The ethical conundrum of implementing and developing Neural Network Technologies

What do two opposing normative ethical view's on Neural Networks mean for the implementation and development of the technology?

Authors
Jonatan Langer
Karrar Adam
Kyle Johnson
Matthias Kaas-Mason
Robert Špralja

Supervisor
Sune Thomas Bernth Nielsen

International Bachelor of Natural Science
Roskilde University
December 18, 2020
Total character count (including spaces): 73723

Abstract

This report attempts to analyze ethically problematic cases in two technologies that use Neural Networks as a part of it's decision making process, specifically Facial Recognition and Automated Vehicles. The report covers technical background of both technologies as well as a theoretical background of normative and applied ethics. It goes on to analyze cases of the trolley problem, mundane problems, Bias and Prejudice, and Privacy issues through each theories' perspective. The report goes on to analyze the case studies as a whole through applied ethical principles, and by contrasting the different normative ethical perspectives on specific cases of this technology. This report concludes that the theories imply a need to limit the use of the technology until its further developed, or the ethics for neural networks has been consolidated.

Contents

Chapter 1

Introduction

Neural Networks are an incredibly powerful yet flawed technology. Over the past few years the prevalence of this technology in our daily lives has increased. However unnoticed this technology is, it has had a dramatic impact on us and will continue to do so. Examples of its use is in translation, analytics, and social media recommendations but what we will be focusing on in this paper is the use cases of self-driving cars and facial recognition. Despite the ethical issues this technology faces, there seems to be no slowing down it's progression. Although many companies that use neural networks in their products have ethics councils, the running, decisions and process of these councils is not always transparent.

We will give an overview on the technical challenges and limitations associated with this technology. As well as analyse the use of this technology (in the specific use cases we defined) from the perspective of normative ethics. We will also analyse this technology through the lens of applied normative ethics, as it offers consolidated yet different perspectives to normative ethics.

Since the prevalence of these issues will only increase with the increased usage of this technology, it is vital that we look at the ethical concerns and considerations. In this project we aims to analyse some of these issues from an ethical perspective and try reach a consensus or at least document the differences between how each of these ethic schools of thought view the problems brought about by the usage of the neural networks. Which leads to our research question:

Research Question and Sub-Questions

What do two opposing normative ethical view's on Neural Networks mean for the implementation and development of the technology?

1. *Which aspects of the technology make it ethically questionable?*
2. *What are the differences and similarities on the opposing views of the problems?*

3. *Do the schools agree or disagree whether Neural networks should be implemented? Developed?*

Methodology

To answer this question our report will provide a review of both Artificial Intelligence (specifically Neural Networks) and Ethical Theories. To achieve this we will present two of the main normative Ethical Schools of thought. We will also explain the background behind applied ethics to help us answer this question. We will then cover the technical background behind artificial intelligence, highlighting the technology's technical problems behind the ethical issues. Finally, we go on to use this background knowledge to analyse 2 specific use cases for this technology and their consequences.

The analysis of the two case studies will involve looking at each of the specific problems caused by AI through the lens of each normative ethical theory; As well as applied ethical principles as these provide different (specific) perspectives on the issues. We will analyze the differences for how each of these schools of thought view these problems. Then in the discussion we cover how these normative theories determine what is and isn't safe. We also discuss the ethical principles derived from both of these theories to see whether they determine each case to be safe to implement today. Our conclusion will then surmise whether the theories views on this technology say it is safe to develop, both now and in the future, as well as consolidate the differences between the theories.

Chapter 2

Theory

2.1 Ethical Schools of Thought

Being the social creatures we are, it is not hard to understand that we as humans have tried to create methods for judging an action. Today, the field of ethics involved in the methods of differentiating right from wrong is referred to as Normative ethics [1]. Normative ethics focuses on determining ethical principles for right and wrong. An example of a principle would be "Do unto others what you want done to you". Using solely this principle of normative ethics we could determine if any action is ethically correct. However this is just one example of a normative theory, most other ethical theories generally focus on more than one principles or character trait [1].

This section of the report will cover two opposing schools of thought: Deontological and Consequentialist Ethics [2]. This section will also cover Applied Ethics and normative principles that have been derived from both of these theories.

2.1.1 Duty Ethics Theory

Duty ethics, also called Deontological ethics (Deon meaning duty in Greek), is based on the idea that all actions must conform to moral rules [3]. What is important to note is that these theories ignore the consequences of an action, rather they focus on the action itself. There are three main theories in Deontological ethics: Agent-centered, Patient-centered, and Kant's Categorical imperative [2, 3].

Agent-Centered Theories

Agent-centered theories revolve around the idea that an action or intention from an individual can be moral based on specific obligations or permissions of the individual or "agent" [4]. These obligations and permissions apply to only the agent in question, although others may share the same reasons for an action [2]. The main concept here is that morality is a personal thing, the reasoning behind an action is ones own [3].

There exists three kinds of Agent-Centered theories. The First focuses on the intentions of the agent. For example, in with this theory simply having the intention to do a bad act is bad in itself [2].

The second focuses not on the agents mental state but solely on the cause itself. If the act itself is evil regardless of the intention then it is wrong. For example, killing someone directly is always wrong, however doing nothing to stop someone from dying with this view is not unless you have a relationship with them [2].

The third theory is a combination of the past two. This one focuses on both the intent and cause. For example, this theory says nothing on an act of evil or the intent to commit evil, but rather the execution of an evil intent is evil. Therefore only acting on evil intentions is in itself evil [2].

Patient-centered

Unlike Agent-Centered theories, Patient-centered theories are focused on individual rights [3]. A right is generally defined as a justified claim against another person. For example, an individual has a right to not be harmed by another person [5].

One of the core rights in patient-centered theories is the right not to be used solely as a means for good consequences without ones consent [2]. For example, imagine a case where a doctor can transplant organs from one healthy person to save five sick ones. Patient-centered theories would argue that this is wrong because you are using the one to save the others effectively accelerating the death of the healthy one [2]. In this case the healthy person has the right not to be used without consent so therefore it becomes the doctors duty not to use him [3].

Kant's Categorical Imperative

Immanuel Kant, an 18th century philosopher, believed that there was a core principle that covered all other duties, he called this the "Categorical Imperative". This Imperative differed from other core values because it simply states a pure duty that is good in all context is inherently good. This concept was explained in two formulations of his imperative.

The first formulation of the imperative uses the idea of a Maxim, or a principle for acting in a certain way to achieve a certain goal [6]. The formulation goes as follows, "Act only in accordance with that maxim through which you can at the same time will that it become a universal law." The first formulation essentially means that you cannot make a rule for yourself that is not universally applicable [6, 2].

The second formulation goes as follows, "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end."[6]. This

means that for a principle to be good then it must be practically necessary, with or without it being necessary for the agent to achieve some other end [7]. Essentially, Kant meant that people should always treat people with respect and never as objects, that can be used without consent [2, 6, 3].

2.1.2 Consequentialist Ethics Theory

One common way for people to determine their actions or moral standards is by measuring the consequences of their actions or potential actions. This is essentially an 'ends justify the means' type of ethics. Teleological theories ('Teleo': Greek for end) determine whether an action is morally good or bad by weighing the benefits of said action against the negatives. For instance most of people would agree that lying is wrong but if telling a lie would help save a person's life Consquentialism would argue it is right thing to do. Within this School of thought there are 3 core theories: Egoism, Altruism, and Utilitarianism. Each is based on the frame of reference that we are applying them from.

Ethical Egoism

Ethical Egoism is a widely controversial theory that focuses on ones own interest [8]. Its arguments state that an action is right only if it is acting in the agents best self-interest [5, 8]. A person following solely this theory would always judge an action on this sole principle. If it is advantageous for them to do an action then the correct course is to do it. For example, a child giving a pencil to a classmate (an action that would get praise from a teacher) is argued to be morally right because the child wanted the praise from the teacher. The morally right behavior has nothing to do with the agent in question helping someone, it is simply the fact that the action benefited the agent themselves[8, 5].

Ethical Altruism

On the other end of the spectrum of consequentialist theories lies Ethical Altruism. This theory defines moral behavior to be what is in the best interest of everyone but the agent [5]. Using the same example as the previous subsection: the act of giving the classmate the pencil would instead be morally right because it is the choice with the greatest benefit to everyone except the agent. However, the outcome is exactly the same for everyone but the moral justification for the action is very different [5] (notice how in this case the child still got praise from the teacher, it just isn't a factor in this moral equation).

Utilitarianism

In the middle of the spectrum of Consequentialist theories lies Utilitarianism. This theory can be described as a mixture of both Ethical Altruism and Ethical Egoism [5]. The defining principle in this theory is that peoples actions should always be in the best interest of society [9] (the agent and everyone else).

Take into consideration the example of the child giving their pencil to their classmate, this time, from the Utilitarianism perspective. In the scenario the child has two choices either to help or not to. In the case that they helps their

classmate, both will benefit from this action (one receiving a pencil and the other a higher standing in the eyes of the teacher), but the child temporarily has less pencils to use. In the case that they choose to keep the pencil, the child will have a surplus of pencils but nobody else will benefit from this action. In weighing the consequences for the agent and all others; the first case has the greatest benefit to society so it is the morally correct action [5, 9].

2.1.3 Applied Ethics

In Section 2.1.2 we highlighted an example of a child giving away a pencil to a classmate. This example is straightforward and it is very easy to see the morally correct action from all three perspectives. However, in the real world the problems are much more complex, and the morally correct action is not always clear. Applied ethics is the process of analyzing these issues with overarching normative principles [5].

Ethical principles Applied in real life

Generally, complex and controversial issues will have a very clear solution when only applying a single principle to them. However these principles will have hundreds of principles supporting them and hundreds opposing them, and hundreds more that just look at it from a different perspective. Therefore the solution generally is to use a few of the overarching normative principles to analyze it and see what is most strongly supported [5]. In our analysis we will use the following principles to determine if Neural Networks for pattern recognition systems is safe from ethical controversy today.

- Personal Benefit: principle of agent benefits
- Social Benefit: principle of benefits for all of society
- Benevolence: principle of helping those in need
- Harm: principle to do no harm
- Autonomy: Freedom to have control over one's own body and actions
- Precautionary: Principle to not implement a technology until it is safe

2.2 Theory of AI

2.2.1 What is AI?

Artificial Intelligence (AI), is a subset of software that attempts to create intelligent agents that can make decisions or perform tasks that usually require human intelligence to perform. An example of this is, video game AI, requires creating a bot that mimics human actions and movements inside of the game, but these bots are not necessarily intelligent [10] (meaning that they don't reason and rather just perform the tasks it is programmed to do). This technology is generally split into 2 categories: Strong AI (AGI: Artificial General Intelligence) and Weak AI (ANI: Artificial Narrow Intelligence) [10].

Artificial Intelligence is already being used to make important decisions, analyse data, and interface with humans, among many other important tasks. The advantage of using this technology is how fast and automatic it is, such that an AI can often outperform humans in any task it is trained or programmed to do [11].

2.2.2 Different types of AI

Weak AI is the only type of AI we have managed to create. And it is what we will be talking about in this paper [10]. Weak AI just tries to mimic human intelligence. Note that the definition of weak or strong isn't based on the processing power or efficiency of the AI, and rather its capabilities. Rather weak or narrow refers to implementing an AI for a narrow task or set of tasks that can be performed by a human [10]. This means that even though chess engines can beat the strongest grand masters with ease, they are still weak AI. This is because they lack the emotions or creativity of a human, and rather operates by generating every possible state the board can be in and assessing whether it is a winning position before making a move.

Strong Artificial Intelligence, is only theoretical, but would be an actual machine representation of a human [12]. With the ability to reason, learn, think and understand to a degree that is indistinguishable from a human. In creating this programmers would be creating a machine that is conscious. While this is an extremely interesting topic, it is beyond the scope of this project, so we will leave the discussion here.

Within weak AI, there are lots of different techniques and technologies used. For example, Natural language processing (chat bots) involves matching words, phrases and sentence structures with a predefined structures to choose or build responses from. Contrasted a chess AI that instead performs rigorous calculations to determine every possible state the board could be in and calculate the best move. Therefore AI is not so much a singular algorithm/technique and more of a general goal that can be achieved in any way possible [10]. One of these programming patterns for Artificial Intelligence is called a Neural Network [12].

2.2.3 Neural Networks

Artificial Neural Networks (ANNs) are an attempt to replicate nature's own super computer, our brains. Modelled after the neurons in our brain, neural networks enable a computer to learn from data given to it [12]. They are an incredibly powerful tool that can be (and are sometimes required) to solve certain problems. While a neural network can be trained to perform any task that can be approximated by a function (something that maps a series of inputs to an output), they are not necessarily accurate a 100% of the time [11]. We will explore how this technology works, as well as what the drawbacks are and how they arise.

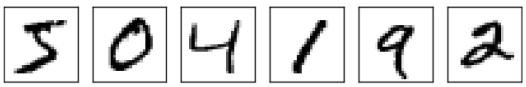

Figure 2.1: Examples of Handwriting
[12]

A great example of how neural networks are useful is handwriting recognition, specifically, we will be looking at recognizing the digits from 0-9 (The same topic [12] uses to introduce neural networks, the video [13] is based on this paper). While it might seem trivial for us to recognize that the first image in figure 2.1 is a 5 and the last is a 2, when you try programming a computer to recognize these things you'll soon be in a hell filled with exceptions, incorrect rules and every other problem imaginable, unless you use a neural network [12]. This is because they can be trained to recognize the digits with actual examples, rather than having to hard code every rule into place [12].

Our brain contains millions of neurons (hence the name neural network), and while they aren't all connected directly together they are all indirectly connected. And it is very much the same with Figure 2.2 and neural networks in general, where you can see there are multiple layers of nodes (the dots) connected by vertices (black arrows) forming a network of interconnected neurons [13]. In any neural network, every node contains a value (usually between 0 and 1), and every vertex has a weight [12, 11]. The amount of layers and nodes in each layer are somewhat arbitrary decisions [13, 12] (but we will discuss this further in the sub chapter 'deep neural networks'), while the most basic neural networks only require 3 layers (input, hidden and output), it is often up to the programmers to adjust these values to get the desired result [11].

When the network is told to make a decision about something it is fed information about what it will make a decision on through it's input layer. This means that somehow a qualitative thing (like a picture) needs to be represented in a concrete, discrete, structure (RGB/grey scale values for each pixel). Once the data is loaded into the input layer, the values for each node in the next (hidden) layer are determined, followed by the next until the output layer is reached. The output of the neural network is generally determined by analysing

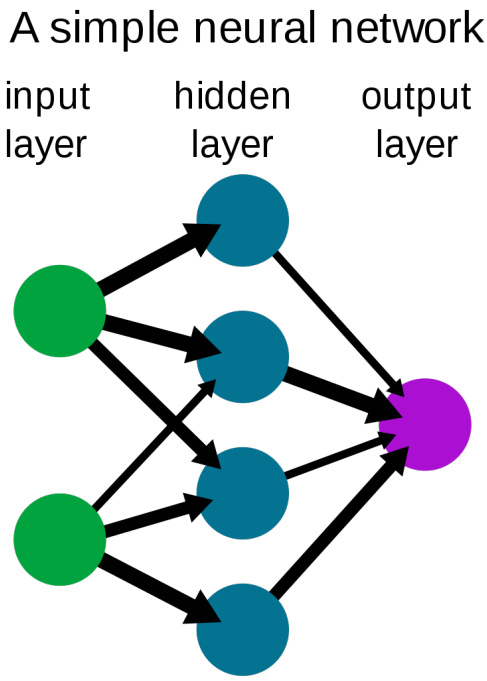

Figure 2.2: A simplified neural network [14]

the output layer, usually each node in this layer is associated (externally) with a possible output (i.e. 10 output nodes, one for each digit that the network can identify) [12]. Usually the output node with the highest value is selected, the associated value with this node therefore becomes the output of the neural network [15].

For our example, we will be assuming the numbers have been written on 28x28 pixel canvas, and it will be a grey scale image (black and white) [12]. This means there are 784 nodes in its input layer (one for each pixel), any node can take on a value between 0 and 1, 0 being black and 1 being white [12, 13]. If this was a color image we would have 784x3 input nodes, as each pixel has 3 values one for the red, green and blue density. The book this example is taken from somewhat arbitrarily choose 2 hidden layers, each with 16 nodes, and our output layer will contain 10 nodes (one for each digit that can be recognized) [13]. This can be seen in figure 2.3.

In figure 2.3 you can see that any single node in one layer is connected to every node in the next layer (to the right). Every one of these vertices also has a value associated with it, the "weight" of that vertex [12]. In 2.3 the different colors for every vertex represents different weights [13]. Additionally, in each node there is a "bias" which acts as a kind of threshold for how much a neuron needs to be activated before it actually contains a value higher than 0 [12].

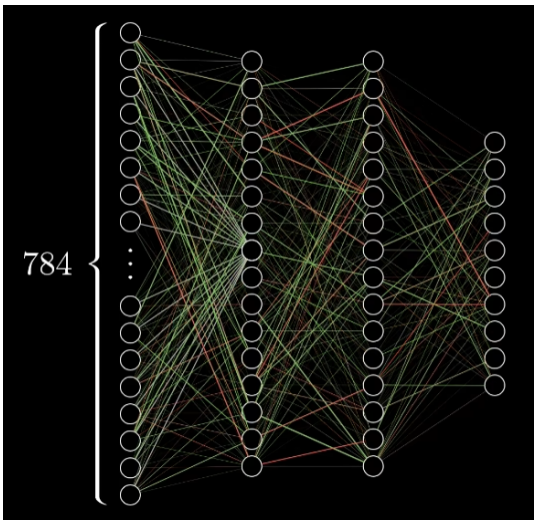

Figure 2.3: An Actual Neural Network
[13]

The way the neural network gets its result is by using each of the input nodes to determine the value for every node in the next layer (to the right). This process is then repeated, using the just determined layer's values to determine the values for each node in the next layer [12]. This kind of numerical shuffle ends at the output node where the output is selected. The actual process of determining the value for an individual node is beyond the scope of this project, but the way it works is by summing the value of every node in the previous layer, multiplied by the weight of the connection between that node and the one we are determining the value for [12, 13, 11]. From this you can hopefully see that the process of machine learning is really just tuning these weights and biases to get a sensible result [10].

Another way of thinking about the adjusting weights and biases processes is: For example, if a person were deciding to go to a movie, they might take whether their partner was going with them and if they enjoy that type of movie into account. Any given person might value these things differently, and be more affected by a certain factor than others. This is analogous to the action of neurons in a neural network because, the neurons value (the person's choice) is determined by the weights (how much they value a factor) and the values of the previous layer of neurons (the actual factors affecting their decision). Then by adjusting the weights between these inputs and the output we can end up with a different range of results depending on each persons network. So, if going with your partner mattered a lot then the weight for this node will be higher than the one for liking the genre of movie, and it plays a larger role in the decision making process [12].

2.2.4 Machine Learning

As previously explained machine learning is less of the sci-fi definition of autonomous machines acting and behaving like humans, and more of the very real process of tuning each of the weights and biases within a neural network. This is done through the use of extremely large sets of data (a data set of over 10,000 images was used to train the neural network in 2.3 to get it to recognize images [13]). An alternative approach to machine learning is that during operation the AI can train itself, by analysing its own performance. We will go into more detail on this further on.

These large data sets contain a possible input, as well as the expected output (for that input), when the neural network calculates its result it is then compared with the expected output [15]. After applying one of several different techniques (all of which are essentially just function approximators of different complexities: see Newton-Raphson Method, or Neuroevolution [13, 12]), the network's weights and biases are adjusted and then tested again. For example the initial training for Google Translate's AI took just over 10 days on an 8 core Centeral Processing Unit (CPU) [12].

2.2.5 Training Neural Networks

Neural networks, can do anything, so just how do we get them to do one thing? We need to train them. What this means is that we give the network some data, and allow it to produce an output, based on it's output we then go over the network again and tweak all the weights and biases accordingly [12]. This method is called back propagation, because it involves propagating the result backwards, and tuning the weights based on what we expected to see [12]. The technical details on how this is actually implemented is beyond the scope of this project. Back propagation is a form of gradient descent that aims to lower a loss function (a measure of accuracy of the network) [11], and thereby increase the accuracy of the network. However there are several ways this can be achieved [15], and depending on what data is available and what the task is different techniques are used:

Supervised Learning

This type of learning, is based off of data that has already been labelled by a human with the correct output. The data is fed into the network, and then the network's answer is checked against the human's, and the results are back-propagated, to retrain the algorithm [11]. This type of learning is best for classification and regression, therefore is best when there is a set of reference points, from which the network can learn [15].

Unsupervised Learning

Unsupervised learning is the polar opposite of supervised learning, in this case the data is unlabelled and/or uncategorized (for example when training an algorithm without knowing the answer) [11]. The neural network is trained on data without a predefined answer, and it tries to find structure in the data without a specific outcome in mind [15]. This can be trained to do several different things,

for example: clustering (grouping data together), anomaly detection (detecting data that is different), and association (relating data to other data) [15]. The problem with this kind of training is that there is no predefined truth, so it is difficult to measure the accuracy of the network, but it is useful in the case that data is hard to get or very expensive [11].

Semi-Supervised Learning

Just like the name sounds this method of training is based on having both labelled data and unlabelled data [15]. This type of learning is used when good labelled data is hard (but not impossible) to get. For example, a radiologist could label a few MRI's (as it would be expensive and time consuming to do many) that can be used to train a neural network (along with other unlabelled MRI's) [15]. This small amount of labelled data can improve the networks accuracy when compared to unsupervised learning, but still has the same drawbacks.

Reinforcement Learning

Reinforcement learning is used when the neural network is trained with another algorithm/method that evaluates the success of the neural network or the neural network's choice (note that it doesn't necessarily operate in this order) [11]. For example, when training neural network to play chess, it's very hard to get human labelled data, and rather you can have it play against itself and have it tweak itself based on when/how the algorithm wins. This is useful in the case that you can evaluate the network's performance but it is hard/impossible to get training data (for example in a game you can't necessarily create training data on what the best move is, but you can analyse the algorithm's move once it has been made) [15].

2.2.6 Deep Neural Networks and Deep Learning

There are several different types of neural network, and each is called a different thing depending on how it acts, or works. However, a deep neural network is generally defined as a neural network with several layers. This is often paired with the term 'deep learning' which encompasses several techniques and algorithms that can be used to train these deep neural networks.

These deep neural networks use several hidden layers, as it allows for more abstraction, and are therefore potentially/theoretically more accurate and faster (as opposed to those with fewer layers with as meany neurons as necessary) [12]. It has also been mathematically proven that in some specific circumstances a shallow network requires exponentially more nodes than deep networks [12].

Most commercial products use deep neural networks as opposed to shallow ones because they allow for more abstraction and therefore operate more reliably. In the example of writing recognition (figure 2.3 from [13]), each consecutive hidden layer has a more abstract and specific job, that eventually leads to the abstraction of the result [12]. For example the first hidden layer could recognize different edges (curved line, short line, etc), and the second would recognize different shapes those edges make when put together (circle, cross, etc), leading

to the output layer where the different shapes are put together and turned into a result [12].

There are however drawbacks to using deep neural networks. Because of the many layers, we find that our typical learning algorithms (gradient descent, and back propagation) aren't training the layers at equal speeds (the later layers might train quickly while the earlier ones get stuck and seemingly learn nothing, or vice versa). This causes the deep networks to operate at the same level as their shallow counterparts [12]. The way we overcome this is through deep learning, which is a set of techniques we can use to make sure the layers are all trained [12].

2.2.7 Pros and Cons to this technology

Neural networks are extremely flexible and can be trained to perform tasks that cannot be easily programmed, and any numerical data can be used to train it for any task that has such data. Its particularly useful when there are a lot of inputs (for example an image), because of how it can split tasks into a network of simpler calculations it is also reasonably fast once trained. However, there are also issues with using this type of technology.

Training and Accuracy

The process of training a neural network can be very Central Processing Unit (CPU) intensive, this means it can be both expensive and slow to train a neural network properly [11]. Unlike traditional programming neural networks can make unpredictable mistakes, either because of how it is fitting the data (which we will cover shortly), because it wasn't trained enough, or because it just made a mistake. For example Google researchers fooled a image classifying AI that could successfully analyse pictures, by just changing a few pixels of that image [16].

The core issue with using training data is that the training data has a direct effect on the performance of the network: so if the data is flawed this flaw is passed on to the network [12]. And while one possible solution is to provide more training data (or more accurate/appropriate data), there are also drawbacks to this. This is because the training data needs to be created/analysed by a human before the network can operate on it, making it costly to attain huge amounts of data (assuming we need a lot of labelled data) [15]. The other issue is that the more data that is used, the longer the training process will take [11]. These factors mean there are disincentives to training a network more than appears necessary.

Overfitting and Underfitting Data

The problems with data do not end with just what the data is, but also how the network interprets this data when training it. This issue is known as underfitting or overfitting the data [11]. Overfitting is when the network is too complex and follows the training data too much. This means it could be interpreting random patterns in the training data rather than the intended pattern/result [12]. A

more technical definition for this is that the network has a high accuracy score on the training data but a low accuracy score on test data [11, 16][1]. Underfitting is a similar but opposite problem, in this case the network is too simple to pick up the underlying pattern in the data [12]. This can be caused by the algorithm making incorrect assumptions about the data (oversimplifying it). The technical definition for underfitting is: testing the network and seeing it has low accuracy on the training and test data [11].

Black Boxes

The final core issue of neural networks is that they are black boxes [11], this means that while we understand the core technology behind how it works, and how we change the network's values. But we do not understand why the biases and weights are assigned their specific values or how we can tweak these values ourselves to achieve specific results. This also means that we can not know how each independent variable (pixels in a canvas) affects the dependant variable (what number the neural network chooses) [11].

[1] Here accuracy is a measure of how close the Neural Network's result is to the expected result

Chapter 3

Case Studies

3.1 Automated Vehicles

While attempts at automated vehicles could have already started in the 1920s [17], they didn't start becoming a reality until recently. Now automated vehicles across the world have driven millions of kilometres [18]. In the future automated vehicles might revolutionize transportation and lead to lower emissions, faster transportation, and even lives saved. Lightly automated vehicles are already used every day, but more advanced driverless versions are yet to be available commercially.

3.1.1 Theory

Automated vehicles are categorised into 6 levels by the Society of Automotive Engineers (from level 0-5). Vehicles designated as level 0 are not automated what so ever. Vehicles designated as level 1 can offer lateral or longitudinal support motion. Conversely, vehicles designated as level 2 can offer both lateral and longitudinal support motion. Furthermore, level 3 vehicles can perform the whole dynamic driving task in suitable condition. If it anticipates that the conditions won't be suitable in the near future it issues a request to intervene in appropriate time. Vehicles designated as level 4, can perform the entire driving task in suitable conditions, and if the driver is unresponsive it will achieve a minimal risk condition. It may even delay the human's request to drive manually to achieve a minimal risk condition. Vehicles designated as level 5 can perform the entire dynamic driving task, and can drive without a human back-up driver i. e. are driverless. We are mostly concerned with level 4 and 5 automated vehicles [19].

Automated vehicles operate with two main systems: the perception system and the decision making system. These two main systems are further divided into many subsystems.As you can see in Figure 3.1.1 the the perception system send information to the decision making system which operates in sequence. We will briefly explain each subsystem:

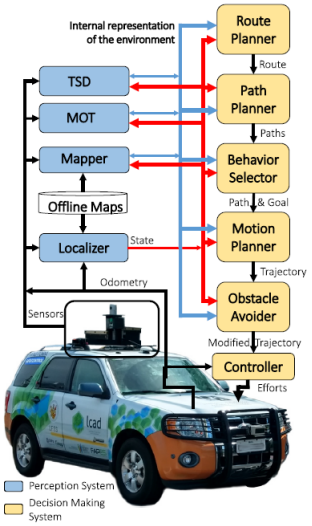

Figure 3.1: Example of an outline of a the subsystems of an automated vehicle. TSD denotes Traffic Signalisation Detection and MOT denotes Moving Object Tracking [20].

Localizer

The localizer is responsible for estimating the vehicles orientation and position relative to a road or map [20].

Mapper

The mapper is responsible for computing maps from offline or online map databases, it is essential for the vehicle not to collide with static obstacles such as sign posts or pavements [20].

Moving obstacle tracking

The MOT is responsible for the detection and tracking of moving obstacles. This subsystem is essential to for the avoidance of moving vehicles or other moving objects [20]. Huval et al. propose a neural network based approach at obstacle tracking [21].

Traffic signalisation detection

The TSD is responsible for detecting a recognising traffic lights, signs and other parts of the road like pavements [20]. This can be done by using neural networks specialising in detecting traffic lights, signs or pavements, like it was done by Liu et al. [22].

Route planner

The route planner is responsible with the computation of a route from the vehicles initial position to its destination, this is done using various path-finding

algorithms [20].

Path planner

The path planner is responsible with the computation of a set of possible paths which are part of the route, this can also be done using various path-finding algorithms [20].

Behaviour selector

The behavior selector is responsible for choosing the current behavior of the vehicle. For example choosing the lane, traversing an intersection, stopping at a traffic light or stop sign [20].

Motion planner

The motion planner is responsible with the computation of the trajectory of the vehicle, where the trajectory must follow the path, it simulates the physical motion of the car so it can be adjusted [20].

Obstacle avoider

The obstacle avoider is responsible The obstacle avoidance and control subsystem or the obstacle avoider is responsible with avoiding obstacles. It adjusts the trajectory if obstacle avoiding is necessary [20].

Controller

The controller is responsible actually responsible with operating the vehicles controls, like adjusting speed, or turning to keep true to the trajectory provided by the motion planner and obstacle avoider [20]. This can possibly be done with neural networks. For example as proposed by Guidolini et al. [23].

From the given examples we can see that some implementations of automated vehicles use neural networks for their perception subsystems and that they also could be used in the controller subsystem in the decision making system.

3.1.2 Benefits

According to an estimation by the World Health Organisation more than a million lives are lost in road incidents, and more than 20 million are injured globally. Of these 93% are caused by human error. So, because automated vehicles avoid human error (like drunk driving), wide adoption of automated vehicles would lead to millions of lives being saved [24]. Because automated vehicles can also drive more efficiently, a wide adoption of automated vehicles will lead to a lesser environmental impact because of the added efficiency [25]. Furthermore, because many traffic jams can be averted with automated vehicles, because all the vehicles could for example start moving at the same time, significantly reducing the jam. Simulation research has also shown that even with 5% vehicles being automated, a reduction in traffic jams is expected [26].

3.1.3 Ethical Issues Arising From this Technology

The Trolley Problem

Automated vehicles pose ethical concerns. Let us take this situation as an example: An automated vehicle with a passenger is in a situation where a crash is imminent, but the autonomous system has a choice between killing a pedestrian or driving the vehicle into a wall killing the passenger (as in the Figure 3.2 [27]). What is the automated system to choose? What if there are two pedestrians, or multiple people in the vehicle? The car might choose to kill less people but that might be unappealing to customers, causing them to buy the vehicle from a competitor, so surely the vehicle should be programmed to always protect the passengers first. But what if the pedestrians have a relationship with the passengers. For example what if two parents are in the vehicles and their 2 children are the pedestrians. The vehicle does not know who the pedestrians are, and might kill them, whereas a human driver probably would not. This problem commonly known as the Trolley problem [28, 29] might require the automated vehicles to have a moral system implemented.

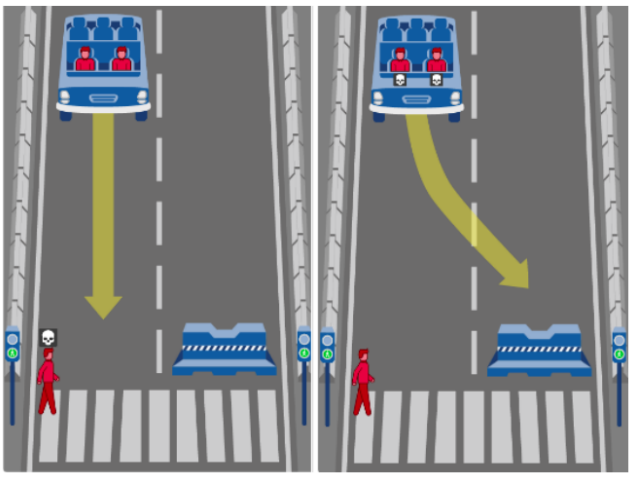

Figure 3.2: Trolley Problem has a choice between killing a pedestrian or driving the vehicle into a barrier killing the passenger [27]

The Mundane Problem

Now let's take another example: An automated vehicle is approaching a cross walk with limited visibility. There is a chance that a pedestrian might cross the street and cause a high risk situation. The vehicles can be programmed to always stop on cross walks but that will cause much time to be wasted which is in itself unambiguously morally negative (both because it wastes time of the passengers, but presumably more energy will be expended while the vehicle is idling then if it wouldn't have stopped). This kind of problem is intuitive to human drivers, but what is intuitive to humans is hard to implement into machines. This kind of problem is called the mundane problem according to Himmelreich [30]. While mundane problems are mundane, they occur much more frequently than trolley

problems and as such should be considered. More generally mundane problems are problems which are intuitive to human drivers, but when automated vehicles are applied to them their decisions can have significant effects both because they occur very frequently and because there will be many other vehicles with the same system in place.

3.2 Face Recognition

The Technology of facial recognition has gotten more and more advanced over the last few decades. It has reached a point where the technology is being implemented in security-law enforcement sector, the Health sector as well as in the Advertisement sector. The similarity between all of these uses and versions of the technology is the use of DeepLearning with Artificial intelligence. Using DeepLearning AI, companies like Facebook, Google, and Amazon have managed to create systems that can identify people in various situations with near human accuracy [31].

As this technology gets better the demand for it will also increase, an example of this is in airports all around the world. With the amount of people traveling across borders increasing, it is expected that by 2030 there will be 720 million people traveling by air in the European Union [32]. With numbers like this it can be seen why facial recognition is being implemented but this brings up problems with using this technology in this type of environment.

3.2.1 How it works

With the growth of this technology comes an increased interest in the way Facial recognition functions. Facial Recognition as well as all bio-metric identification technologies have two main methods of operation: obtrusive and unobtrusive [33]. The difference between the two lies in the level of interaction from an individual in order for it to function properly. In an obtrusive bio-metric identification system the user must stop what they are doing and voluntarily provide the system with information. An example of this is a retinal scanner, where the user must stop what they are doing and position themselves to allow the technology to scan their retina.

An unobtrusive bio-metric identification system is much more interesting as it requires the use of machine learning and artificial intelligence in its operation. In an unobtrusive bio-metric system the user is not required to consciously input any information. Using artificial intelligence the facial recognition technology can characterize an appearance by creating weighing a face print of a target against trained data to find the most probable match without the need for the target to stop their activity [34]. This technology is usefully installed in many airports and other high traffic areas [33].

The Face Print

Every persons face is a unique combination of shape, color, and topography. For the average human brain, identification of a person's unique facial pattern is a hardwired ability. However, for a machine to do this same task it is necessary for it to learn the intricacies of many faces. In order for it to do this the Neural network being used is taught to identify key features of the face using databases of images in a process called facial landmark recognition [34].

In facial landmark recognition the Neural network assigns annotations to key features such as eye fissures, lip shape, and eyebrow size to name a few. Once the Neural Network is able to autonomously identify these landmarks it is then capable of facial detection. For this technology to be useful, however, the Neural Network must be capable of both facial alignment and facial recognition [34].

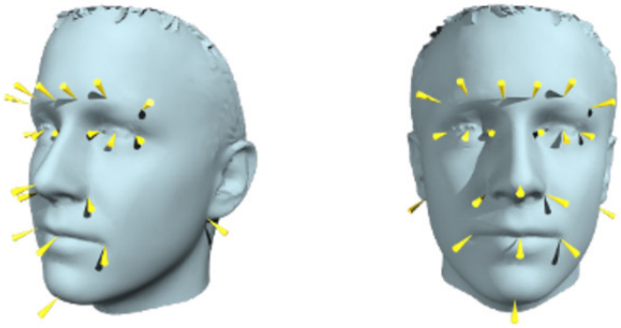

Figure 3.3: Annotated landmarks using Facial Landmark Recognition [34]

Facial Alignment and Recognition

Once the Neural Network is capable of facial detection the process of facial alignment becomes key. In a standard neural network the process of facial alignment requires the use of facial landmarks identified autonomously beforehand. These landmarks can then be manipulated against known face prints from databases to align the faces, this being in the case where the target face print and the comparison face print are in a different pose or expression relative to each other [34].

Using this alignment technique it is now possible for the Neural network to compare the target face print against a database of known face prints each indexed with a discrete value. The Neural network will then return a probability distribution of confidence over the indexed face prints. This percentage represents the confidence the Neural network places on a specific discrete value after weighing the face prints against each other. If the confidence percentage is high enough then the target can be assumed to be a match [35].

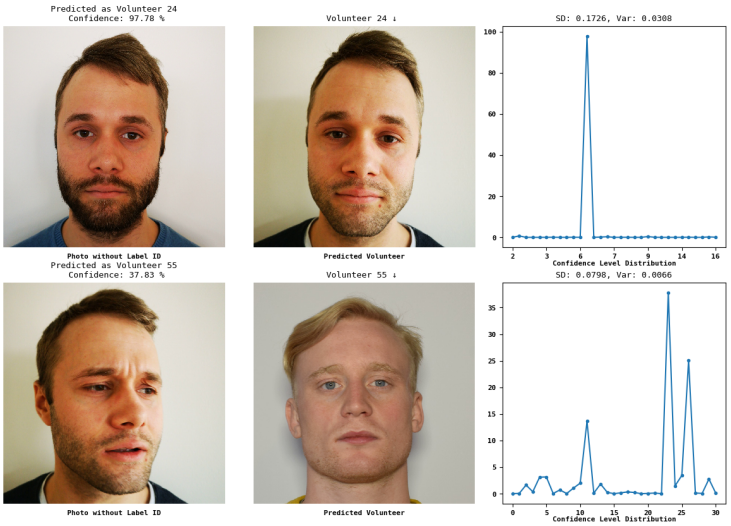

Figure 3.4: Probability Distributions in a match case and non-match case [35]

3.2.2 Tech Issues

Being a rather new technology it is not hard too understand that Facial Recognition technologies are not perfect. In fact, there are a few technical issues that can lead to larger problems.

Error

Before we get into the problem it is important to explain the different possible outcomes a facial recognition software can come to. These are one of 4 decisions: 1. correct positive 2. correct negative 3. false positive 4. false negative. A correct positive is when the system matches a face with the right identification it is looking for, this could be correctly identifying someone in an airport as someone with a criminal record. A correct negative is the same as a correct positive just that it is correctly identifying that this person does not have a criminal record. A false positive is much more important as it is when someone is misidentified as someone else, this could be when someone is identified to the system to have a criminal record when in fact they are not the person it matched them too. A false negative can be more dangerous in this case as it is when someone who is meant to be identified is not, this could be someone with a criminal record is not identified as that person.

Airports are places where many people from all around the world travel through, meaning there is a very diverse group of people this technology is used on. This creates a problem because "Across demographics, false positives rates often vary by factors of 10 to beyond 100 times" according to a study by the National Institute of Standards and Technology (NIST) on 8.5 million people using 189 different commercial algorithms [36].

Function creep

The underlying causes behind errors such as those mentioned above could be explained by the phenomena of function creep. Function creep is when a program designed for a limited task set gains additional unanticipated tasks [37].

In Facial Recognition Technology based in Neural Networks, Function creep can occur a number of ways. Firstly, by expanding the data that the Neural Network has access to. Secondly, by expanding the purpose of the technology to include new tasks rather than recognition. Thirdly, by giving new users access to the technology resulting in various uses that may differ from the programs original intent [37].

In general any change or update to the programs function will result in the program taking on unanticipated methods and biases in its analysis of data. This brings up problems from certain ethical perspectives that will be discussed in the next chapter.

3.2.3 Benefits

The main benefit in the use of facial recognition technology is the security it provides in public and private spaces. It is commonly used in preventing violent crime by analyzing behavior and flagging people that act suspiciously. It is also used in investigating and preventing acts of terrorism. Facial Recognition technologies have been implemented to watch large events and crowds (especially in airports) specifically for this purpose. The technology can also help identify and locate wanted persons and known suspects. A rather unknown use is its ability to help located missing persons. This tool is commonly implemented by police forces and other agencies to great benefit [38].

3.2.4 Ethical Problems Arising from this Technology

Privacy and Surveillance

For facial recognition software to work it needs to have a large database of faces so that the program can learn and compare. This means that companies or governments are motivated to store large amounts of information about people so that their system will work as well as possible. Some of these databases are publicly available and some are kept private for the companies that use them. As we have seen in the past many private databases have had breaches and now more and larger databases are made than ever before.

Recently, the Chinese Government has started a social credit system that tracks citizens data and movement to determine a "Credit Score" for individual citizens. This program is being developed and expanded currently with 200 million surveillance cameras though out the country in 2018 to 626 million cameras being installed by the end of 2020 [39].

Bias & Prejudice

In orthodox surveillance systems without the use of Neural Networks or Pattern recognition, the system will require a human to be watching the screens. A human will have innate preconceptions about potential suspects that he sees. These biases and prejudices are generally referred to as operator bias [40].

On the other hand, In Facial Recognition using artificial intelligence the operator is not present as the process is unsupervised. This would lead to the assumption that bias is not present in this technology. However, the base of Neural Networks is code and an authors values and biases will be present in the written code [41]. This bias is often referred to as author error [40].

This operator bias is also present in the Neural Network's training data. In the last decade there have been many cases exhibiting racial bias by facial recognition software [42]. This is mostly due to function creep. As seen in section 3.2.2, when expanding the purpose of the system unexpected errors can arise from it. If the system has been trained from a database of predominantly Caucasian faces it will have a harder time identifying faces of other ethnicity's [42].

Chapter 4

Analysis

This analysis of Neural Networks is based on several key assumptions. Firstly, the agent in every case we are analyzing is the system using the Neural Network, with the exception of the applied normative principles section below. Secondly, each theory used is considered equal to the others. Thirdly, The normative principles' collective opinion, chosen in Section 2.1.3, supersede those of the individual theories. These Assumptions and the restrictions that result from them will be discussed in Chapter 5.1.

4.1 Deontological view

4.1.1 Agent-centered

As covered in Section 2.1.1, Agent-centered duty theories focus on the intent and cause of the agent in question.

Trolley Problem

According to [29], the driver of a trolley would be acting immorally whether they chose to switch to a track with fewer people or not. Therefore, it would seem that there is no morally correct course of action. However, according to [2], an agent may have relative reasons to be permitted to do a specific action that others may not be. Therefore the Agent-centered theory would say that the intent of the vehicle should be to protect any passengers, thus killing bystanders. Due to the existing relationship between the car owner, or user, and the vehicle.

Mundane Problem

In the example when approaching a cross-walk with limited visibility, the car must slow down with the intention to preserve the lives of possible pedestrians who could cross. By doing this it would waste the users time, but with the intention to not kill.

Bias and Prejudice in facial recognition

According to [43], an agent would not be causing an evil when their act simply enables some other agent to cause an evil. Biases that are inherent in the

program then become defined as unintentional. In the surveillance systems described in [38] defined in Section 3.2.3, the program simply flags an individual as a potential suspect. It is up to police and other individuals to follow up on this information. The agent is not forbidden from selecting a person to then potentially be harmed [44], because whether the person is harmed or not is not the agent's decision.

Privacy

The problem of privacy becomes an issue ethically because the agent is violating the privacy of all those under surveillance to collect information. In the case explained in Section 3.2.4 the Social Credit system attempts to give value to peoples actions based on what the system sees and defines as morally good. According to [2], an agents obligation is based on intended causes. In the case of the credit system the program intends for the information gathered to be judged. Again in this situation the program is not acting immorally for selecting a person to be judged.

4.1.2 Patient-centered

As covered in section 2.1.1, Patient-centered duty theories focus on individual's rights and not on the intention.

Trolley Problem

According to [2], Patient-centered theories justify the action based on whether another person's body, labor, talents are used without their consent. In the case of the Trolley problem the vehicle is acting immorally whether it chooses to continue into bystanders or crash itself to save them. This is because the vehicle would be using either the passengers life or the bystanders'. Therefore the only morally correct act is on the condition the vehicle had no passengers and the crash would not use anyone's body, labor, or talent.

Mundane Problem

In the example of approaching a cross-walk with limited visibility the vehicle ought to again slow down, but with greater emphasis of achieving a balance between not doing harm, and not wasting the user's time.

Bias and Prejudice

Facial Recognition technologies today are not perfect, see Section 3.2.4. According to [5] an individuals right implies another individuals duty to protect it. People have a right against harassment, discrimination and prejudice therefore it is the systems duty to protect people from that. However, since the system is inherently biased as of today the Facial Recognition technology must be acting immorally.

Privacy

The case of privacy is not as straight forward as the issues of Bias. In many locations where facial recognition systems have been implemented, the people under observation have waived their right to privacy. According to [2], an individual has the right not to be used without their consent. Therefore in those situations the system is acting morally. In any situation where people do not consent to waive their right to privacy, the system is in the wrong.

4.1.3 Kant's Categorical Imperative

As covered in section 2.1.1 the agent ought to follow Kant's Categorical Imperative ("Treat people as an end, and never a means to and end").

Trolley Problem

Kant's Second Formulation of the Categorical Imperative, according to [6], in simple terms says not to use people as merely a means to an end. In the case of a trolley problem, the second formulation would argue that whether the vehicle chooses to kill the bystanders or the passengers it would be acting immorally. This is because in either case the Vehicle is using someone else as an means to an end. The First formulation would also agree, since the maxim of killing the few to save the many is not universally applicable.

Mundane Problem

According to Kant's categorical imperative it is clear that for example in the example of approaching a cross-walk with limited visibility, the vehicle should not slow down because it would use the passenger as a means to an end (wasting their time, to protect a possible pedestrian) instead of as an end.

Bias and Prejudice

Facial recognition systems have innate biases and prejudices, written or trained into them by their developers. According to [45], the principle of discriminating against a specific group of people in order to ensure the safety of the whole is not universal. However, the system itself is designed to protect people, thus treating them as an end. The act of surveillance, as seen in Section 3.2.4, according to [45] then becomes a moral act through the second formulation.

Privacy

Privacy in areas of surveillance, is controversial. In the case that people knowingly and willingly consent to their faces and information being observed and analyzed, the act of breaching the individuals privacy would be seen as moral and necessary. This is supported by [45], with Kant's Second formulation. In this formulation the agent, the surveillance system, is not treating individuals as "merely" a means because the individuals willingly participate.

In situations where individuals do not knowingly and willingly consent the act of invading their privacy, according to [45], then becomes immoral because the agent is using the individuals as a merely a means to an end.

4.2 Consequentialist view

4.2.1 Ethical Egoism

As covered in section 2.1.2, according to ethical egoism the agent ought to take actions which best serve their best-interest..

Trolley Problem

From a the perspective of an ethical-egoist, the driver of a trolley would choose the course that results in the least damage to them self [8]. The ethical egoist would say that the agent, the automated vehicle, should protect itself and its interests. Therefore, the any act that protects the vehicle and its passengers (who would be considered to be the interest of the vehicle) is considered moral.

Mundane Problem

In the example of approaching a cross walk with limited visibility the vehicle ought to do what has the greatest benefit to it. Whether the vehicle chooses to slow down not to be responsible of killing, to not slow down not to be responsible of wasting the passengers' time, or a to take a compromise is unclear because we do not know which one gives the greatest benefit to the vehicle.

Bias and Prejudice

According to Rachels in [8], the ethical egoist serves their best interest. In the case of Facial recognition biases and prejudices would be against the best interest of the agent. This is because minimization of bias and prejudice would lead to better performance from the system. Thus, prejudice and bias in facial recognition systems today are not in the best interest of the system, showing this act to be immoral.

Privacy

As with the section above, the best interest of the program must be taken into consideration. The consequences of breaching an individuals privacy is a better functioning surveillance system. According to Macnish in [40] it is in the Systems best interest to function at its highest potential. Therefore the act of breaching an individuals privacy is morally correct.

4.2.2 Ethical Altruism

As covered in section 2.1.2 according to ethical altruism the agent ought to take actions which benefit everyone else the most, while ignoring themselves.

Trolley Problem

In the case of the trolley problem, if there are more pedestrians in danger of death than there are passengers in the vehicle, the vehicle ought to be programmed to crash into the wall killing the passengers and saving the pedestrians. According to Fieser in [1], the agent should behave in a way that benefits

society over oneself. Therefore the vehicle ought to be programmed to save as many lives as possible, with no weight placed on the vehicle itself.

Mundane Problem

In the example of approaching a cross walk the vehicle ought to do what has the greatest benefit to everyone else excluding itself. Whether this is stopping or slowing down is unclear.

Bias and Prejudice

In the case of the surveillance systems today the system is full of small biases. However, the technology, according to [1], should benefit everyone but the agent. Therefore ethical altruism cannot say have a true opinion on this topic because the information on whether this hurts more than it helps is not available. As on one hand the fact people are being discriminated against by a system needs to be considered, on the other hand, their safety also needs to be considered, and in this case they are at odds with one another.

Privacy

Privacy issues unlike bias issues, can be judged by ethical egoism. Since the consequences to the agent are not taken into consideration; The act of invading the privacy of all individuals has to be weighed against the benefits. According to Msafiri in [38], the benefits of facial recognition to the people under surveillance are plentiful. Therefore act of invading privacy becomes moral when weighed against the many benefits.

4.2.3 Utilitarianism

As covered in section 2.1.2 according to utilitarianism the agent ought to take action which benefit everyone the most including themselves.

Trolley Problem

In the Trolley problem through the utilitarian perspective the correct course of action is always to protect the lives of the most people possible. However, unlike the altruistic perspective the vehicle may take itself into consideration for the decision [9]. Therefore if the passengers and bystanders were of equal number the car could be permitted to save the life of its passengers.

Mundane Problem

In the example of approaching a cross walk with limited visibility the vehicle ought to do what has the greatest benefit to everybody including itself. It is not clear whether that is to slow down or not.

Bias and Prejudice

With biases in surveillance systems, the utilitarian perspective argues the action with the most beneficial consequences for all is the correct one. In this case then the benefits of potential lives being saved outweighs the potential cases

of flagging false- positives. Thus the act of having prejudices in the system is permitted morally [9].

Privacy

The issue of privacy like that of Bias and prejudice considers the benefits and consequences. The act of invading every one's privacy to protect them is also morally permitted by the theory of utilitarianism.

4.2.4 Applied Normative Principles

In this section we are analyzing the case studies as a whole, using the normative principles of Deontological and Consequentialist theories.

Personal Benefit

As we saw in 3.1.2, Automated Vehicles will provide countless benefits for the individual consumer as well as to the private companies that develop and build them. Therefore we can say that Automated vehicles today, follow the principle of Personal benefit.

In facial recognition technology, the agent will benefit if they are in the organization or group the surveillance is protecting. In most cases this is true and therefore the technology follows the principle of Personal Benefit as well.

Social Benefit

Much like the personal benefits of Automated vehicles the social benefits of more efficient traffic systems and lifesaving potential are plentiful. It is then easy to say that Automated vehicles follow the principle of Social Benefit.

With Facial Recognition Technology then the Social benefits include security, safety, and order. As seen in most of the Consequentialist theories the technology is generally considered to be very beneficial. Therefore we can say that Facial Recognition Technology also follows the principle of Social Benefit

Benevolence

Automated vehicles have been designed with the intent to save lives and provide better traffic efficiency to consumers. When we consider the developers of the technology as a the agent we see that the technology is following this principle.

In Facial Recognition, the designers and developers of the technology also have the intention to protect lives and benefit individuals. When considering them the agent we can say the implementation and development of the technology to be benevolent.

Harm

Automated vehicles are built to be as safe as possible. However, as shown in Section 3.1.3, the vehicles are bound to run into scenarios such as the Trolley Problem however rare they may be [28]. In these situations there will always be a casualty or injury. When harm will inevitably befall individuals involved then we cannot say that Automated vehicles can be implemented in a way that does no harm

It is similarly hard for facial recognition to do no harm. In this case the facial recognition's bias and prejudice, in surveillance, will inevitably flag an innocent person. However rare, false-positives from systems in use can harm individuals being observed [9]. Therefore we cannot say that Facial recognition can be implemented while following the Principle of do no harm.

Autonomy

Autonomous vehicles similarly to regular vehicles do not remove a persons autonomy in most cases. In situations like the trolley problem, however, the act of allowing the vehicle to make the choice removes the persons control over their body. In mundane situations as well, any approach to common traffic responses that are not controlled by the passenger remove a persons Autonomy.

Facial recognition similarly hinders autonomy. In the case of the Chinese Social Credit system [38], citizens under surveillance will increasingly see benefits for specific actions and not for others. This predetermined choice of good and bad behavior, removes the individuals autonomy to make decisions and act in accordance to their own values. Therefore the technology does not follow the principle of autonomy

4.3 Precautionary Principle

The Precautionary principle, refers to the idea that a technology should be proven to be safe before it is implemented in society [46]. In both cases of Facial Recognition and Automated Vehicles the technology as of today is not proven to be safe. There are many interpretations of the Precautionary Principle and according to most the development of the technology is not categorically forbidden.

Chapter 5

Discussion & Conclusion

5.1 Discussion

5.1.1 Explanation of assumptions

In the Analysis section we show the assumptions that were made and this section will explain in more detail why these were made. The first of these assumptions is that the agent is the system that is using this neural network. According to [6] to be a person one does not have to be a human and an "artificial intelligence computer systems could count as a person". This is done because we want to analyze the system that is using this neural network and not how someone is using the system. By viewing the actions and consequences of the system alone we can get a more accurate idea of the ethical theories views on the system alone. An example of this is for facial recognition in the analysis where we view the actions of the system to chose someone but not view what will happen to that person by police after they are chosen.

The second assumption that we make is that all of the theories carry the same weight. This is done so that we can create a better comparison of how the different theories view these issues and where they differ from each others views. If we were to weigh different theories by different amounts it could create an environment where some theories are viewed as less than others creating some bias in the analysis and comparison of these theories.

The applied ethical principles that were described in Section 2.1.3 supersede the ethical theories (Deontological and Consequentialist) because they were derived from them. The first two of these are derived from Consequentialist theories, these are personal and social benefits. The remaining four are derived from Deontological theories, these are benevolence, harm, autonomy and precautionary. By having broad ethical principles from both theories of normative ethics we can created an analysis of the issues that will be representative of more ethical theories. This will create a better representation of a real world view of this technology and the ethical issues that are involved.

By using these assumptions it creates some restrictions and limitations to our analysis of the theories and their views on the issues with the technology. These are that when we assume the agent is the system we ignore the potential misuse of the technology. We explain above why we make this assumption as we want to focus on the system used when in proper ways as technology can be used for unethical purposes even if this was not the intended purpose. An example is self driving cars do not have as much room for misuse as facial recognition systems have. Another limitation is that there are hundreds of applied ethical principles and we have had to chose a small enough number of them that we can give a proper analysis, but also have a wide enough range of principles to make it cover a large portion of ethics.

When we look at the principles of applied ethics we chose to expand the definition of the agent to include the developers. This was done because we were discussing the intent of the agent and by adding the developers to the agent we were able to talk about the reasoning behind the creation and development of the technology. If we look at self driving cars we can see that they will save a lot of lives because of how many lives are lost to human error and this is one of the main motivations for the technology.

When it goes with the cases of automated vehicles, we have chosen a simple case for each problem. This was done for simplicity because covering all the possible cases would be beyond the scope of this project. As such we assume that the conclusions we can draw from the cases apply to the entire problem to some extent.

5.1.2 Ethics and Technology

Technology exists all around us, it is beneficial to most but almost always has unintended consequences. For example, in the case of Facial Recognition, according to Macnish in [40], biases will always be programmed into systems because they are written by a person with their own personal values. These biases produce unintended results such as false-positives. When developers produce these types of technology their aim is to produce something beneficial. The main theories of Normative ethics differ, sometimes drastically in their opinions [2, 5, 3]. Where these schools of ethics agree is where more concrete statements can be made about a technology. For example, in every theory considered in the Sections 4.1 and 4.2, said that in the case of privacy: when individuals willingly participate the act of breaching privacy was moral in every theory. Considering the ethical views then helps justify further development and implementation of the technology in cases where the ethics agree.

5.1.3 Current state of the technology and the future of it

In 1943 neural networks were first modeled but widespread use of the technology is fairly new so problems have arose from this as seen in the Analysis section [47]. These problems come from different factors such as the technology not being fully developed, the widening uses of this technology and the regulation of it. Since this is a new technology there are issues such as incorrect object detection which can be dangerous [48]. Other issues have come up since the usefulness of

this technology is very apparent in today's world. To compound thishe amount of different use cases has risen which causes function creep issues. This is where the technology performs worse as it is used in situations it was not trained for, such as facial recognition systems being used on populations not well represented in the training data. Another issue is the regulation of this technology in regards to when it can be used, this is because neural networks require large amounts of data and usually regulation lags behind new technologies widespread use.

All of the issues that were just talked about are problems that can likely be fixed in the future of this technology. With improvements to the effectiveness of neural networks and how training data is used the error rates can be kept very low. The use cases of these neural networks also needs to be kept relatively small as the weak artificial intelligence we currently use does not work well outside of its specific use case. This could possibly change if the currently theoretical general artificial intelligence is possible to make but even without this it is possible to have a low error rate. Lastly, as the use of this technology gets regulated and becomes more normalized the issues with privacy will be minimized as people will have a better understanding of what and how their data is being used and how they can opt-in or opt-out of this.

In the future when this technology has had time to develop this will likely change some of the ethical theories views on if it should be used or not. An example of this is a patient-centered Deontological view on the issues of bias and prejudice. Currently this theory views the use of facial recognition unethical because it is wrong to use a system that discriminates, and has potential issues with privacy. In the future when the technology has improved and this bias has been removed or reduced to be negligible, and the regulation has caused the technology to only be used in places were people have waved their right to privacy this ethical theory will view the use of facial recognition as ethical. Another example is how Ethical Egoism looks at facial recognition. With the current technology this theory can not agree on how it sees the system as a whole since it disagrees on the issues. When this looks at bias and prejudice it this makes it unethical because of how the system works with the current bias. On the other side this theory currently views privacy as it is not an issue and is currently ethical to use. As seen the advancement of this technology can have a large effect on the ethical views of it and has a large positive change in the views as well.

5.2 Conclusion

Problematic cases in the use of Neural Networks span a myriad of technologies, only two of which were analyzed in this report. Facial Recognition software installed in surveillance systems, as described by Kshetri in [39] and Macnish in [40], come with the issues of Privacy and Bias. Meanwhile, Autonomous vehicles come with problematic trolley cases as well as more common mundane situations [28, 30].

Through ethical analysis of these two technologies we found the schools of normative ethics when applied alone, as in sections 4.2 and 4.1. have many a multitude of opposing opinions. Trolley problem cases, for example, are handled differently under Deontologcial Ethics than Consequetialist Ethics when one considers the Agent to be the Technology using the Neural Network. When the role of the agent is the user or developer of the technology, we can see that the two technologies follow the the principles of Personal Benefit, Social Benefit, and Benevolence. However, as shown in section 4.2.4, these technologies do not follow the Principles of Autonomy, and to do no harm. These violations of applied normative principles occur because of human errors as well as technical limitations as described in sections 3.1.3 and 3.2.4. Because of the general disagreement from the two normative theories, the precautionary principle would advise against implementation of the technology. It is our belief that as this technology develops these technical limitations may be overcome, potentially to the point of making human error redundant. Therefore, because of the theories of Consequentialism and Deontologys' general disagreement about the morality of specific situations as well as the opinions of the applied normative principles leads us to conclude that Neural Networks, while currently implemented and used all over the world, should be limited in use until further developments consolidate ethical principals or decide on fixed rules and understandings for this technology.

References

[1] James Fieser. *Ethics*.

[2] Larry Alexander and Michael Moore. "Deontological Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University, 2020.

[3] Daystar University Christine Wandolo. "DIMENSIONS OF DEONTOLOGICAL ETHICS". In: *Tura Journal of Humanities and Social Sciences* 2 (2018), pp. 11–16.

[4] David Mcnaughton and Piers Rawling. "Value and Agent-Relative Reasons". In: *Utilitas* 7.1 (1995), pp. 31–47. ISSN: 17416183. DOI: 10.1017/S0953820800001825.

[5] Wrenn. *Internet Encyclopedia of Philosophy*. December 2020.

[6] Andrew Chapman. "Deontology : Kantian Ethics". In: *1000 word philosophy* (2014), pp. 1–9.

[7] Allen Buchanan. *Categorical Imperatives and Moral Principles*. Tech. rep. University of Minnesota, 1977, pp. 249–260.

[8] James Rachels. "Ethical Egoism". In: *Ethical_Theory_An_Anthology*. Ed. by Russ Shafer-Landau. John Wiley & Sons, Inc, 2013, p. 812. ISBN: 9780470671603.

[9] John Stuart Mill. "Utilitarianism". In: *Ethical_Theory_An_Anthology*. Ed. by Russ Shafer-Landau. John Wiley & Sons, Inc, 2013, p. 812. ISBN: 9780470671603.

[10] Mat Buckland. *Programming Game AI by Example*. eng. Wordware Publishing, Inc, 2005. ISBN: 1556220782.

[11] Balaj Venkateswaran Giuseppe Ciaburro. *Neural Networks with R*. Packt Publishing, September 2017. ISBN: 9781788397872.

[12] Michael Nielsen. *Neural Networks and Deep Learning*. Determination Press. 2015. URL: http://neuralnetworksanddeeplearning.com (visited on 12/10/2020).

[13] Grant Sanderson (3blue1brown). *Neural Networks*. figure. YouTube. 2018. URL: https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi (visited on 12/15/2020).

[14] Wikipedia, User: Wiso. *Neural Network Example*. [Online; accessed December 1, 2020; Public Domain Image]. 2008.

[15] Isha Salian. *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?* NVidia Corporation. August 2018. URL: https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/#:~:text=In%5C%20a%5C%20supervised%5C%20learning%5C%20model,and%5C%20patterns%5C%20on%5C%20its%5C%20own. (visited on 12/16/2020).

[16] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, et al. *Intriguing properties of neural networks*. 2014. arXiv: 1312.6199 [cs.CV].

[17] "Phantom auto will tour city". In: *The Milwaukee Sentinel, 8th December* (1926).

[18] Waymo. *On the Road*. 2018. URL: https://web.archive.org/web/20180323062918/https://waymo.com/ontheroad/ (visited on 12/17/2020).

[19] SAE International. "Surface Vehicle". In: *SAE International* 4970.724 (2018), pp. 1–5.

[20] Brian Paden, Michal Cap, Sze Zheng Yong, et al. "A Survey of Motion Planning and Control Techniques for Self-driving Urban Vehicles". In: (2016).

[21] Brody Huval, Tao Wang, Sameep Tandon, et al. *An Empirical Evaluation of Deep Learning on Highway Driving*. 2015. eprint: arXiv:1504.01716.

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, et al. "SSD: Single Shot MultiBox Detector". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Cham: Springer International Publishing, 2016, pp. 21–37. ISBN: 978-3-319-46448-0.

[23] Guidolini R., De Suoza A. F., Mutz F., and Badue C. "Neural based model predictive control for tackling steering delays of autonomous car". In: (2017).

[24] World Health Organization. *Global Status Report on Road Safety*. 2015.

[25] Shaheen S. Greenblatt J.B. "Automated Vehicles, On-Demand Mobility, and Environmental Impacts". In: *Curr Sustainable Renewable Energy* 2 (2015), pp. 74–81.

[26] Raphael E. Stern, Shumo Cui, Maria Laura Delle Monache, et al. ""Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments"". In: *Transportation Research Part C: Emerging Technologies* 89 (2018), "205–221". ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2018.02.005.

[27] Sven Nyholm and Jilles Smids. "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?" In: *Ethical theory and moral practice* 19.5 (2016), pp. 1275–1289.

[28] Judith Jarvis Thomson. "Killing, Letting Die, and the Trolley Problem". eng. In: *The Monist* 59.2 (1976), pp. 204–217. ISSN: 0026-9662.

[29] Judith Jarvis Thomson. "The Trolley Problem". In: *The Yale Law Journal* 94.6 (1985), pp. 1395–1415. ISSN: 00440094.

[30] Johannes Himmelreich. "Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations". eng. In: *Ethical theory and moral practice* 21.3 (2018), pp. 669–684. ISSN: 1572-8447.

[31] Jonathan P. How. "Ethically Aligned Design". In: *IEEE Control Systems* 38.3 (2018), pp. 3–4. ISSN: 1066033X. DOI: 10.1109/MCS.2018.2810458.

[32] Carl Gohringer. "The application of face recognition in airports". eng. In: *Biometric technology today* 2012.7 (2012), pp. 5–9. ISSN: 0969-4765.

[33] Alex Sandy Pentland and Tanzeem Choudhury. "Face recognition for smart environments". In: *Computer* 33.2 (2000), pp. 50–55. ISSN: 00189162. DOI: 10.1109/2.820039.

[34] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization". In: *Proceedings of the IEEE International Conference on Computer Vision* (2011), pp. 2144–2151. DOI: 10.1109/ICCVW.2011.6130513.

[35] Abdul Halim and Karrar Adam Mahdi. "Mathematical model of Convolutional Neural Network for face recognition". In: *Roskilde University student projects* (2019).

[36] Patrick Grother, M Ngan, and K Hanaoka. "Face Recognition Vendor Test (FRVT) Part 3 : Demographic Effects". In: *Nistir 8280* December (2019), https://doi.org/10.6028/NIST.IR.8280.

[37] Philip Brey. "Ethical aspects of facial recognition systems in public places". In: *Journal of Information, Communication and Ethics in Society* 2.2 (2004), pp. 97–109. ISSN: 17588871. DOI: 10.1108/14779960480000246.

[38] Fulgence S. Msafiri. "Naval Postgraduate". In: *Security* June (2004), pp. 1–47.

[39] Nir Kshetri. "China's Social Credit System: Data, Algorithms and Implications". In: *IT Professional* 22.2 (2020), pp. 14–18. ISSN: 1941045X. DOI: 10.1109/MITP.2019.2935662.

[40] Kevin Macnish. "Unblinking eyes: The ethics of automating surveillance". In: *Ethics and Information Technology* 14.2 (2012), pp. 151–167. ISSN: 13881957. DOI: 10.1007/s10676-012-9291-0.

[41] Henry Lowood. *Introduction .. To Classify Is Human from Sorting Things Out: Classification and Its Consequences (review)*. Vol. 42. 2. 2001, pp. 392–394. ISBN: 0262024616. arXiv: arXiv:1011.1669v3.

[42] Ayanna Howard and Jason Borenstein. "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity". In: *Science and Engineering Ethics* 24.5 (2018), pp. 1521–1536. ISSN: 14715546. DOI: 10.1007/s11948-017-9975-2.

[43] H. L. A. Hart and Tony Honoré. *Causation in the Law*. Oxford University Press UK, 1959.

[44] J. J. C. Smart, Bernard Williams, and Anthony Quinton. "Utilitarianism; For and Against". In: *Philosophy* 49.188 (1974), pp. 212–215.

[45] Immanuel Kant. "Groundwork of the metaphysic of morals (HJ Paton, Trans.)" In: *NY: Harper & Row* (1964).

[46] Didier Bourguignon. "The precautionary principle". In: *Publications office of the European Union* (2015). DOI: 10.2861/821468.

[47] Bohdan Macukow. “Neural Networks – state of Art, Brief History, Basic Models and Architecture”. In: *Computer Information Systems and Industrial Management*. Ed. by Khalid Saeed and Władysław Homenda. Springer International Publishing, 2016, pp. 3–14. ISBN: 978-3-319-45378-1.

[48] Manikandasriram Srinivasan Ramanagopal, Cyrus Anderson, Ram Vasudevan, and Matthew Johnson-Roberson. “Failing to Learn: Autonomously Identifying Perception Failures for Self-Driving Cars”. In: *IEEE robotics and automation letters* 3.4 (2018), pp. 3860–3867. ISSN: 2377-3766.