabilities has made a big step forward with the Face ID technology breakthrough from Apple. This technology is a biometric facial recognition algorithm that performs user's authentication and can be adapted to many use cases like:

- Unlocking devices like phones or computers
- Unlocking Doors and Systems
- Validating online transactions (especially the financial ones)
- Authorizing online purchases payments
- Mass Surveillance in airports, Railway, stadium, government offices, and business establishments, or for the whole population like in China (citizen score).

Many of these use cases show that face recognition can be added as an extra layer of security to a range of different systems due to the uniqueness of faces. This extra layer provides even more reliability which can only be seen as beneficial. In addition, it is highly efficient so the time cost of implementing it is minimal. The networks themselves however require sufficient computational power to be able to run. This is something that can be avoiding using cloud-based solutions which mean using a remote system to carry out the computational expensive part of running the model and the results can simply be passed back to users.

CNNs also provide a solution of processing large amounts of data which would have been very difficult manually. With the idea of mass surveillance for, in example airports, threats such as criminals can now easily be identified through comparisons of features to set of features linked to certain individuals contained in criminal databases. Moreover, the process of boarding can be sped up significantly using facial recognition rather than having to check boarding passes which provides a greater efficiency.

However, they are some concerns that come with CNNs being used in this way which is why I will come back to the topic of surveillance when addressing the risks.

3.3 - Recurrent Neural Networks and Natural Language Processing

Recurrent Neural Networks (RNNs) are a form of machine learning algorithm that are ideal for sequential data such as text, time series, financial data, speech among others. This structure differs from both GANs and CNNs as the they make use of data which isn't sequential, mainly images.

RNNs are ideal for solving problems where the sequence is more important than the individual items themselves.

They function in a similar way to feed-forward neural networks (which are neural networks that don't use backpropagation, instead only pass data forward through the network) in the sense that they have a similar architecture with the difference that they are able to retain information of outputs. For example, if a feed forward neural network was given the work 'language', it would process the 'l' then the 'a' then the 'n' and so on. A RNN instead would process the 'l' and save the output then move onto the 'a', remembering the output of l and continue. In this way, an RNN has a short-term memory which makes it useful when processing sequences as it is able to remember previous parts of the sequence.

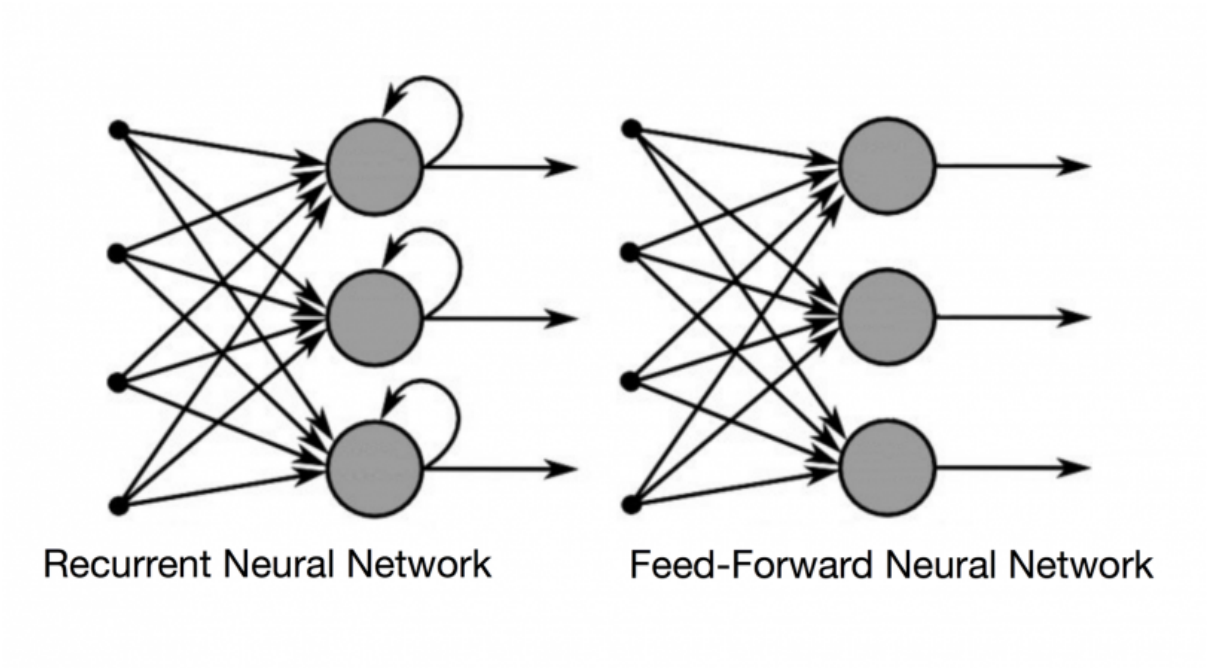


Figure 5 Structure of RNNs

[19]

The implementation of a short-term memory creates a powerful network which has the ability to create content by itself that can be almost indistinguishable from human created content.

Taking specifically the creation of books, RNNs are particularly useful in this area since they can take sentences to be sequences of words and create text based on this. A novel titled 'The Road' by Ross Godwin and Kenric McDowell is an example of such a book created by an RNN.

Ross Goodwin drove from New York to New Orleans in March 2017 with an AI in a laptop hooked up to various sensors, whose output the AI turned into words. As writing is a task done in continuous sequences it was framed by an RNN. The model had been trained on previous text data. Data was collected in forms of scenes, speech captured inside the car and coordinate values.

As understood from data collection, the book has been written on the basis of the mood set by the surroundings and how people's conversation changed with it. Despite ambiguity in the book, it provided descriptions and passages that, to readers, were interesting.

The project of creating the book with AI was definitely a success as it was funded by google and there were good critic reviews. Brian Merchant from the Atlantic said, "there are some striking and memorable lines".

The novel was published in 2018 by Jean Boîte Éditions. This is also one of the first novels completely written by AI. [20]

This success means that the models can be leveraged to create their own books, however, more usefully can be used to aid writes.

With the foundation of GPT-2 and now the revolutionary GPT-3 (which are both top performing RNNs), the writing industry can expect a major involvement of AI in their work. Creative artists often spend time discarding scenes or creations from the final cut. If Neural networks are used to decide which sections are better than others, informed decisions can easily be made in regard to which sections to include as well as how parts can be improved allowing writers to easily improve their content.

This application of RNNs highlights the capabilities of the network in the field of natural language processing (NLP). This field is concerned with how computers analyse large amounts of natural language data such as text.

Some other areas in this field where RNNs are particularly useful include:

- Search - autocorrect and autocomplete
- Language translation (i.e., Google Translate)
- Chatbots
- Targeted advertising
- Voice assistants (i.e., Alexa)
- Grammar checkers
- Email filtering

The range of applications in NLP exemplifies the usefulness of RNNs in relation to daily life, providing small efficiencies and profit gains for companies with better advertising. [21]

4 - Risks of Neural Networks

Now that we have both a background on the topic as well as knowledge of the benefits that Neural Networks provide, we can start to investigate the different types of risks that they can bring as well as specific examples of these risks.

4.1 - Deepfakes

What are Deepfakes? They are content such as images, videos or sound generated to imitate real content and they aren't restricted to being created using AI, they can be created by people using things like photoshop, however, with the rise of GANs, they are becoming more and more realistic and can be created more easily.

Deepfakes highlight the dangerous side of neural networks. The idea of being able to create data which is seemingly real using the discriminator and generator over many iterations, makes things like fraud very possible. Take this very recent example that was published in Forbes: "Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find". A shocking and poignant example of the extent to which GANs can be taken advantage of. Forbes went onto describe the events:

In early 2020, a bank manager in the Hong Kong received a call from a man whose voice he recognized—a director at a company with whom he'd spoken before. The director had good news: His company was about to make an acquisition, so he needed the bank to authorize some transfers to the tune of $35 million. A lawyer named Martin Zelner had been hired to coordinate the procedures and the bank manager could see in his inbox emails from the director and Zelner, confirming what money needed to move where. The bank manager, believing everything appeared legitimate, began making the transfers.

What he didn't know was that he'd been duped as part of an elaborate swindle, one in which fraudsters had used "deep voice" technology to clone the director's speech, according to a court document unearthed by Forbes in which the U.A.E. has sought American investigators' help in tracing $400,000 of stolen funds that went into U.S.-based accounts held by Centennial Bank. The U.A.E., which is investigating the heist as it affected entities within the country, believes it was an elaborate scheme, involving at least 17 individuals, which sent the pilfered money to bank accounts across the globe. [22]

This is only one possible use of GANs for criminal purposes, there are a range of other duplicitous activities which can be carried out using GANs, some of which may not even be widely known yet, and some which are bound to crop up in the future.

[23]

4.2 - Adversarial attacks

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake; they're like optical illusions for machines. [24]

Neural networks themselves aren't perfect and hence make these mistakes. However, since they are not the same as humans, the mistakes they make might seem completely illogical. For example, students at MIT showed that making small changes to an image of a toy turtle causes a neural network to classify it as a rifle. This is not a mistake a human would make and raises issues when implementing neural networks in real life.

Figure 6 Researchers at labsix showed how a modified toy turtle could fool deep learning algorithms into classifying it as a rifle

[25]

Adversarial examples such as that of the toy turtle cause neural networks to make these highly irrational mistakes and of course, these can be leveraged and taken advantage of.

in 2017, researchers at Samsung and Universities of Washington, Michigan and UC Berkley showed that by making small tweaks to stop signs, they could make them invisible to the computer vision algorithms of self-driving cars. This means that a hacker can force a self-driving car to behave in dangerous ways and possibly cause an accident. As the examples below show, no human driver would fail to notice the "hacked" stop signs, but a neural network could perfectly become blind to it.

Adversarial attacks are not limited to computer vision. They can also be applied to voice recognition systems that depend on neural networks. Researchers at UC Berkley developed a proof-of-concept in which they manipulated an audio file in a way that would go

unnoticed to human ears but would cause an AI transcription system to produce a different output. For instance, this kind of adversarial attack can be used to change a music file in a way that would send commands to a smart speaker when played. The human playing the file would not notice the hidden commands that the file contains.

The dangers here are clear however, when I was researching about these kinds of attacks, I wasn't able to find any solid examples where the attacks had happened in real life and caused damage. This indicated to me that despite their potential, adversarial attacks aren't currently feasible to be consistently and effectively carried out. Researching further I was able to find a scientific paper on 'Adversarial Attacks on Medical Imaging' which highlighted that they are only being explored in laboratories and research centres. There's no evidence of real cases of adversarial attacks having taken place.

Given the great amount of research that is currently being conducted on the topic with, for example, research at IBM and MIT being able to formulate a proposed solution to adversarial input attacks; It seems unlikely that they will cause significant issues and therefore the risk they pose is limited however an example where they cause damage could crop up at any time since the full application of defences against them aren't in place so they cannot be completely ignored. [26]

Moreover, developing adversarial attacks is just as hard as finding and fixing them. Adversarial attacks are also very unstable, and they can only work in specific circumstances. For example, a small change in pixels from a result of a change in lighting may disrupt an adversarial attack. This fact contributes to the idea that they aren't fully feasible in real-life at this moment in time.

4.3 - Data Poisoning

Data poisoning involves tampering with machine learning training data to produce undesirable outcomes. An attacker will infiltrate a machine learning database and insert incorrect or misleading information. As the algorithm learns from this corrupted data, it will draw unintended and even harmful conclusions. [27]

Deep learning algorithms have no notion of morality, common-sense and bias that humans possess. They only reflect the hidden biases and tendencies of the data they train on. In 2016, Twitter users fed an AI chatbot deployed by Microsoft with hate speech and racist rhetoric, and in the span of 24 hours, the chatbot turned into a Nazi supporter and Holocaust denier, spewing hateful comments without hesitation.

Because deep learning algorithms are only as good as their data, a malicious individual that feeds a neural network with carefully tailored training data can make it to take on harmful behaviour. This kind of data poisoning attack is especially effective against deep learning algorithms that draw their training from data that is easily and publicly available.

An interesting example was the case of two brothers who weren't twins, didn't look alike and were years apart in age. The brothers initially posted a video that showed how they could both unlock an iPhone with Face ID. But later they posted an update in which they

showed that they had tricked Face ID by training its neural network with both their faces. Despite being a harmless example, it shows the danger of data poisoning and how it can be used.

Similar to adversarial attacks, I was unsure as to whether these data poisoning attacks were currently being carried out as the articles and sources, I came across on this specific issue described the attacks but didn't give examples. However, I was able to come across an article written on SCO online which gave specific examples of attacks, and this indicated to me that data poising was in fact a current risk and the research on protecting against it was necessary to address it as a current issue.

The examples that were given were the following:

A real-world example of this is attacks against the spam filters used by email providers. In a 2018 blog post on machine learning attacks, Elie Bursztein, who leads the anti-abuse research team at Google said: "In practice, we regularly see some of the most advanced spammer groups trying to throw the Gmail filter off-track by reporting massive amounts of spam emails as not spam [...] Between the end of Nov 2017 and early 2018, there were at least four malicious large-scale attempts to skew our classifier."

Another example involves Google's VirusTotal scanning service, which many antivirus vendors use to augment their own data. While attackers have been known to test their malware against VirusTotal before deploying it in the wild, thereby evading detection, they can also use it to engage in more persistent poisoning. In fact, in 2015 there were reports that intentional sample poisoning attacks through VirusTotal were performed to cause antivirus vendors to detect benign files as malicious.

[28]

Both attacks were on Google a large company which was almost powerless against the attacks. This suggests that smaller organizations using neural networks would be completely vulnerable against these attacks and therefore puts them at a significant risk to businesses who could have their services rendered useless as a result of these attacks.

4.4 - The Black-Box Nature of Neural Networks

Neural networks can be seen as black boxes due to the difficulty in figuring out exactly what they are doing behind the scenes to find relationships between data and give certain output. This can lead to certain issues.

An example of this issue is a neural network being used to determine car insurance pricing, the creator can manipulate the network to give higher probabilities for certain people to have higher prices and therefore creating bias however, due to the black box nature, this may not be consciously picked up. [29]

Although this issue in itself doesn't prove to be a threat, it can be seen as a risk i.e., if the network doesn't behave correctly due to bias, it can cause unfair results in the processes the network is involved in. Therefore, preventative should be taken to regulate this aspect and reduce bias.

4.5 - Deep Learning-Based Malware

Something that many people would have experienced or heard about is malware and for most people, it is commonly heard of and today there are various methods to protect against it, from ant-virus software to simply knowing the characteristics of a dodgy website or email. However, with the rapid rise of AI, malicious users have been given a new way to infect computers with malware. IBM's DeepLocker is a project that highlights the capabilities of neural networks in malware. The key usage is hiding a malicious payload within a deep neural network, which makes it very hard to detect.

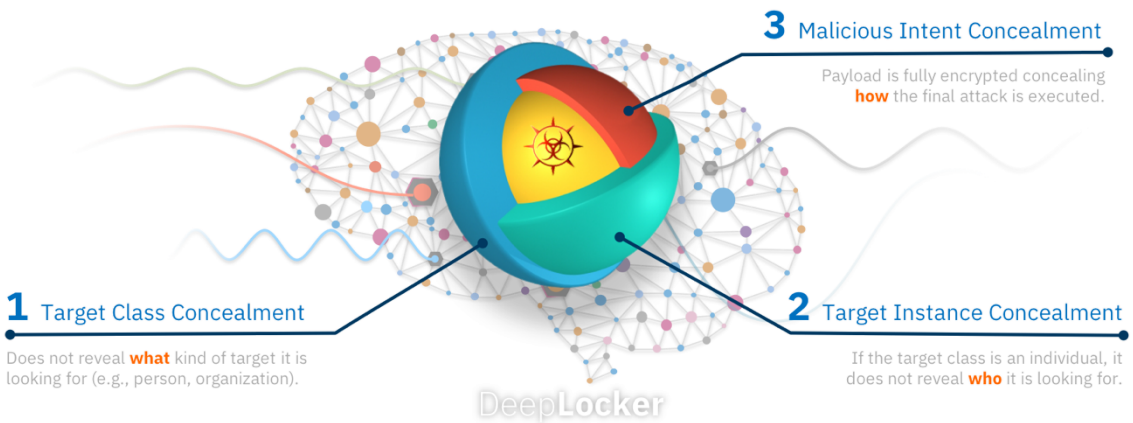


Figure 7 Deep Locker Structure

[30]

The model is trained to behave normally unless it is presented with a specific input: the trigger conditions identifying specific victims. The neural network produces the "key" needed to unlock the attack. DeepLocker can leverage several attributes to identify its target, including visual, audio and geolocation. As it is almost impossible to exhaustively enumerate all possible trigger conditions for the AI model, this method would make it extremely challenging for malware analysts to reverse engineer the neural network and recover the details, including the attack payload and the specifics of the target.

DeepLocker can leverage the black-box nature of deep neural networks to conceal the trigger condition. A simple "if this, then that" trigger condition is transformed into a convolutional neural network that is very hard to decode. [31]

This example takes advantage of the previously mentioned black-box nature of neural networks and therefore means that this property can pose a risk as it allows for such malware to be implemented. This suggests that measures to regulate networks need to be

put in place otherwise attackers would have free reign in this area of using neural networks for malware.

4.6 - Facial Recognition and Privacy

Previously I mentioned face recognition allowing for improved efficiencies in mass surveillance and this being a positive use for airports however, this can be seen as a negative use of CNNs due to the issue of privacy.

Taking an article written on CNN Travel:

In April 2019, traveller MacKenzie Fegan was left surprised and confused when she boarded a JetBlue flight from the United States to Mexico, without handing over her passport, or travel documents.

"There were plastic barricades across the front of each lane, I look to my right, and the gate opens," she tells CNN Travel. "I was like, 'What, just happened?' There was no boarding pass scan, nothing like that."

The traveller went on to tweet:

“I just boarded an international @JetBlue flight. Instead of scanning my boarding pass or handing over my passport, I looked into a camera before being allowed down the jet bridge. Did facial recognition replace boarding passes, unbeknownst to me? Did I consent to this?” [32]

This incident makes it clear that companies and organizations can easily use neural networks without the public being actively aware, hence violating privacy however since neural networks are relatively new, it is difficult to deal with this issue in a concrete way with the lack of regulation and laws on the topic.

An even more worrying application of mass surveillance is in China where the government can be regarded as a one-party dictatorship [33]. Their use of CNNs for mass surveillance gives them more power and invades privacy even further. With the idea of the social credit system, CNNs being able to detect and recognise people and therefore give indications to their activities and purchases, the government is more control, and this limits the freedom of citizens. This introduces the idea of being always watched and this is an intimidating idea that has been given power from CNNs being developed, a major drawback of CNNs.

5 - How Risks are Mitigated

5.1 - Protecting Against Deepfakes

Despite the intimidating factor of deepfakes, there are efforts to tackle them. Google in recent years released a dataset of deepfakes along with tools to help research in mitigating the effects of deepfakes and creating solutions to counteract them.

Cybersecurity specialists in particular are now tasked with the difficult job of spotting deepfakes. This also means that specialists must take into consideration that there is a large possibility that deepfakes have come from within the organization itself which means that there is likely to be a trail which cyber forensics can uncover and therefore use to solve the issue internally.

When analysed, multiple facial deepfakes gave unnatural looking eyes or movement of facial features. This still may be hard to detect by an average person.

Aside from addressing them directly through detection methods employing cyber security best practises such as:

- Employing a zero-trust philosophy especially in companies and high-stake situations such as the bank-transfer situation earlier mentioned
- Verifying sources as legitimate
- Using fingerprints or watermarks when transferring images, thereby making it difficult for attackers to create deepfakes

These measures help defend against social engineering, which is where attackers target people as the weak point of systems, so that incidents such as the money heist are prevented.

Overall, the defence systems already in place will work to prevent deepfake phishing and social engineering attacks. Deepfakes are still in the earliest stages, so cybersecurity has the advantage of preparing defences as the attacks improve.

5.2 - Protecting Against Adversarial Attacks

After coming across a scientific paper on the adversarial attacks [34], I found out that GAN's, specifically APE-GAN, can be used to take adversarial input and convert it to normal input so that it passes through the network correctly which addresses the issue directly however, after further testing of this method, the paper concluded that in fact the APE-GAN wasn't as effective as it seemed and only seemed to work when trained in specific scenarios with specific data so although the idea is there, the implementation isn't good enough to be reliable. The evidence that the paper provided gave me confidence in its conclusions.

There are other methods that can be employed to deal with adversarial attacks, these include penetration testing which involves seeing how the network responds to adversarial

examples and giving additional training to accommodate these flaws which is a short-term fix for the specific adversarial attacks that have been accounted for and are previously known, however new attacks would still pose a threat.

Overall, there isn't a concrete way that completely solves the issue of protecting against these attacks and as a result, the best way of dealing with them is to first identify situations in which they would be a high-risk and come up with contingency arrangements in the situation that the networks don't work as a result of being fed with these adversarial inputs.

However, ensuring these arrangements are considered is not going to be universally put in place so legislation in this area would be a good progression towards ensuring safer systems.

5.3 - Protecting Against Data Poisoning

One well-established method is using training data filtering. This technology focuses on the control of the training data, using detection methods and data sanitization to prevent the system from being attacked. The defender separates adversarial examples from normal ones and removes these malicious samples. Poisoned samples are generally crafted by statistic knowledge of training data and therefore can be detected simply.

This practise is already common but to ensure data poisoning isn't conducted, this practise needs to be employed by all important neural network developers where the networks themselves play a significant role in daily life so generally bigger companies and organizations need to make this standard. Again, this brings the need for legislation in the area to enforce certain rules in protecting against data poisoning so that neural networks aren't vulnerable and don't result in damage caused to companies and consumers.

Another method that I was able to source from a scientific paper was the traceback of Data Poisoning Attacks in Neural Networks. This method developed by researchers essentially trimmed off all non-poisoned data examples till the poisoned data examples remained which were responsible for the attack. This allows for the examples to be eliminated and training to be carried out again without these examples so that the network can function correctly. The researches claimed that their system achieved over 98.4% precision and 96.8% recall across all attacks which are impressive metrics. Due to the source being from the University of Chicago and numerous examples of tests and data being given, this meant that the source was reliable, and the figures could be taken to be accurate.

This means that methods to protect against data poisoning do exist and can be successfully implemented which means that as a risk, data poisoning is being addressed and shouldn't be something to worry about.

5.4 - Protecting Against Deep Learning Based Malware

Dealing with upcoming attacks similar to DeepLocker earlier mentioned, developments in AI used to detect malware would be incredibly helpful, additionally monitoring apps and how they behave across devices would give an indication to whether they contain these kinds of

malware which are difficult to detect and decode. These are things that anti-malware companies would have to develop so for now, all we can do is avoid downloading the malware in the first place which comes back to the previous point addressing social engineering by using measures already in place and this is the best we can do.

5.5 – Addressing the Black Box Nature of Neural Networks

With the AI black box problem becoming an increasing concern, AI developers are now turning their attention to solving it. Ideally, we want a way to be able to understand and explain what is happening in the networks however, we don't possess methods the knowledge to do this currently.

Until such functionality becomes available, the black box problem provides a reason to remain cautious of neural networks.

5.6 – Protecting Privacy

The main approach to protecting privacy of data used in neural networks is through legislation, as seen with other privacy laws. Most existing privacy laws however are based on a model of consumer choice: "notice-and-choice" (also referred to as "notice-and-consent"). We encounter this approach in a barrage of notifications and banners linked to lengthy and uninformative privacy policies and terms and conditions, which most of us don't read. [35]

This means that there needs to be a change in privacy laws or the implementation of new laws which ensure that the general public's privacy is met properly. Currently, there are no Privacy Laws that address AI directly.

Despite the current situation with AI Laws, new Laws will be coming. The time period they will start may not be now, but they will inevitably start to be introduced in the foreseeable future.

Most governments are currently using a "wait and see" approach to laws and regulations in this field and this makes sense seeing how it took years for laws to be put in place for the use of phones and driving. It is too early for them to make fully informed decisions; however, they could be taking more action as a preventative measure for future incidents.

The European Union is the most active in proposing new rules and regulations, with existing or proposed rules in seven out of nine categories of areas where regulation might be applicable to AI. Other governments aren't as active. [36]

6 - Conclusion

Neural networks provide benefits from small improvements in efficiency for daily tasks such as autocomplete to significant contributions to the field of medicine, including being able to detect cancerous tumours, clearly the range of fields that neural networks contribute to is wide reaching. They are pushing the boundaries of what is possible, with traditional programming, image recognition, book creation and advanced predictions would not be possible, at least not to the accuracies that neural networks reach. With the current rate of research, they are going to become even more powerful and will move onto to solve more challenging problems at a larger scale in the real world, benefiting society greatly, this could be through further automation, allowing us to focus our attention on more useful things, or even developing robots with highlight improved capabilities that can carry out dangerous tasks. The idea that neural networks can be intelligent essentially means there is no limit to what they are capable of given sufficient computational power which is something that is always increasing. The benefits that neural networks provide are vast and they will definitely be a large part of our future, however this future will only be possible with a strong grip on the risks involved.

Leaving risks unregulated and left to develop on their own could lead to a loss of control and can result in issues such as adversarial attacks becoming prominent, limiting the use of neural networks, and so limiting the benefits they provide. What Is currently being done to address these risks gives me confidence to assume that some of them such as data poisoning won't develop into problems which we will have to watch out for however others, more needs to be done. Should we really be waiting for an unfortunate outcome that requires action in terms of law making? I believe this is simply unnecessary and measures should be put into to place. These would include a list of testing models for robustness, including adversarial examples. As well as enforcing contingency arrangements in the event that models fail in a critical situation. An upside though, is that a significant amount of research is going into preventing threats such as deepfakes and adversarial attacks so despite the lack of laws, the technological side is likely to continue to advance at a rate that prevents the threats posing significant risks.

This leads me to the conclusion that the risks do not outweigh the benefits, given the range of preventative measures that are both being implemented as well as being developed and the sheer amount of use cases that are provided by them. I have only scratched the surface in terms of the uses and possible uses of neural networks, there are numerous applications that provide further benefit and many still yet to be discovered and implemented. Moreover, the work being done on neural networks and the research being put in stretches further than just neural nets to the field of AI as a whole and gets us closer and closer to understanding how the human brain works which in itself will allow us to improve the way we live, being given the power of understanding how any why we make the decisions we make.

7 – Evaluation

I have found the project to be an enjoyable experience, exploring an area which I had previously only experienced through code. Being able to research about it gave me a more solidified understanding of the core concepts behind neural networks and created the pathway into the cyber security elements involved as a result of particular risks which I had never come across before such as adversarial attacks.

Learning about the different types of networks as well as their particular uses led me to realize how integrated neural networks are in our daily lives and this made me wonder, if they are used so widely then someone must have found vulnerabilities in different aspects of them. Which ultimately led me to going into greater depth on the risks as opposed to my initial ideas and research questions, considering AI more broadly.

Through completing the project, I was able to develop my skills in putting together a dissertation as it is something I have never done before. The process of researching, collecting relevant information and critically evaluating the information was something I was able to work on and develop. Particularly when using different sources, I found myself thinking about how reliable the different parts of the sources were and whether a different kind of source was needed to verify certain information, for example the articles I was considering often had a component of opinion which led to certain points being focused on more than others, as opposed to scientific papers which were more objective in their points and provided much more concrete evidence to their claims, hence I used a range of papers for different section of the project such as the risks, some of which I had never come across before so reliable sources were highly important because of this.

The process of planning out the project was also an important component as I was able to split the main problem into smaller subsections which I could individually research and formulate points on. It also allowed me to reach a conclusion based on both the benefits and risks effectively, being able to draw on several points made throughout the course of the dissertation.

The presentation was also something which I enjoyed as It allowed me to engage with an audience on the topic of neural networks, highlighting different points which I found to be particularly interesting throughout the dissertation.

However, If I had more time, I would have liked to cover more uses perhaps to strengthen the benefits and highlight how useful neural networks can be. In addition, I would have liked to cover some more current techniques which are being developed to mitigate risks as they were highly interesting to learn about and would have also strengthened the argument further.

Overall, the experience of completing an extended project dissertation has left me more knowledgeable and fluent in the field of AI, specifically neural networks and it will be useful when I continue my studies at university, moving on to study computer science where this topic is likely to be visited. It has also taught me valuable skills in management, organization and critical thinking which will carry forward into the future.

8 - Bibliography

[1] SAS, "Artifical Neural Networks What They are and Why they Matter," [Online]. Available: https://www.sas.com/en_gb/insights/analytics/neural-networks.html#:~:text=Neural%20networks%20are%20computing%20systems,time%20%E2%80%93%20continuously%20learn%20and%20improve.. [Accessed 01 07 2021].

[2] IBM, "Machine Learning," [Online]. Available: https://www.ibm.com/uk-en/cloud/learn/machine-learning. [Accessed 11 07 2021].

[3] M. Nielsen, Neural, Nielsen, Michael.

[4] 3Blue1Brown, "Neural Networks Video," [Online]. Available: https://www.youtube.com/watch?v=aircAruvnKk. [Accessed 20 07 2021].

[5] IBM. [Online]. Available: https://www.ibm.com/topics/cybersecurity#:~:text=Cybersecurity%20is%20the%20practice%20of,sensitive%20information%20from%20digital%20attacks..

[6] CodeCademy, [Online]. Available: https://www.codecademy.com/learn/paths/build-deep-learning-models-with-tensorflow. [Accessed 21 07 2021].

[7] "neural-net-history," [Online]. Available: https://www.skynettoday.com/overviews/neural-net-history. [Accessed 10 09 2021].

[8] "History-Neural-Nets," [Online]. Available: https://towardsdatascience.com/a-concise-history-of-neural-networks-2070655d3fec.

[9] Educative, "What is a multi-layered perceptron?," [Online]. Available: https://www.educative.io/edpresso/what-is-a-multi-layered-perceptron. [Accessed 10 07 2021].

[10] Pathmind, "The MNIST database," [Online]. Available: https://wiki.pathmind.com/mnist. [Accessed 12 07 2021].

[11] Pathmind, "GANs," [Online]. Available: https://wiki.pathmind.com/generative-adversarial-network-gan. [Accessed 10 10 2021].

[12] Google Developer, "Gan Structure," [Online]. Available: https://developers.google.com/machine-learning/gan/gan_structure. [Accessed 10 10 2021].

[13] J. Brownlee, "What are generative adversaril networks?," [Online]. Available: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/#:~:text=GANs%20are%20an%20exciting%20and,to%20night%2C%20and%20in%20generating. [Accessed 11 10 2021].

[14] MobiDev, "GAN technology use cases for business application," [Online]. Available: https://mobidev.biz/blog/gan-technology-use-cases-for-business-application. [Accessed 12 10 2021].

[15] Towards Data Science, [Online]. Available: https://towardsdatascience.com/top-10-cnn-architectures-every-machine-learning-engineer-should-know-68e2b0e07201#:~:text=The%20latest%20work%20is%20called,accuracy%20on%20isolated%20character%20recognition..

[16] Medium, "CNN Architecture," [Online]. Available: https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697. [Accessed 11 11 2021].

[17] TowardsDataScience, "Applied Deep Learning," [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2. [Accessed 12 11 2021].

[18] "Lung Nodule Classification," [Online]. Available: https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9/figures/11. [Accessed 7 12 2021].

[19] Builtin, "Reccurent neural networks and lstm," [Online]. Available: https://builtin.com/data-science/recurrent-neural-networks-and-lstm. [Accessed 10 12 2021].

[20] Medium, "Interesting Novels written by artificial intelligence," [Online]. Available: https://medium.com/the-research-nest/interesting-novels-written-by-artificial-intelligence-d407e330fe07. [Accessed 17 12 2021].

[21] Analytics Vidhya, "Applications of NLP," [Online]. Available: https://www.analyticsvidhya.com/blog/2020/07/top-10-applications-of-natural-language-processing-nlp/. [Accessed 18 12 2021].

[22] "Forbes Deepfake," [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/. [Accessed 10 10 2021].

[23] CNN, "Pentagons race against deepfakes," [Online]. Available: https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/. [Accessed 11 10 2021].

[24] OpenAI, "Adversarial Example Reserch," [Online]. Available: https://openai.com/blog/adversarial-example-research/. [Accessed 17 10 2021].

[25] Labsix, "Physical Objects that fool neural nets," [Online]. Available: https://www.labsix.org/physical-objects-that-fool-neural-nets/. [Accessed 20 10 2021].

[26] IBM, "Deep learning hacks," [Online]. Available: https://www.ibm.com/blogs/research/2020/12/deep-learning-hacks/. [Accessed 10 09 2021].

[27] International secuirty journal, "What is data poisoning," [Online]. Available: https://internationalsecurityjournal.com/what-is-data-poisoning/#:~:text=Data%20poisoning%20involves%20tampering%20with,unintended%20and%20even%20harmful%20conclusions.. [Accessed 22 10 2021].

[28] CSO Online, "How data poisoning attacks corrupt machine learning," [Online]. Available: https://www.csoonline.com/article/3613932/how-data-poisoning-attacks-corrupt-machine-learning-models.html. [Accessed 16 10 2021].

[29] Codecademy, "Dangers of Black Box," [Online]. Available: https://www.codecademy.com/article/dangers-of-the-black-box. [Accessed 15 07 2021].

[30] Security intelligence, "Deeplocker Overview Image," [Online]. Available: https://securityintelligence.com/wp-content/uploads/2018/08/deeplocker-overview-chart.png. [Accessed 19 08 2021].

[31] S. Intelligence, "Deeplocker how ai can power a stealthy new breed of malware," [Online]. Available: https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/. [Accessed 27 08 2021].

[32] CNN Travel, "Airports facial recognition," [Online]. Available: https://edition.cnn.com/travel/article/airports-facial-recognition/index.html. [Accessed 12 12 2021].

[33] "What Type of Government Does China Have," [Online]. Available: https://www.worldatlas.com/articles/what-type-of-government-does-china-have.html. [Accessed 20 12 2021].

[34] T. L. P. A. G. Austin Short, "Scientific paper on Defense against adversrial attacks," [Online]. Available: https://www.osti.gov/servlets/purl/1569514. [Accessed 12 11 2021].

[35] Brookings, "Protecting privacy in an ai driven world," [Online]. Available: https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/. [Accessed 19 12 2021].

[36] Forbes, "Ai laws are coming," [Online]. Available: https://www.forbes.com/sites/cognitiveworld/2020/02/20/ai-laws-are-coming/. [Accessed 27 12 2021].

9 - Project Proposal Form

Pearson

Project Proposal form

Learner Name	Vivian Lopez	**Learner number**	5183
Centre Name	St. Olave's Grammar School	**Centre Number**	14285
Teacher Assessor	Mr Jewson	**Date**	21/12/21
Unit	1		
Proposed project title	To What Extent do the Risks of Neural Networks Outweigh the Benefits?		

Section One: Title, objective, responsibilities

Title or working title of project (in the form of a question, commission or design brief)

To What Extent do the Risks of Neural Networks Outweigh the Benefits?

Project objectives (eg, what is the question you want to answer? What do you want to learn how to do? What do you want to find out?):

I want to learn more about neural networks and the ways in which they are utilized in ever

If it is a group project, what will your responsibilities be?

N/A

Section Two: Reasons for choosing this project

Reasons for choosing the project (eg, links to other subjects you are studying, personal interest, future plans, knowledge/skills you want to improve, why the topic is important):

I have an interest in AI and machine learning, I am studying Computer Science and Mathematics which both have links to AI and machine learning with neural networks falling under machine learning.

I enjoy coding as well and am familiar with pre-built frameworks needed to create neural networks.

I intend on studying Computer Science at university. In completing this project, I hope to deepen my knowledge on neural networks and the effects they have in the real world.

Section Three: Activities and timescales	
Activities to be carried out during the project (eg, research, development and analysis of ideas, writing, data collection, numerical analysis, rehearsal techniques, production meetings, production of final outcome, administration, evaluation, preparing for the presentation, etc): An activity that I will carry out is research around the topics, reading articles and books that will help me to gain a deep understanding of neural networks. I will carry out the write-up after completing research. Meetings. Evaluation. Presentation preparation (script, power point, rehearsing)	How long this will take: Research - around 40 hours Write-up - around 20 hours Meetings - 5 hours Evaluation - 10 hours Presentation preparation - 25 hours

Milestone one: First Draft Sumbission

Target date (set by tutor-assessor): 14/12/21 (Before Christmas Holidays)

Milestone two: Presentation

Target date (set by tutor-assessor): 31/01/22

Section Four: Resources

What resources will you need for your research, write up and presentation (eg, libraries, books, journals, equipment, rehearsal space, technology and equipment, venue, physical resources, finance):

I will use books, the Internet (Sites such as BBC,Codecademy), Online lectures and cours

What your areas of research will cover?

I should initially cover research on neural networks, how they are created and some history behind them. I will make sure I know the different types and how they function and the advantages and disadvantages that come with them. I should cover specific examples of both positive and negative instance of neural networks being used which will help me to answer my title question.

Comments and agreement from tutor-assessor

Is the learner taking this project as part of the Diploma? Yes/No

If yes, which Diploma are they taking? _____

Comments (optional):

Is project derived from work which has been/will be submitted for another qualification? Yes/No

Which qualification (title and unit)? _____

Comments (optional):

I confirm that the project is not work which has been or will be submitted for another qualification and is appropriate.

Agreed: (name) 21/12/21 (date)

Comments and agreement from project proposal checker

Comments (optional):

I confirm that the project is appropriate.

Agreed: (name) 21/12/21 (date)

10 - Activity Log

Pearson

Project Activity Log

Learner Name	Vivian Lopez	**Learner number**	5183
Centre Name	St. Olave's Grammar School	**Centre Number**	14285
Unit Name	Dissertation	**Unit number**	1
Teacher Assessor	Mr Jewson		
Proposed project title	To what extent do the risks of neural networks outweigh the benefits?		

This form should be used to record the process of your project and be submitted as evidence with the final piece of work.

You may want to discuss:

- what you have done (e.g., from one week to the next)
- if you are working in a group, what discussions you have had
- any changes that you have or will need to make to your plans
- what resources you have found or hope to find
- what problems you are encountering and how you are solving them
- what you are going to do next

Date	Comments
4/04/21	After deciding on a topic for my EPQ, I started drafting my project proposal form. Because I am planning on studying Computer science at university, and AI was an area in which I wanted to gain more knowledge about, I decided to base my EPQ on this. At this point I wasn't completely sure on the specific areas that I wanted to cover and how I could formulate a question, so I began with 'How has AI evolved over time?'. I went with this as my initial research and after completing the rest of the PPF, I submitted it. I also got notified of my supervisor today. I am planning to start contacting them later when I am in the process of writing my dissertation.
30/06/21	I began my online CodeCademy course on creating neural networks in python using a library called TensorFlow. This began with some brief articles on basics on neural networks, some risks including the black box nature of neural networks and began developing the core principles on perceptrons, leading to simple logic gates. This introduction gave me some background knowledge on perceptrons and helped me begin research as this was the simplest points of neural networks. It also allowed me to include some points on risks of neural networks, namely the black box nature and
15/07/21	I have almost completed the CodeCademy course, I am on the final project which involved classifying images of lungs as having Covid-19, Pneumonia or as healthy. This project ties together all the components of the course, mainly utilizing a Convolutional network which I have decided to include in my research since it through the project I am doing, it seems to be an extremely useful neural network architecture.
20/07/21	I have completed my CodeCademy course and have been able to gain a large amount of knowledge on neural networks and in particular, different types of neural networks. Moreover, I have encountered a range of different use cases of neural networks, specifically convolutional neural networks. These have including mainly classification of images as well as a type of problem called a regression problem which involved predicting a continuous variable which was in my case a grade for students given data. This course taught how neural networks, mainly the coding aspects however, through physically programming networks, I have a good grasp of how they work and function. However, since I want a better understanding, I have also started reading a short book generally on AI titled "Artificial Intelligence a Short Introduction" by Oxford. This gave both a background on the history aspect of neural networks as well as some applications and different fields in which neural networks are used. However mainly the focus was AI in general, so this gave me background on the general field of AI.
21/07/21	Due to the section on dangers of neural networks in the course, I wanted to explore this aspect more and this

	required a change in focus for my research question. Hence, I changed the title from 'How has AI evolved over time?" to "Are Neural Networks Dangerous?". I now submitted an updated version of my PPF with the updated research question and updated fields on the supporting information. The main reason I made this change is because I felt that the question was more targeted, and I could create an argument for it as opposed to generally considering AI and I felt the original question was too broad.

Date	Comments
22/07/21	Due to my recent change in research question, I began researching different risks as this was a big part of my new research question. To do this, I started looking for books however there was limited literature on the risks specifically, so I decided to visit websites and read articles instead as they provided information directly. Alongside this, I wanted to get a deeper understanding on the inner working of neural networks, to do this I started watching videos on a YouTube channel by a creator called '3blue1brown'. One of his videos detailed how neural networks functioned through the example of character recognition which I found to be quite intuitive, so I included this as part of my research. In the video, a link was made to a book titled 'Neural Networks and Deep Learning' by Michael Nielsen. It indicated that it went deep into the fundamentals along with the mathematics of neural networks. This is something I was interested and so I decided to read this next
23/07/21	I started reading the book I outlined in my last record. It goes through the example of character recognition with significantly more detail and incorporates a step-by-step python implementation of the neural network that doesn't use any libraries but instead creates the network from the base up. I am following this as well.
28/07/21	I am almost finished reading the book I started reading in the previous record. The mathematics got to a very complicated stage in the section on backpropagation and gradient descent and I wasn't able to comprehend all of it but I got a gist of what is the goal, the general idea of how it reaches the goal and the general mathematical concepts it uses to reach the goal. These concepts include multivariate calculate, vectors, and different vector operations. I am also following the python program. It is fairly simple to follow. The book also incorporates exercises that test my understanding at different stages. I enjoy some of these, however some can be quite challenging given that my mathematical knowledge it not at the level required. An example of these challenges involves is explaining how a particular activation function works (the sigmoid function). It was interesting to see how graphs and graph sketching fits into neural networks.
30/07/21	I finished reading the book and it has given me a much deeper understanding of neural networks and how they function. This is the level of understanding I was hoping to gain when I started out. I am now continuing to find relevant articles on risks since I spent most of my time on the book and getting to grips with the content on neural networks. Some risks I have found include deepfakes and adversarial attacks which are highly interesting, I haven't known about the technology behind them before. Deepfakes specifically, I learnt through a few different articles on how generative adversarial networks created them. I did come across generative adversarial networks in my CodeCademy course, but this was very brief. Adversarial examples also were intriguing, and I think that they will be a big risk I will definitely include in my dissertation due to the great risk they pose to neural networks.
15/08/21	I have started working on the dissertation today. Beginning with defining key terms. I have started here because I realize that my dissertation will be fairly technical and as a result, defining key terms will be helpful for my assessor. I am drawing on research as well as the knowledge I have gained from completing my course, reading the books as well as articles.
25/08/21	Today I went onto creating a section on explaining how neural networks works. This similar to the last section as I was able to draw on my previous knowledge. I am trying to be concise and clear so that it simple to understand, as this is important so that the function of the neural networks can be clearly explained using this.
10/09/21	I have now moved onto a section on the history of neural networks. I will be explaining how some key components such as perceptrons and backpropagation came about which I hope will help in making these complex ideas and concepts easier to understand. To help me with this section I found some useful articles that described the history, I picked key moments which I thought were important in the development of neural networks such as the development of the perceptron algorithm as well as the development of backpropagation. I decided not to include the complex mathematics as this wasn't the focus of my research question.
5/10/21	This week I decided to complete the risks sections on neural networks, covering deepfakes, adversarial attacks as well as another two risks I found through articles: data poisoning and deep learning-based malware. Before I could explain the risk of deepfakes I explained how generative adversarial networks worked as they were the basis for deepfakes. I tried to explain this as simply as possible like the initial section not to make the description seem too convoluted and complicated as this would have taken away from readability and clarity. Over the course, I also found an interesting real-life example for deepfakes, which involved a significant amount of money being stolen. This highlighted to me raw possibility of deepfakes at their worst and this contributed a significant amount to the risks section on deepfakes. Going onto covering the other risks, I described and explained each of them however, I wasn't able to find solid examples and therefore I decided to research them more.

6/10/21	I found some scientific papers on some of the rest of the risks (adversarial attacks malware and data poisoning). They provided me with more of an objective view of the risks contrasting with the articles that had more of an opinion-based approach to the risks. This allowed me to have confidence in them and their reliability. I am considering adding more risks later on but for now I am going to move onto considering how the risks are being protected against. I am planning on doing a section on each of the risks.
10/10/21	I have added sections on all the different risks on how to protect against them. In addition, I have continued to complete a section on the 'Benefits of Neural Networks Today'. In this section I mentioned some key examples of neural networks uses which I thought were important. To do this, I drew on the research I did earlier on the different types of neural networks.

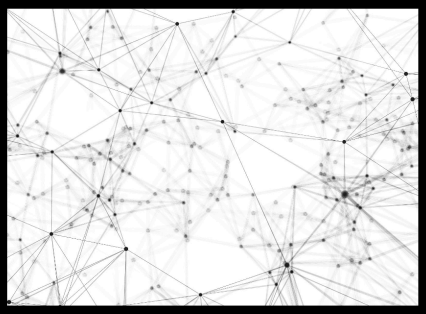
Date	Comments
17/10/21	Today I added an overall judgement on both the risks and benefits came to a conclusion on my research question.
21/10/21	I got in contact with my supervisor today, we scheduled a date for an initial meeting on the EPQ. The meeting was planned to get gauge of my progress and generally plan the dates for the next drafts.
8/11/21	Today I had my first meeting with my supervisor. I brought a copy of my project so far which I got generally positive feedback, with there being a defined structure. We decided on submitting a first draft before the Christmas holidays, so I started planning for this.
14/12/21	Today I submitted the first draft for review. After adding some detail to different sections. Including the risks since this was the main focus of my project.
21/12/21	Today during the holidays, I received detailed feedback on my first draft. Since I was going on holiday. One comment was that the depth varied in different parts, and this is something I realised when comparing my risks to the uses. The content on uses was much thinner so this is something I planned on re-working in terms of structure by adding sections for different networks, describing them, and attributing the relevant uses to them. Also, the history of neural networks section was commented on as being too detailed which was fair given my question was on the dangers of neural networks, hence I cut out some of the sections in it and placed more emphasis on the important parts of the history such as the perceptron algorithm. Also, concerning the structure, my supervisor indicated that uses may be better put along with the history section and this made sense, so I made this change. The risks and threats were described as good, so I didn't plan to spend too much time changing or re-working this section. One point made was that it was confusing whether some threats were hypothetical or not and whether some were currently active threats. To rectify I planned to add and tweaking some section so that it was clearer about what threat they actually posed. I did this by adding information from scientific papers and real-life examples to make the threats clearer, as well as evaluate which weren't significant threats due to the fact that they weren't currently being executed. The explanations for protection against threats was described as thin so I planned on adding to these and I found more articles, websites and papers that would help with this. The final main point was that the question may better be approached by comparing the risks and benefits. I took this suggestion and re-worked the question to: "To what extent to the risks of neural networks outweigh the benefits?". I created the new PPF according to this question.
01/01/22	I finished implementing the changed I planned out in the previous section as well re-structure my dissertation so that it had a better sequence, and I also added an abstract as this was something that was not included in my first draft. I also decided to add diagrams to better illustrate certain structures such as the different architectures of neural networks.
02/01/22	I now started working on my presentation which was due to be on the 31st of January.

27/01/22	I have now completed my presentation; I did this by first writing down a script of what I generally wanted to cover so that I could get my thoughts down and then I used this to create a PowerPoint slide. For the actual presentation I planned to not use flashcards since I was hoping to be confident with the content in my script. I also rehearsed my presentation for the first time, using my script initially to help me build confidence. I included some interactive elements in my presentation, namely a website by CNN which had a 'Spot the Deepfake' activity. I thought this would allow the audience to better engage with the presentation.
31/01/22	I was due to have my presentation but volunteered to have it switched as a result of fewer people on the following day. So, instead I rehearsed my presentation again, at this point I was feeling confident with my presentation.
1/02/22	I gave my presentation today and overall am happy with how it went however I did overrun; this is something I could have prevented if I practiced with some spare time and used better time during rehearsing my presentation however, I feel it didn't significantly affect my performance.
27/02/21	I sumbitted my project today after having updated some formatting and doing a final proof-read.

11 - Presentation Slides

NEURAL NETWORKS

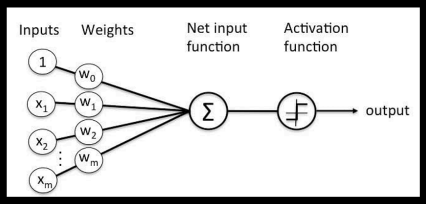
TO WHAT EXTENT DO THE RISKS OUTWEIGH THE BENEFITS?

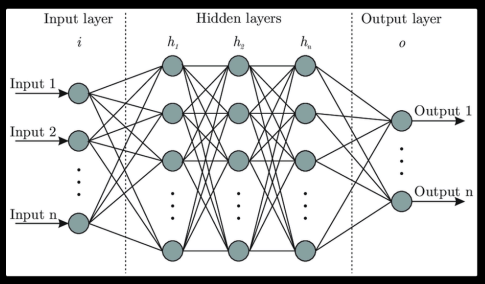


What Will Covered:

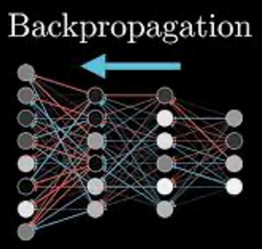
- Neural networks explained through the perceptron model
- Examples of risks in neural networks
  - Deepfakes
  - Adversarial Attacks
  - DeepLocker
- How risks can and are being addressed
- Examples of benefits
- Judgement

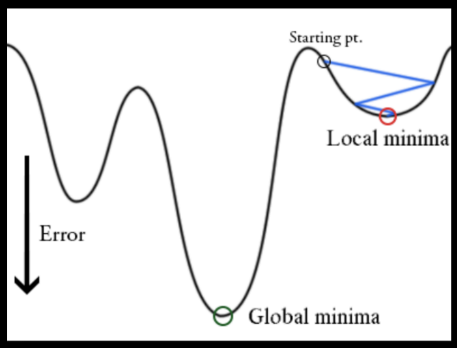
The Perceptron Model



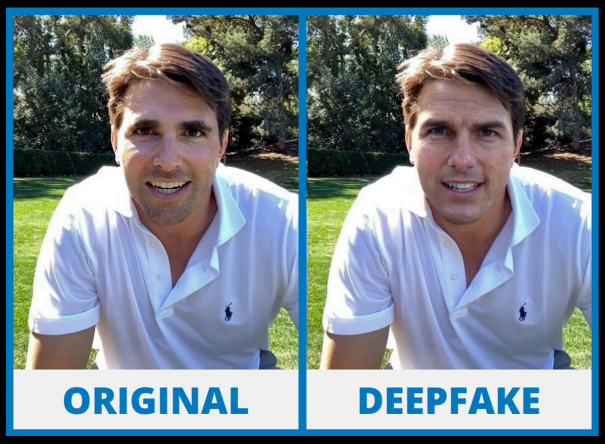


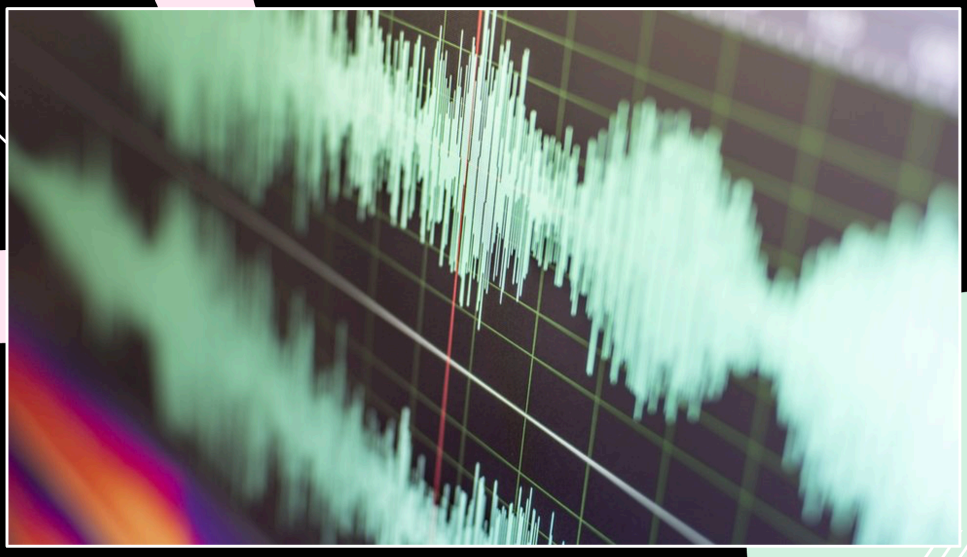
Backpropagation





Risks - Deepfakes

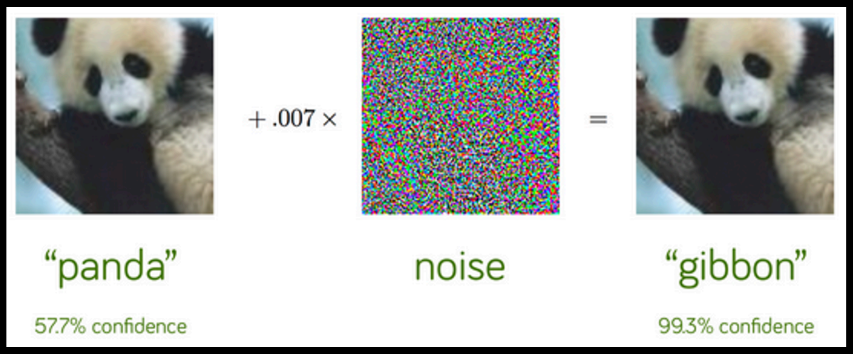




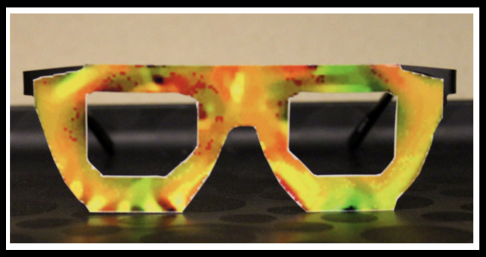
Spot the Deepfake

- https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/

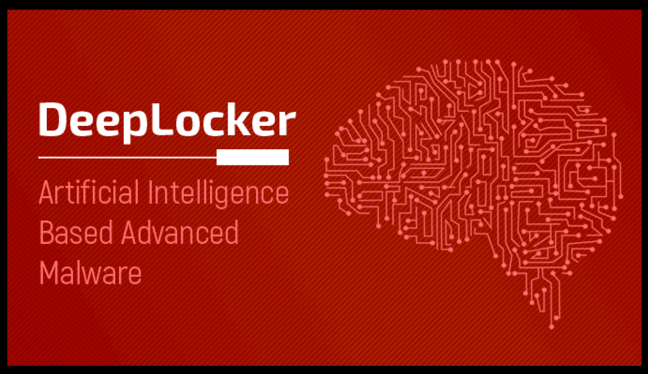
Risks - Adversarial Attacks



Further Examples of Adversarial Attacks



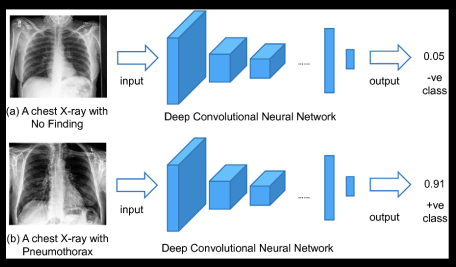
Risks - DeepLocker



How Risks can and are being addressed

- Deepfakes:
  - detecting unusual facial features
  - Standard security practice
- Adversarial Attacks:
  - Scanning for adversarial input and amending, with APE-GAN
  - Penetration Testing of models
  - Lack of proper defense
  - Risk assessment of models
- DeepLocker
  - Flagging unexpected actions taking by applications
  - Monitoring and analyzing data

Examples of Benefits



Judgement

Thank you for listening!

- Any questions?