Keele
UNIVERSITY

This work is protected by copyright and other intellectual property rights and duplication or sale of all or part is not permitted, except that material may be duplicated by you for research, private study, criticism/review or educational purposes. Electronic or print copies are for your own personal, non-commercial use and shall not be passed to any other individual. No quotation may be published without proper acknowledgement. For any other use, or to quote extensively from the work, permission must be obtained from the copyright holder/s.

Conscious robots: what happens when a philosophical confusion becomes a societal reality?

Sila Ozdemir

Thesis submitted for the degree of Doctor of Philosophy in Philosophy

June 2023

Keele University

ACKNOWLEDGMENTS

First, I want to offer a special expression of gratitude to my supervisor, Prof. James Tartaglia. My research could not have been completed without his continuous help and suggestions.

I am also grateful to the lecturers of Philosophy Department at Keele University, especially Dr. Stephen Leach and Prof. Sorin Baiasu who helped me during my research.

In addition, I would like to thank to Antonio Giuliani for his constant help and advice.

I want to express my thanks to the Turkish Government that gave me financial support for my PhD degree in one of the best universities in the United Kingdom under the best supervisor and lecturers.

Last but not least, I want to offer a special thanks to Dr. Sophie Allen and Prof. Emil Višňovský for their helpful comments and suggestions to finalise this thesis.

ABSTRACT

This thesis is an investigation into philosophical issues surrounding the production of human-like robots. I will argue that there is no reason to think conscious robots will ever be built, but that supposedly conscious robots that are able to emulate consciously guided behaviour will cause severe problems for societies of the future. In chapter one, I look at the history of machines and robots and show how changing conceptions of the body, along with technological development, has led us to expect conscious robots. In chapter two, I look at some philosophical theories of mind – behaviourism, identity theory, functionalism, eliminativism - to see how they encourage the idea of conscious robots and conclude that functionalism is the theory which mainly does. In chapter three, I discuss the main objections to functionalism that have arisen in the literature and conclude they are mainly correct and cannot be answered. So, a robot that satisfies the functionalist theory of mind will only be a supposedly conscious robot. In chapter 4, I look at the main ethical theories of Western philosophy and conclude that they make the idea of robots as moral agents very dubious. In chapter 5, I look at the harm that supposedly conscious robots may inflict on societies of the future. Then finally in chapter 6, I argue that a Singularity will never happen – supposedly conscious robots will never outdo the intelligence of human beings.

Keywords: Robots, Artificial Intelligence, Consciousness, Mental States, Understanding, Intentionality, The Technological Singularity.

TABLE OF CONTENTS:

INTRODUCTION

Humans have always been fascinated by robots since ancient times and have kept trying to build one more capable than the last (Truitt 2015; Mayor 2018). Science fiction today is more interested than ever in exploring the possibilities of human-machine relationships. These stories have gained huge popularity with both fans and futurists in complex, emotional, and thought-provoking ways. And, of course, recently we have seen many improvements in robotics (Kurzweil 2014; Ford 2015). From basic calculation tools to computers, from robotic arms in workplace to driverless cars, from health care robots to killer robots, robot technology is rapidly improving. Today, there are some robots that make some people think that they might be conscious and perhaps in the future there may – arguably – not remain any significant difference between the behaviour of humans and robots. They can reply to your questions, tell you a story or joke, clean your house, serve you, prepare your coffee, wash your car, remind you to take your medicines, present a funeral, etc. They might appear to be capable of doing anything that a human being can do. Therefore, some people might think that they are genuinely conscious. Most probably in the future those who believe that they are conscious will treat them as friends, collaborators, family members or companions and, in performing many tasks, may come to trust them more than humans.

One of the important companies that develop robots is Hanson Robotics. On their website, there is a title that attracts attention immediately. The title says that ‘we bring robots to life’. The company’s aim is to ‘create socially intelligent machines who care about people and improve our lives’. The company claims to make human-like robots that will develop cognitive abilities and simulate human mental states. Then, robots will interact with humans and evolve or improve themselves from this interaction. The website also says that they have a plan to make a ‘surreality show’ which is called ‘Being

Sophia'. They claim that this show will be about the robot Sophia's journey to become a 'super-intelligent benevolent being' and eventually, she will become a 'conscious, living machine'[1].

A 'conscious, living machine'… Is it possible that a robot can become conscious and socially intelligent and live among us so as to enrich our lives? What if there are already conscious robots because we humans are actually conscious machines? (Dennett 1994). These are the questions that I try to answer. In order to do so, I firstly ask what we mean by robots/machines; is it something just 'self-moving', is it a tool, or is a human being a machine? Secondly, I ask what we mean by consciousness: is it defined only as observable behaviour or is it something more than behaviour, identical to the brain or the functioning of the brain? What we understand by these terms will shape our answer as to whether a machine/robot can be a conscious living.

I begin my discussions by examining the history of machines and what it is that we understand 'robots' to refer to. In the ancient world the human body was seen as special, and machines were regarded only as tools or as the toys of the Gods. In the 17th century, this understanding changed: René Descartes claimed that animals and the human body are machines. He claims that the human body is a machine controlled by the human mind (Descartes 1637; 1641). Later, this understanding changes once again, and Julien Offray de La Mettrie claims that there is no difference between animals and humans, so, a human being is a machine (1747). Our understanding of machines has changed in time and is affected by changes in science, technology, and religion.

After reviewing historical changes in our ideas about machines/robots, and discussing dualist and materialist views of mind, I will move onto different materialist

[1] This information comes from the 'Hanson Robotics' website: https://www.hansonrobotics.com/

approaches to the mind/consciousness/mental states. What might the human mind be? What makes someone/something conscious? In philosophical discussions, sometimes a human mind is thought of as a separate substance from the human body, as dualism claims (Descartes 1637, 1641); sometimes the mind is equated with exhibited behaviour, as behaviourists (influenced by empirical psychology) claim (Ryle 1949); sometimes the mind is equated with the brain itself, as identity theorists (influenced by neuroscience) claim (Place 1956; Smart 1959). A currently popular theory (influenced by technological and scientific advances) is that mental states are best understood as functional states (Putnam 1967). If the mind is basically behaviour as behaviourism asserts, then we can build conscious robots because if a robot can exhibit predictable dispositions to behave in particular ways in particular situations, then it can be claimed to have a mind. Or, if the mental states are identical to the brain states as the identity theory claims, then if we can build a brain that has the same neurophysiological aspects of the human brain, and we can then build a conscious robot. Or, if the mental states are functional states, then if we can emulate the functions of the brain, we can then build a conscious robot. But what if consciousness does not actually exist in the way that eliminative materialists claim (Churchland 1984)? What if the mind cannot be equated with the brain? In the second chapter, I discuss the persuasiveness of the materialists' arguments.

Among these four theories (behaviourism, the identity theory, eliminative materialism and functionalism), I believe that functionalism seems to offer the most popular and plausible support to the idea that we can build a conscious robot; therefore, I devote relatively more time discussing this theory. Functionalism supports the common belief that better technologies will help robots to really understand. This theory depends on the analogy between a computer and a brain. According to functionalism, the mind works like a computer. The human brain is equivalent to the hardware of a computer

whereas the human mind is the software/program/code. Whereas computers are physical tools with electronic material which perform computations on inputs to supply outputs, brains are physical tools with neural material which perform computations on inputs that produce behaviours. So, the argument goes, anything can have a mind as long as it is programmed appropriately and so, in principle, machines can have consciousness (Dennett 1984; 1991).

Functionalism claims that mental states can be identified with their function instead of their material. That is to say, it claims that what matters to consciousness is not biological make-up but causal structure and roles; therefore, a non-biological system might be conscious as long as it is created or programmed correctly. If we can copy the functioning of a human brain, then, according to functionalists, this would produce real understanding and consciousness. For this reason, many have claimed that functionalism supplies a good reason why it is not impossible to build a conscious machine. But can this be right? Are mental states functional states? Is the mind a computer program? Can consciousness be computed/calculated? Is consciousness only symbol manipulation? Can computers have semantics (meaning)? Is formal symbol manipulation (syntax) enough for understanding (semantics)? Is simulation duplication? Can qualia (conscious experiences) be accounted for with functional role? I discuss these and related issues in the third chapter and provide an argument against functionalism (Block 1978; Chalmers 1996; Nagel 1974; Searle 1980; 1992).

Even though there are many rational objections to functionalism, some people will still claim that the mind is computable and strongly maintain that robots can be conscious as long as they are programmed correctly. Technology is improving and we will see many more innovations in the near future. Perhaps, robots will be on the streets and part of our lives very soon. We will see many robots in society and people will treat

them as if they are friends, family members or companions. That is to say, they will treat robots as if they are subjects. So, in the fourth and fifth chapters, I look at the contemporary issues and issues of the near future, that is to say, ethical issues relating to robots once they have become part of our society. For example, we live in society under moral rules. We, conscious humans have moral understanding. What about robots? Will they have moral understanding, too? Is it possible for robots to be morally good? Should humans have any concern to act towards them ethically? Is there any common moral rule that a programmer can use in designing a robot? Is there any ethical theory that allows robots to be moral agents? In chapter four, I discuss whether robots can be ethical or not and whether there is any reason to act ethically towards robots. These questions are examined in relation to three ethical theories – consequentialism (especially, utilitarianism) (Bentham 1789; Mill 1863; Wallach and Allen 2009), deontology (Kant 1785; 1793; Ulgen 2017), and virtue ethics (Aristotle; Foot 1995, 2002).

Related to ethical issues, in chapter five, I examine the consequences of treating robots as if they are subjects – specifically in relation to driverless cars, robots in the workplace, sex robots and killer robots. People will very likely vary considerably in their attitudes towards robots. Some will really like them and think that they are just like humans. They will even treat robots as if they are moral agents. They will try to not blame robots for any damage, and some will try to always blame and punish them for any incident. Some will choose them as romantic partners, and some will use them to kill other humans.

Like it or not, it is very likely that robots will increase in number and strength in our society (Ford 2015). With the new technological advances, they may be able to improve themselves (with machine learning). So, in the final chapter, I shall discuss whether, in the future, robots will surpass humans and perhaps even take over our planet

(Bostrom 2002; 2016; Kurzweil 2005). Is there any possible way for robots to reach human-level intelligence? (Dreyfus 1972). There are different suggestions as to the route to human-level intelligence (Chalmers 2010; 2022). Some argue that evolution made humans intelligent, so it may make robots intelligent too. Others argue that we can make robots with human-level intelligence. Of the latter arguments, the whole brain emulation argument is possibly the most promising. So, I shall focus on the question of whether, if we can emulate the functioning of a human brain, we can indeed create human-level intelligence.

CHAPTER 1: HISTORY OF MACHINES

INTRODUCTION

Technology has affected the conditions and the quality of human life. Humans are 'homo biotechnologicus', a biological being that cannot exist in the absence of technology (Višňovský 2015: 232). We create tools, machines, computers, tablets, smart phones, weapons, etc. by depending on scientific knowledge and material that we have in that moment in order to satisfy practical aims. One of the most astonishing devices that humans have manufactured are robots.

In different ways, robots have always amazed us. Sometimes it is their mechanism that surprises us, sometimes their advanced skills, sometimes their appearance, but mostly we are surprised by their apparently intelligent actions. Sometimes they even act *as if* they have mind – sometimes they can even have a conversation with you. Such robots were anticipated in science fiction, but now they have become a reality. The change from science fiction to reality began with Unimate, which was the first industrial robot, patented by General Motors in 1954 (Nof 1999: 5). After two years, the usage of robotics in the industry exploded, and since then they have been used in many different places.

These machines are manufactured with the aim of improving human functions. They help us with our daily tasks; they ease our lives. They can make calculations better than we do. Today, we use them in various areas from factories to the service sector, from agriculture to underwater, from hospitals to the military. Their functions may vary depending on whether they are used for defence, labour, sex, or entertainment. They can help us with washing cars, packaging food, or building new machines. They can go beyond our physical limits in carrying or pulling heavy weights. They can calculate the

salary of an employee, they can diagnose medical conditions, or they can tell us the expected time of a train. They can even help us to discover new planets where we might live in the future. These tasks were previously done by only humans (Kirk 2003: 13-15). These machines are becoming more common than before and are much improved upon previous models. And it seems that they will continue to make us, who cannot live without technology, ever more amazed.

Have you ever thought when the first robot might be built? You may think that we do not need to go back that much. But in fact, they have existed very long time ago but under another name. Today, what we call 'robot' is just one example of an 'automaton'. The word 'automaton' or 'automata' (plural) originally comes from ancient Greek.[2] An automaton means a moving mechanical tool which is made in mimicking of humans. Automata are designed to function according to a set of predetermined instructions. The term is applicable to a manufactured object which mimics a natural living form (Truitt 2015). So, robots are any device that are able to mimic human capabilities. In the 20th century, this word has been joined by others, such as 'robot', 'android' and 'cyborg' (Truitt 2015: 2). A cyborg is a half robot and a half human being whose physical abilities are increased with mechanical aspects that are built in the body. An android (a humanoid robot) is a robot that has the appearance of a human.

The beginning of robots can be traced back to ancient times. The ancient Greeks were said to be the pioneers of automata. In Greek mythology, there are mentions of automata that were built by gods and sculptors. It could be claimed that Greek mythology provides us earliest science fiction. So, in this chapter, firstly, some early examples of machines will be introduced. What I argue in this chapter is that automata were developed

[2] Homer was said to be the first who used this word to define 'automatic door' and 'automatic movement of wheeled tripods' (Homer *Iliad* 5. 749; 18. 376).

as toys and tools that were used to augment human abilities, to astonish and frighten others, to deceive, to hurt and kill in ancient times. In addition, I will claim that in the ancient world, the human body was thought to be special – hence the rarity of human dissection in this era. Later, I will continue with the robots/machines that were built from the Medieval period to the Modern Age. It will be argued that the idea of the automaton changed in the 17th century, when living things began to be considered as machines. I will discuss the arguments of dualist philosopher René Descartes who sees the human body and animals as a machine. Later, in the 18th century, the materialist philosopher Julien O. de La Mettrie extended this argument by claiming that there is no difference between animals and humans; therefore, humans are machines, too.

1.MACHINES FROM ANTIQUITY TO THE MODERN AGE

Any device which makes tasks easier can be called a machine. If we speak more precisely, a machine is a physical system which has certain structural and functional features and uses power to apply forces and control motion in order to perform an 'action'[3]. A machine requires energy sources that are input in order to accomplish the task which is its output. It works with the help of biological creatures, humans or animals, natural forces such as wind, water, and/or chemical or electrical power. It might have some sensors or a computer system that allow it to 'perform actions'; together these are often called mechanical systems. There are simple machines as well as the complex ones. For example, simple machines, such as the lever, wheel, wedge, pulley, and inclined plane, are the ones that help us to reduce the amount of energy required for a certain task and to multiply the force by exploiting the laws of physics (Usher 1929: 47). The lever

[3] I would like to clarify one important point here. During my thesis, you will notice that I am using quotation marks while talking about the capabilities of robots or machines. When I do it, I mean that it is something simulated, not genuine. For example, I often write machines can 'learn', or robots can 'decide', or they can 'choose'. I mean they are done unconsciously; this practice will be justified by the arguments of Chapter 3.

can help us to move a heavy object, for instance. Of course, time has progressed, technology has developed, and these inventions have developed or been replaced. Today we use these simple machines in more complex ways. Artificial devices such as engines and motors, vehicles such as cars, airplanes, boats, home and office appliances or devices such as computers, television, sewing machines, agricultural machinery, water treatment systems and robots are various examples of machines.

In ancient history there were various machines built that worked by steam, water, or wind power. One of them was a water clock. This tool measures time by the means of flowing water. This is arguably one of the oldest machines in history. It was believed to be created in Babylon and Egypt in roughly 1400 B.C. (Ewalt 2012). The Antikythera device that was made roughly 100 B.C. is believed to be one of the oldest analogue computers in history. It was a device calculating the position of the sun. There were a lot of examples of machines that were used in the courts. One of them was a throne that was said to raise the Byzantine emperor to the ceiling. Stories about this throne claim that the palace of the emperor had mobile statues of roaring lions (Littlewood, Maguire and Wolschke-Bulmahn 2002: 128-129). Most of the time, these machines were used to impress people by mechanical developments and to display the owners' wealth and power. In 18th century France there were machines that interacted with visitors at the chateau of the Count of Artois (Cave and Dihal 2018: 474). They used to reprimand them and sometimes soak them with water.

In the medieval period, the creation of machines was not encouraged, because it was believed that only God had the power to create. Nonetheless, it is said that in this

period, Roger Bacon and Albertus Magnus made a bronze head which could reply to your questions[4] (Cuervo 2017).

In the Renaissance, humanism placed emphasis on the value of reason and the human spirit. Attitudes to machines also changed. The Renaissance was an age of artistic achievement. There was a proliferation of mechanized angels in the churches, hydraulic spring-powered, clockwork automata, etc. One of the most famous inventors of this period is Leonardo Da Vinci. He sketched a clockwork knight (automa Cavaliere) which could sit, moves its head and jaw, and waves its arms. It is not known whether it was constructed, but if so, it might be the first humanoid robot that was actually produced (Ewalt 2012).

All of these machines were seen as tools; even though they were created by mimicking humans and animals; however, prior to Descartes, nobody seemed to have thought that living things could be machines. Descartes (1637) claimed that living beings were themselves, in part, machines. What he claimed, basically, is that the human body and animals are machines that are created by a divine power. He added that human beings in their entirety are not machines because they have a mind (soul, spirit). According to Descartes, the human body is a machine that is controlled by the mind. (We will return to discuss this idea in more detail very soon.). Thus, Descartes did not entirely ignore divine power. He did not think that a human being was entirely mechanical.

Descartes's ideas seem to have inspired inventors to think of building biomechanical automata. The first successful biomechanical automaton was built in 1739, by French inventor Jacques Vauncanson. It was a digesting duck which could flap its wings, eat, and digest food. Basically, the food was just collected in an inner container,

[4] Just like the 'Siri' of today. Siri is a system in smart phones which can answer any question in the language that we speak. So, this bronze head can be claimed to be the proto-Siri.

and the pre-stored faeces were produced from a second; therefore, there was obviously no actual digestion.

In 1770, Wolfgang von Kempelen built a machine that was called the Mechanical Turk. It was a chess player who used to beat most people at the game. Until one article in 1857 revealed how the machine worked, nobody thought it would be a hoax. There was a hidden person who was playing the chess inside the machine. In 1801, one of the most important steps in the history of computing hardware was invented by a French weaver, Joseph Marie Jacquard. This was an automated loom (the Jacquard loom). The machine was controlled by a chain of cards, consisting of several punched cards tied together into a continuous sequence. Jacquard's looms were small and only controlled a few warps ends, and it required repetitions across the loom width (Essinger 2004). Another landmark in the history of technology occurred in 1898 when Serbian-American inventor Nikola Tesla invented tele automaton, in the form of a radio-controlled boat, a remote-controlled technology (Ewalt 2012).

Then, technology continued to improve even more, and humans or 'homo biotechnologicus' created much better machines depending on their scientific knowledge. At the beginning of the 20th century, the machines became much more developed than before. The mechanization of physical work has begun, and industrialism began to replace workers. This was the time when the term 'robot' was used for the first time. Karel Capek introduced the word 'robota' into popular consciousness in 1921 (Hockstein et al. 2007). The term 'robota' is a Czech word meaning hard, relentless work. When Capek, introduced the term of 'robot', he was not aiming to introduce that term into modern languages; on the contrary, he was simply protesting the sharp rise of the automata in modern technology and the evolution of automata with higher capabilities.

Actually, the idea of creating robots dates back to ancient times. It seems really surprising because we do not readily associate robots with antiquity. However, it is undeniable that they appear in some ancient Greek texts.

In the *Iliad*, (c. 8th century B.C.), it is said that the god of the forge and master of technology and creation Hephaestus created golden maidens to help him when he was having a problem with walking on his own. He needed to be assisted by the golden maidens because he was limping across his hall. It is said that these 'golden maidens have understanding in their hearts, speech, and strength in them, and they know handiworks' (Homer *Iliad* 18. 480-490).

In the *Odyssey*, (c. 8th century B.C.), it is said that when the main character Odysseus entered in the palace, he noticed the watchdogs that were built by Hephaestus. These watchdogs were gold and silver, and ageless to protect the palace (Homer *Odyssey* 90-100). Even though the text does not describe them explicitly, it is implied that these dogs could move.

Hephaestus was the creator of the most captivating devices, which imitated natural bodily forms and possessed something like a mind and voice, and strength and knowledge (Kang 2011: 15-22). Most often it is the devices of Hephaestus that appear in Greek mythology. Another device that he created is Pandora that was built under the instruction of Zeus as a gift from the Gods to man. It can be inferred that she was a humanoid robot endowed with mind, speech, strength, knowledge of crafts from the gods, and the ability to 'initiate action'. In myths and legends humanoid automata often inspect, punish, and guard people and their abilities are magnificent and surprising (Truitt 2015: 60).

Perhaps, one of the most famous automata in Greek mythology, is Talos, which was a bronze giant shaped like a human being that was fashioned by Hephaestus. This was the first killer robot. Talos was created to defend the island of Crete against pirates. Talos patrolled the kingdom of Minos, the son of Zeus, three times a day and when there were captives, Talos would crush them and roast them alive. Talos was an animated machine, an imagined robot and automaton, able to perform human actions, programmed to recognise trespassers and designed to repel invasions. Talos was powered by an internal system of divine ichor, which was the blood of the immortal gods (Mayor 2018: 7-8).

Talos is mentioned in several sources, including a poem called *Argonautica* written by Apollonius of Rhodes in the 3rd century B.C. In the final section of the book, Jason and the Argonauts are turning back home with the precious Golden Fleece, but there was no wind to fill their sails; therefore, their ship, called the Argo, was becalmed. They found a sheltered bay on Crete which was defended by Talos. Talos noticed them and started to throw rocks from the cliff. Medea, a sorceress, suddenly appeared to rescue them before Talos roasted them alive. Medea prepared a plan to destroy Talos using her mind control and her special knowledge of the robot's physiology. She noticed that the ankle of the robot is the point of physical vulnerability (Buxton 2013: 88-94). Medea used some magical words to invite evil-disposed spirits and focused on Talos' eyes. She spread a kind of ominous telepathy, which confused him. When Talos tried to pick up another rock to throw, he dropped it on his ankle and opened his single vein. He collapsed onto the beach. It seems that Talos is represented as a human being that has a physical vulnerability, can get confused, and tries to defend itself. That is to say, human features were credited to Talos even in ancient times.

There is an older story about Talos in which Medea played on the emotion of Talos and Talos was demonstrated to have a human's fear, hopes, intelligence and volition (Mayor 2018: 10-11). Medea made Talos hallucinate his own death and persuaded him that she could grant immortality. There was one condition: the seal around his ankle had to be removed. When she removed it, the ichor flowed out, and he began to die. Thus, she overpowered him. Human beings are obsessed with eternal youth and immortality. So, here we can see that Talos is attributed human features again. Interestingly, in this story, Talos again seems to have consciousness and instinct because he acquiesced to the persuasion of Medea using his rational volition and agency.

Talos is not the only ancient automaton that had human features. For instance, the Argo, the ship that was used by Jason and the Argonauts when they were turning back home with the precious Golden Fleece is attributed the characteristics of living beings. The Goddess Hera ordered a wooden ship and so the Argo was built by Argus (one of the Argonauts) with the help of the Goddess Athena; therefore, it was a blessedly-inspired ship and possessed speech. Greek sources often mention statues, which are made by wood, metal, and marble, and that can move their heads, eyes, limbs; furthermore, they can sometimes sweat, shed tears, and make some sounds (Bremmer 1989: 13-15).

The Gods were not alone when they were creating automata. In the 4th century B.C., a wooden statue, which was carved by Daedalus, was said to be able to talk (Mayor 2018: 91). It is believed that it would run away if it was not fastened to the ground!

Another ancient story is that of Pygmalion, a sculptor who was smitten with a statue that he had made (Mayor 2018: 105-128). That it is possible to imagine that a human can fall in love with an automaton demonstrates that it is possible to imagine a human attributing attractive human features to an automaton.

In Greek mythology, in order to compensate for humans' vulnerability, legends tell of how the powers of gods and animals were given to automata. Thus was created, in answer to our hopes and desires, an ideal servant that always obeys the rules, a perfect soldier who never gets tired or feels scared, a wondrous partner that will always stay with you. But essentially, the automata were only tools.

Myths are stories which might be thought of as containing early history, including supernatural events, and which are intended, at some level, to embody truth. We have just given examples of the myths that include very early stories about robots.

2.THE HUMAN BODY AND ANIMALS AS MACHINES

Now, let us return to the 17th century to Descartes. There were some machines contemporaneous with Descartes; however, we ought to bear in mind that they were not the same as today. For instance, they were hydraulic rather than electrical. There were some sculptures in the royal parks powered by the flow of water, for example. When you stepped on a particular paving stone, a mechanical swordsman would appear with a sword in his hand. Descartes gives the example of a statue of Neptune which waves his trident when anyone inadvertently walks on a pressure pad (Descartes 1677: 100-101). He thinks that these machines are giving particular reactions to particular stimuli. Our bodies sometimes work in the same way. For instance, if you kick someone's knee, the leg rises. So, why should not our bodies be machines?

According to him, the human body is a machine which is controlled by the mind. Nevertheless, he does not think that a human being is just a machine. Human actions do not consist of only reflexive behaviours. Human behaviour can be regular, creative, and communicative, and can include the use of language. Use of language sometimes seems to depend on a reflex: for instance, when someone asks 'How are you?', we can just

answer ‘I am fine, you?’; however, should we wish to, we can be more discursive; we can choose what we are going to say, but machines do not have this capacity.

He notes that ‘the soul moves the body and the body acts on the soul’ and ‘the thought can move the body and feel the things which happen to it’ (Descartes, Elisabeth and Shapiro 1643: 65; 69). So, there is a clear separation between the mind and the body in Descartes’s dualist view.

Descartes’ inquiries begin by trying to doubt everything, including the existence of himself and God. He uses doubting as a method to find certain knowledge. So, the main argument of Descartes’ dualism (Cartesian dualism) is related to his project of doubting everything which helps him to reach his famous idea ‘cogito’: ‘I think; therefore, I am’ (cogito ergo sum). The Cogito shows (arguably) that Descartes exists essentially as a thinking thing, and since the characteristic feature of the mind is that it thinks, it is immaterial. He cannot doubt his existence; therefore, he knows that he exists (Descartes 1637: 18-22). However, he can doubt his body. Thus, he concludes that the mind and body are separate. According to Descartes’ dualist philosophy, because minds do not belong to the physical realm, they cannot be touched or seen. They can only be known through introspection, which is the mind’s own awareness of itself, its ability to ‘look inside’ itself.

Another reason that he thinks they are separate is that while the body is divisible, the mind is indivisible. ‘I am unable to distinguish any parts within myself when I reckon the mind or myself inasmuch as I am a thinking thing. I consider myself as single and complete. Even though the mind looks like being united to the body, I notice that if any piece of the body is cut off, nothing will be hence taken away from the mind.’ (Descartes 1641: 80).

The other thing that leads Descartes to the idea of mind-body separation is that the mind is not influenced by all parts of the body. It is influenced, he thinks, by a small central part of the brain which co-ordinates messages from other parts of the body (Descartes 1641: 92-103).

What Descartes understands by a body is whatever has specifiable form and a describable place and can occupy a location in such a way as to exclude any other body; it can be perceived by touching, seeing, hearing, tasting, or smelling, and might be moved in different ways, not by itself but by anything else comes into connecting to it. The ability of self-motion, like the ability of sensation or thought, is entirely foreign to the nature of a body. (Descartes 1641: 63-69).

So, in Descartes's view, the human body is a machine that is created by God and governed by the mind. God creates machines. We humans also create machines. Therefore, it seems that we are both engineers; God is an engineer that made us whereas humans are engineers that made our own machines. This idea is something very different than the perspective of previous periods. In the medieval period it was believed that there were significant differences between God's machines and machines that humans create. According to Descartes, a machine of God is 'incomparably better ordered' and 'its movements are far more wondrous' (Descartes 1637: 31). Prior to Descartes, it was thought that humans make things from other things, but God makes something from nothing. However, Descartes begins the process of blurring these differences. Descartes locates the difference more squarely in the quality of our creations – God's creations are better than ours. But, in that case, could humans build conscious machines?

Descartes's answer is that we could certainly make an 'animal': for he believes that animals are essentially machines (Descartes 1637: 31-32). Descartes does not think

that we can attribute understanding and thought to animals; however, he believes that some of the animals are stronger than us and there might be some that have an instinctive cunning capable of deceiving the intelligent human (Descartes 1646: 189-191).

Descartes thinks that mental activity influences our behaviours and thence our impact on the world, and conversely, the world influences our sense organs, and finally our minds and our decisions (Kirk 2003:30). There are two things which cause our actions: one of them is mechanical and corporeal – it depends on the power of the spirit, and on our organs and might be named as the corporeal soul; the second is the incorporeal mind that Descartes explains as a thinking substance. Animals are animated only by the first thing which is the corporeal and mechanical (Descartes 1649: 212-216).

So, we could build a 'monkey'. This is actually quite an interesting point if you think about the time that Descartes made this argument. Today, we may see some robots that are shaped as animals that have some astonishing movements even though they do not look exactly like the animals they are supposed to resemble. In some respects then, it can be said that humans can build 'animals'.

There was in fact a long tradition claiming that humans are special, being different from animals in that they possess rationality. So, this rationality makes humans separate from animals. Descartes added something different to this tradition. He said that the human body is a machine, too.

So, the question now arises: can we make a human being? Descartes (1637: 32) thinks if any machines resemble our bodies and act like us, we ought to consider two things which might lead us to think they may not be a real human being. The first thing we should consider is that they can never use words in order to communicate ideas to other people. Even though machines can utter some words, those words are not reasoned,

and so may not be appropriately correct answers. Secondly, even if machines do the same things that humans do, and just as well, they will fail in others. For machines can do only one task whereas humans are multi-tasking. You might think that we have machines that can do a lot of things today. But first we need to consider the time when Descartes was living. In that time, machines were able to perform only one task. Furthermore, it seemed doubtful that a machine could ever adapt to unforeseen events. Descartes argues that even if a machine does many things, even better than we do, eventually, it will fail since it does not do so consciously; it is governed by the disposition of its organs[5] (Descartes 1637: 33). Arguably, Descartes's point still applies to some extent, since machine creativity still cannot rival our own. According to Descartes (1637: 33), however, this will always be the case because machines will never be able to actually think. We humans use language to declare our thoughts to others by using our reason, understanding, and minds. Machines do not act on the basis of their own reasoning; they just act according to their predispositions. That is to say, although machines declare some 'thoughts', they are actually declaring whatever their programs have told them to say.

What about animals? Can they use language? Descartes' response is no. According to Descartes (1637), even those who are stupid or mad can arrange different words together and compose a sentence to convey their ideas; however, there is no animal doing this because animals lack ideas to convey[6]. He continues that although there are some animals which can speak, it does not prove that they have intelligence. For example, a clock can count the hours and measure the time more accurately than we do; however,

[5] Today, this may instead refer to the programming, with the disposition of their organs, for Descartes, being essentially their programming.
[6] We should recall that Descartes lived in the 17th century. In the 20th century, apes have been taught sign language.

it consists of wheel and springs (Descartes 1637: 33-44). However, this does not mean that it is as intelligent as a human.

So, he claims that animals are machines (1637: 32). He compares mechanical automata to non-human animals and claims that a real monkey or dog might be fooled by an artificial automaton, but a human would not be fooled. A human would notice that an automaton never has a genuine conversation[7]. According to Descartes, only humans possess minds (1637). A human is a creature which can think, whereas animals do not have a mind and cannot think. Even if animals could speak, it would not mean that they are thinking. And all animals lack consciousness and mentality, and their behaviour can be fully explained mechanically. On the other hand, humans have immaterial souls (minds); they are conscious, and have free will; therefore, they can be wicked or virtuous.

3. HUMANS AS MACHINES

The French materialist philosopher, and physician Julien de La Mettrie rejected this distinction between animals and humans and raised an objection to Descartes' view about the mind. La Mettrie believes that the human body and mind both work like a machine and that humans are just complex animals. He claims that human works like a machine because thoughts depend on bodily actions. Simply put, the organisation of matter at a complex level produces human thought, which is not a special faculty distinct from the workings of the material world[8].

[7] Three hundred years later, this idea began to be referred to as the Turing Test, named after Alan Turing who can be claimed to be the pioneer of 'artificial intelligence'. In this test, there is a computer, a human being, and a human questioner. The questioner tries to decide which one is a human being and which one is a machine by communicating with them. This test analyses whether people can understand if they are speaking to a machine or a human being. The test is often interpreted as proposing that if a computer can pass for a human, then we should deem it to be conscious and in possession of a mind.

[8] He arrived at his conclusion after discovering that his bodily and mental illness were related to each other.

According to La Mettrie, the body is like a watch, and a human is like a group of springs that enclose each other. The famous remark of La Mettrie is that 'the brain has muscles for thinking as the legs have muscles for walking' (in Crane 1995: 5). Therefore, a human being is a machine, and in the whole universe, there is only a single substance that is merely modified differently (La Mettrie 1747). 'The human body is a machine which winds its own springs' (Mettrie 1747: 94). The mind is like the body, in that it has its own equivalents of contagious diseases and scurvy. For instance, a human who is going from one climate to another feels the alteration, in spite of himself/herself. He/she is like a plant which is walking, that has transplanted itself because a plant will either improve or degenerate when the climate changes (Mettrie 1747: 98).

La Mettrie explains the similarities between humans and animals by claiming that the form and structure of the brains of some quadrupeds are very nearly the same as the brain of a man (Mettrie 1747: 99). The fiercer animals are, the less brain they have. This organ increases in size in proportion to the gentleness of the animal and the more one loses in instinct, the more one gains in intelligence.

La Mettrie (1747: 108) thinks that we can teach an ape how to pronounce and consequently to know, and the ape can learn a language. However, if that were the case, we should then have to ask whether they really learn it consciously or whether they just mimic?

La Mettrie (1747: 112) thinks that imagination is the highest product of the soul and the person with the most imagination ought to be regarded as having the most intelligence or genius. A soul is an enlightened machine, however, so ultimately 'soul' is an empty word (Mettrie 1747: 129).

CONCLUSION

Human beings have been fascinated by machines/robots throughout history. In the beginning, especially robots were just mythical creatures created by Gods or supernatural powers or they were thought of as mythical objects made by human beings with the help of magic or Gods. Later on, they were created to do the tasks of human beings and to ease their lives. With their aid, diseases have been diagnosed, and new buildings have been built; furthermore, they have enabled heavy objects to be moved and new planets to be discovered. While we tended to associate them only with science fiction, robots have become incorporated into daily life over the last century, and in the future, some believe there may even arise a new social class consisting of robots.

Automata were thought to take their strength from demons, the movement of the cosmos, and the secret of powers of natural substances or mechanical technology. After the 1750s, the word 'automaton' became more conspicuous because machines began to feature in philosophy, science, and medical discourses. In the 18th century, the definition of the 'automaton' became a 'self-moving' machine built for the particular aim of mimicking a living creature.

In the 17th century, Descartes describes a human being as a combination of the material body, which is an automaton created by God, and the immaterial soul, which controls the human body. Even though he thinks that the human body is a machine, he does not claim that a human is just a machine. According to him, animals, which lack consciousness, are machines. On the other hand, in the 18th century, La Mettrie claims that there is no difference between animals and humans; therefore, we can assert that human beings are machines and both the mind and the body work like a machine.

As we can see, what we understand about the notion of machines has changed over time. In the ancient world, the human body was considered special, and robots were just the toys of the Gods whereas in the time of Descartes the human body was thought of as a machine controlled by the mind or the soul. Later on, human beings began to be thought as machines. It might plausibly be claimed that it was the dualist approach of Descartes that inspired the idea that living things as biological machines which functions like a clockwork, an idea which La Mettrie then just extended. The aim of the next chapter will be to examine the latter claim in more detail.

CHAPTER 2: RISE OF MATERIALISM

INTRODUCTION

The previous chapter has ended up with the idea that humans are machines. On the one hand, Descartes believes that humans have minds, they have rationality, they have understanding; they use language; therefore, even though their bodies are machines, they themselves cannot be. On the other hand, materialists only believe in the physical world and physical person, that is to say, not an immaterial mind; there is no separation between the mind and the body. Thus, Descartes's dualist approach has been criticized by materialists continually since the 17th century.

Dualism is a theory where behaviour, actions and what is done by humans were thought of as the outward expression of what goes on into the mind. According to dualism, the mind and the body are separate (Descartes 1637). The mind is the controller of the body, and the body is merely a machine. This is founded on the belief that human bodies are material substances and are spatially located, and human minds are non-spatial and indivisible. Also, our mental states are entirely private and unobservable to anybody other than the individual; therefore, dualism supposes the mind and the body are distinct. A human being has two parallel histories, one of them consists of what happens in and to her/his body and this is public – external – that is to say, an event in the physical world. The other one consists of what happens in and to her/his mind and this is private which is an event in the mental world. Thus, one of the biggest problems with dualism arises – the interaction problem. If the mind and the body have totally different natures, how can they interact? How would Descartes solve this problem? He argues that the interaction between the mind and the body occurs via the pineal gland located at the back of the brain

(Descartes 1637). Since this gland is physical, it cannot solve the problem of how the physical and mental interact, so it is a solution that has never been accepted.

This kind of dualism was already established in the world and remains influential. A Persian philosopher from the 10th and 11ths centuries, Avicenna, with his metaphor of the floating man, argued that we can imagine being disembodied, so we could be disembodied. Let us consider X as a person. '1) What X imagines is possible. 2) If X imagines being disembodied, it is possible to be disembodied for X. 3) X imagines being disembodied. 4) Therefore, it becomes possible to be disembodied for X'. According to Avicenna, we cannot disclaim the consciousness of the self (in Goodman 1992: 161), which is similar to way Descartes argues for dualism, and also how David Chalmers argues against materialism in the present day.

Rationalist philosopher and priest, Nicolas Malebranche, appeals to the omnipotence of God to find a solution to the problem of mind-body interaction and claims that bodily actions are brought about by God (Schmaltz 1992: 303); therefore, mind and body cannot enter into the causal relations. Another alternative solution to the problem is Leibniz's parallelism. The main idea of Leibniz is that the mind and the body seem to interact because God created both of them in a pre-established harmony. We can imagine that the mind and the body are like two identical clocks and the clocks are always an agreement since there is pre-established harmony between them; however, they never truly interact with each other (Schmaltz 1992: 310).

Descartes's dualist approach was not embraced by materialists because materialist theories claim that the mind and the body are not distinct, and mental states are nothing more than physical interactions and mental states are identical with our brains, or our central nervous system and our mental states are identical with

electrochemical states of our central system, and we would not exist unless our bodies did. Despite the criticisms, dualism has continued to exist as a strong idea more than three hundred years. It entails the claim that humans are not their bodies; therefore, even if their bodies are machines, it cannot be claimed that humans are machines.

At the beginning of the 20th century, materialist approaches to the mind started to become very popular. Perhaps, we may say that scientific developments also help these approaches to find advocates. Humans started to believe only things that they can find in the physical world that they can test or measure. But if the mind is not separate from the body, if the mind is not an immaterial thing or spirit or soul, then what is it?

There are four materialist or materialist-compatible theories which have an answer to that question that I would like to discuss in this chapter. So, this chapter firstly focuses on behaviourism that reduces the human mind to behaviour. If our consciousness is behaviour, then if something behaves as if it is conscious, then it should be accepted as conscious. Therefore, behaviourism supports the idea that we might build conscious robots. Then, the chapter focuses on the identity theory, which claims that processes of the mind are identical to processes of the brain – the implication being that if we can build a robot which is physically the same as a human being, then we can build a conscious robot. The chapter continues with eliminative materialism (this is an extreme version of the identity theory), which believes humans' common-sense understanding of the mind (folk psychology) is mistaken and certain classes of mental states do not exist. Finally, the chapter discusses functionalism which asserts that a human brain is a computing machine, a theory that also supports the idea of conscious robots.

1.BEHAVIOURISM

The key idea of behaviourism is that our mind is only behaviour, and if the idea is extended to consciousness, that our consciousness is only behaviour. It reduces our

descriptions of consciousness, for example, being in pain to the descriptions of behaviour like clutching at an injury and crying out for help, for example. In order to fully understand what behaviourism (logical/philosophical behaviourism) says, we should look at its roots which come from behaviourism in psychology (psychological behaviourism).

Behaviourism started as a methodological research program in experimental psychology in the 20th century. Psychologists thought that information supplied by introspection was inherently ambiguous and unreliable. Psychology should rather rely on publicly observable and testable, and hence measurable, behaviour. So, the proper subject matter of human psychology is the behaviour or activities of the human being (Watson 1924). According to behaviourists, psychology should not concern itself with mental states that are not manifested in any form of physical behaviour. So, behaviourism aims to explain human and animal behaviour in terms of external physical stimuli and responses, as seen in the work of Pavlov, Thorndike, Watson, and Skinner. For example, Pavlov researched salivation in dogs in reaction to being fed (Todes 2014). He inserted a tube into the cheek of a dog to measure saliva when the dog was fed. Pavlov thought that the dog would salivate in reaction to the food which is placed in front of it, but he noticed that the dog would start to salivate whenever it heard the footsteps of his assistant bringing it the food. Pavlov discovered that any object or event that the dog learned to relate with food would trigger the same reaction. Pavlov focused solely on observable behaviour. So, he just looked at inputs and the outputs and predicted a huge amount of what was going on inside the body.

It was Gilbert Ryle's opposition to Cartesian dualism which combined with this idea from empirical psychology to produce logical behaviourism. The puzzle, as Ryle (1949: 15) saw it, is that the mind is isolated in dualism. The problem appears in the

logical structure of Descartes' theory of the mind, in which the workings of minds are given the attributes which bodies lack, for example, minds are not in space, they are not motions, and they are not observable. Although the human body is an engine, another 'engine' inside of the human body conducts some of its workings. This particular 'engine' is of a special sort, it is not visible, it is not audible, and it does not have any size and weight; therefore, no one can know how it conducts the bodily engine. The dualist theory claims that the content of somebody's own mind is best known by the individual herself/himself. The best way to know one's own mind is introspection because one's inner life is a stream of consciousness. The dualist theory claims that there is no direct way to know the inner life of other persons. We can only make inferences based upon their outward behaviour. Ryle objected to this whole traditional picture.

Ryle makes the criticism that dualist theory fails to prove the existence of minds other than our own[9] (1949: 16). He thinks that dualism is false in principle, namely, there is one big mistake in dualist theory named a 'category mistake'. This mistake arises when it is thought that mental life belongs to a different ontological category than physical life. He gives an analogy of somebody being shown around a university campus (Ryle 1949: 17-20). They are shown the library, the student union, the lecture theatre, scientific departments, and offices. The visitor thanks the guide but then asks them where the University is. The answer is that the University is not something above and beyond the buildings which they were shown. All the things that they were shown constitute the University just as all the parts of the body constitute the individual and the mind is nothing above and beyond the person (constituted by those parts). When we talk about mental states, all we are doing is talking about the behaviour that a certain person exhibits. When

[9] A counter-argument is that the situation is not different for the materialists – I do not see your brain, and if I did, I could not 'read' what it meant.

we talk about somebody's mind, we actually talk about her/his abilities, liabilities and tendencies to do certain things. The mind is no more than propensities to act in certain ways. It is not an object; it is just behaviour. There is, for example, no more to pain than the propensity to cry and recoil, etc... Thus, while Ryle was criticising dualism, he was, influenced by psychological behaviourism, advancing a view that has come to be known as 'logical behaviourism'.

Logical behaviourism claims that mental states might be analysed in terms of behavioural concepts. Any conversation about mental states is nothing more than a conversation about somebody's behaviour. In another words, logical behaviourism is a thesis which takes behaviour as constitutive of mentality and which is about the meanings of the terms or concepts of our mental states. According to logical behaviourists, for instance, when a belief is attributed to somebody, we do not say that s/he is in a particular internal state. They argue that it would be more accurate to say that we characterize the person in terms of what s/he may do in particular cases, their 'dispositions'. Any conversation about mental states is nothing more than a conversation about somebody's behaviour, actual or possible. The idea of logical behaviourism is that any mental predication (for instance, 'x is in pain' or 'x wants a holiday in Maldives') is to be examined and determined as a group of hypothetical or conditional ('if……then….') statements about how x would act in different situations. The description of the situations ('if….') and the description of the behaviour ('then….') will be expressed in terms of physical properties (such as size, shape, and velocity) of the material objects (Grayling 1998: 257-258). For example, 'I am angry' is nothing more than the description of my physical state, like glowering or scowling. So, statements about consciousness can be reduced to statements about behaviour without loss of meaning.

Behaviourists believe that if a machine or a robot behaves *as if* it has intelligence or a mind, it should actually be intelligent or have a mind. Therefore, in that sense, people who are involved in the idea of building conscious robots may follow a behaviourist model. For, as Galen Strawson points out, those who believe that we can produce conscious machines refuse to acknowledge the distinction between behaviours and what is supposed to be behind the behaviour (Strawson 2018: 130-153). Hence, if the binary opposition is denied, we should judge the sentience of robots or machines solely by their behaviour. If we follow this idea, we can also claim that zombies (physical replicas of humans that by definition do not have an 'inner life') have minds and consciousness so long as they clearly exhibit predictable dispositions to act in particular ways in particular situations.

Let us recall one of the examples that was introduced in the first chapter: the digesting duck. We can claim that according to behaviourists, it might have a mind because it was exhibiting predictable responses and behaviour. Indeed, if the mind is inferred only by behaviour, then it should have a mind. If we were simple-minded creatures, we would think that the duck decided to flap its wings, or eat the food and then digested it, and it generally knew what it was doing, but we know that it was just a clockwork duck, which was not capable of eating or digesting the food and it did not have thoughts and feelings (such as consciousness, sensations, emotions, and intentionality). We are persuaded that it does not have feelings and thoughts because we do not believe that it is conscious, it has no mind (or soul) and brain, it is just an automaton. Its behaviour does not convince us that it is conscious. (Kirk 2003: 1-6).

However, the question is: can we really identify the mind with behaviour? I would argue that we cannot. First, behaviourism does not supply a satisfactory explanation of the causal roles of mental properties and mental states: it is not clear how logical

behaviourism can allow for the thought of one mental state which causes, or causally interacts with, another one (Davies 1998: 258-259; Lewis 1966: 20-24). For instance, consider the behaviourist account of pain in terms of dispositions to pain-behaviour. This leaves out the fact that pain causes the behaviour. Likewise, if beliefs and desires are analysed in terms of behaviour, we will not be able to say that behaviour was caused by beliefs and desires. There seems to be a trouble with circularity in the analysis too. In order to analyse beliefs in terms of behaviour, we need to refer to desires and in order to analyse desires in terms of behaviour, we need to refer to beliefs (Chisholm 1957). For instance, if John wants a cookie, he will be behaviourally disposed to get one. But not if he believes they have been poisoned. So now this belief needs to be analysed, too. But then, he might actually want a poisoned cookie, if he is suicidal – so we need to add that he does not desire to die. We can go on like this forever, but the problem remains.

Secondly, behaviour sometimes does not correlate to somebody's mental state. For example, functionalist Hilary Putnam (1968: 332-333) suggests a thought experiment in which there is a race of human beings who show no pain behaviour. These 'super Spartans' or 'super-stoics' do not wince, scream, flinch, or cry; they keep all the pain inside, nonetheless, they do feel pain and dislike it. However, they carry on as normal. Or we can imagine an actor who looks to be in deep pain but actually he is just putting on a show for the audience.

Also, behaviourism does not explain the qualia that our minds perceive. Qualia are the qualities as perceived by a particular individual, such as experienced when I look at the colour blue. The way I personally experience blue is an example of qualia. My feeling of pleasure when I eat a piece of chocolate cake is another example. Therefore, qualia are the constituents of personal conscious experience. For example, let us say John likes X and it makes him laugh. Jade likes Y and it makes her laugh. These are the same

behaviours, but there may be different qualia and therefore conscious experiences. We can behave in the same way, but the underlying experience differs. Therefore, we can claim that behaviourism does not explain our conscious experience.

Behaviourism cannot be right because of the above three reasons. Therefore, if our conscious mind is not synonymous with its behaviour, then it cannot be assumed that we can build a conscious robot. Now, let us move onto another materialist approach to discuss what else the mind could be.

2.THE IDENTITY THEORY

The dissatisfactions that we have discussed with behaviourism led to the identity theory. According to the identity theory, mental states cannot be separated from physical states because each mental state is identical to some kind of physical state in the brain or central nervous system. The identity theorists claim that mental states are inner states producing behaviour, rather than just the behaviour itself. Basically, mental states are physical states of the brain (Place 1956; Smart 1959). In the famous example from John J. C. Smart, pains are asserted to be identical to C-fibres firing. When someone says s/he is feeling pain, they are reporting nothing more than neurological processes in her/his brain. The identity theory sees no problems with the mind-body interaction because mind and brain are one and the same. Thus, it aims to escape the problems of dualism.

In the 20th century, the identity theory became extremely significant and found many supporters. Remember that the theory claims the mind is just the brain, that is to say, mental states are only brain states. So, the theory seems to offer simplicity. Identity theorists propose that there is just one substance. This seems to accord with scientific reduction. For instance, water is chemical compound of hydrogen and oxygen (Feigl 1958; Shaffer 1961), or lightning is an electrical discharge (Smart 1965). Once they are reduced to something, it seems to be easy to understand what once were mysteries. The

identity theorists reduce the mind to the brain; thus, they claim to solve the mystery of the mind. They believe we explain mental states by reducing mind to the brain.

The identity theory is often thought to find support from neuroscience. Neuroscientists try to show the connection between the mind and the brain. For instance, when you drink too much alcohol, this may cause memory loss. Or, when a patient has Alzheimer's disease, some parts of the brain shrink. Or when someone had a car accident and injures her/his head, there occurs some changing in her/his mental states. There seems to be a strong connection between the mind and the brain.

In order to show this strong link between the mind and the brain, neuroscientists and researchers often refer to the story of Phineas Gage (Twomey 2010). Phineas Gage was a worker in railway construction. In 1848, while he was working, he was terribly injured. He and his colleagues were blasting rocks in order to build a railway. The task of Gage was to drill a hole in the body of a rock and fill it with gunpowder, sand, and wick, then press the mixture with the help of a pole. One blasting occurred unexpectedly and the pole entered his left cheek and went out from the top of his head. His left frontal lobe was destroyed. Less than a year later, he wanted to return to the work as he felt strong enough to resume. However, his employer did not take him back. He was in good health condition, but his doctor John Harlow said that his personality had changed dramatically. Before he was a kind, friendly, hardworking, responsible, efficient worker but after the accident, he became rude, impatient, disrespectful, and stubborn even though his general intelligence and memory seemed unaltered. His friends said that he was no longer Gage, but someone else. Today, it is known that frontal lobe has an important role in reasoning, language, and social cognition (Filho 2020: 419-421).

Moreover, today, an electroencephalogram (EEG) shows brain activities and there is magnetic resonance imaging (MRI) which demonstrates the kind of brain events

that are happening when we have positive memories or when we are angry for instance. Different parts of our brain are stimulated according to different mental states. So, with the help of these technological devices, the changes in our brains can be assessed. Because of this close link between the mind and the brain, the identity theory concludes that mental states are brain states, that is to say, the mind is identical to the brain.

What about building a conscious robot? What would the identity theorists say? 'Identity' refers to quantitative or numerical identity. For instance, you say that someone is 'Father Christmas', someone else might say 'Santa Claus', but they are the same. We are talking about one thing in two ways. In the same way, when we talk about mental states, we talk about brain states. Therefore, if zombies have all the behavioural and neural properties attributed to them by those who argue (from the possibility of zombies) against materialism, then zombies are conscious, and they are in fact not zombies (Kirk 1999: 1-16) Therefore, if we are plausibly able to ascribe neural properties to a robot, it must be conscious and should no longer be seen as a robot. The identity theory lends support to the idea which we might build conscious robots, so long as they are physically the same as human beings in certain respects.

Even though the identity theory has been believed to be correct by many, it is vulnerable to some significant criticisms. John Searle asks us to imagine someone who has a problem with vision and is becoming blind in time. So, s/he went to a physician, and they found the problem with the optic nerve that connects the eye back to occipital lobe. They removed the optic nerve and replaced it with an artificial connection that connects the signals from the eye to the occipital lobe and it works successfully. (Imagine that this artificial one is able to not only duplicate the functions but also the mental phenomena). Thus, s/he can see again. Searle says to imagine that a problem occurred again. Then, they found the problem which was the connection between the artificial optic

nerve and the occipital lobe. So, they replaced it, and s/he can see again. This time imagine that s/he has a problem with the occipital lobe, and it is not working. So, they replaced it and you have a new processor now. Searle says that this continues, and the parts of the brain are replaced one by one. However, s/he still has the same memory, hopes, desires, plans, emotions, and s/he still thinks and feels as the same person. But there is one difference now; s/he does not have a real brain anymore because it was all replaced by artificial ones. While s/he would be still having the same mind, but s/he would not have a brain at all (Searle 1992: 29-30). So, we can put the argument as follows: 1) It may be possible to replace the brain while the same mind remains. 2) If this is possible, then the brain cannot be identical to the mind. 3) If the brain cannot be identical to the mind, then the identity theory cannot be correct.

Another objection to the identity theory among materialists is that it seems impossible that there will be just one kind of neurophysiological state identical with every kind of mental state (Block and Fodor 1972; Putnam 1967). For example, we believe that the Moon is the only satellite of the Earth. So, let us say that my belief about this is identical to a certain state of my brain. If this certain belief has a certain physical state in the brain, this certain physical state should occur in everybody's brain. This seems unlikely to happen. The same applies to pain. You may claim that pain is identical to a certain physical state in all humans' brains, but there are other species that can experience pain but might be identified with the other kinds of neurophysiological configurations. This is referred to as the multiple realizability objection. The main idea is that mental states can be realized by different physical states. Let us consider a mental state, 'pain'. According to the identity theory, let us say that pain is C-Fibre stimulation just like lightning is electrical discharge (Smart 1959; Place 1956). So, thinking and feeling are certain types of neurological processes, and if there were these processes absent, there

would be no thinking or feeling. Putnam (1967) argues that it is not the case that pain is identical to C-Fibre as the identity theorists claim. If a brain does not have C-Fibre, then it should not be feeling pain. But this cannot be correct if dogs or octopuses can feel pain too but do not have C-Fibres like a human. According to Putnam, pain could be correlated with distinct physical events in the nervous systems of different types of organisms, but they could still experience the same mental kind of being in pain. Even though the brain structures of all mammals, octopuses or reptiles etc. differ, they can share the same mental kinds or properties because these mental kinds are realized by separate physical kinds in distinct species. To demonstrate, the human heart might be physically different than the heart of a giraffe; however, they both are hearts as they do the same job of pumping blood in the organisms that they belong to. So, if a mental state is multiply realizable by different physical states, then it cannot be identical to a certain physical state[10].

Another objection to the identity theory might be made using Leibniz's Law. Leibniz claims that if two things are identical, then they must have all the same features; therefore, if x is identical to y, then all features of x must be true of y (Loemker 1989). For the idea that the brain is identical to the mind, let us see how this plays out in the example of someone who has a mental state of happiness. The happiness will be about something; for example, Jane is happy that she passed her exam. But how can a brain state be about something, or how can the movement of molecules and nerve impulses be happy about something? Brain states and experiences have different properties. The brain is spatially located in the head of living human beings. When someone feels loved, are we to say that love is located on the right side of their skull!? It may be that something

[10] To further distinguish between behaviourists and the identity theorists: behaviourists would say that what all pains have in common is a certain behavioural property since they assert mental states as a behavioural state. So, this allows behaviourists to claim that other creatures such as reptiles, octopuses, dogs feel pain, so long as they behave the same general way.

happens on the right side of her/his brain when there is love and never on the left, but this correlation can be metaphysically described in any number of different ways, with identity being the most problematic. We can open someone's skull and see her/his brain, but we cannot point to something like feelings; we can look at somebody's brain, but we cannot understand what s/he is thinking. The conceptual thoughts which someone has cannot demonstrate themselves to us. All these things depend on conscious experience.

You might think to reply to this objection by saying that with the new technologies like MRI, today we are able to check someone's brain. For example, we can scan your brain while having happy news, or while eating a piece of chocolate, and observe where the happiness resides in your brain. Thus, we can actually point to happiness in your brain[11]. However, when we scan a brain and see differences in the brain waves for the different mental states, all we are seeing are the nerves. Brain scans or MRIs do not reveal any mental states unless we presuppose that the identity theory is true – nobody would think they were seeing happiness without the benefit of the theory, only what is going on in somebody's brain while they are happy. Given Leibniz's law, there must be a difference because the nerves have different properties from feelings like happiness. Also, we can measure brain waves, but not mental states, which is another difference. We may associate a certain physical state with a certain mental state, but this correlation could be accounted for in many ways. We might think the brain waves are a tool that transmits the message, but the message itself is not the brain waves. So, for example, we may use a computer (tool) to convey our thoughts (message) to someone, but when the computer is broken, it does not mean that our thoughts are broken too (Verschuuren 2012).

[11] This objection was suggested by Professor Sorin Baiasu during my doctoral progression meeting.

Given all these reasons, there is no reason to believe that the mind is identical to the brain. Even if we all copy the neurophysiological states, we cannot create the mind or consciousness. Therefore, the ideas of identity theory give us no good reason to think that we can build a conscious robot.

3.ELIMINATIVE MATERIALISM

Mental states exist according to the identity theory, but they are identical to brain states. There is also an extreme variant of the identity theory called eliminativism that claims that mental states do not exist. In a well-known version, it argues that our common-sense beliefs about the mind create a type of primitive theory, a 'folk psychology', which is radically false (Churchland 1981: 72). It denies that some or all the mental states which are posited by common-sense, for example, beliefs, hopes, fears, and desires, exist and will have a role to play in a mature science of the brain. Remember that Descartes insisted that we can be sure about the content of our minds, but eliminative materialism tries to reverse this by challenging the existence of mental states and eliminating the mind.

Folk psychology (FP) is thought to include both generalizations or laws and theoretical posits which are characterized by our daily psychological words, such as 'belief', 'pain' or 'feel'. The generalizations are considered to define the different causal or counterfactual connections. A classic instance of a folk psychological generalization might be: if somebody has the desire for A and the belief that the best way to reach A is by doing B, then s/he tends to do B. For example, if John wants to eat a cookie, and his belief is that the best way to do it is to go to the kitchen, then he tends to go to the kitchen. Here, we should underline that it is similar to behaviourist theory ('if…then…'). There is another similarity between behaviourism and eliminative materialism too.

Behaviourism reduces consciousness to behaviour; consciousness is nothing more than behaviour. It might mean that behaviourism actually denies the existence of consciousness (Strawson 2018: 135) – but that is not something that behaviourists necessarily accept. Since eliminative materialism denies the existence of consciousness, we can say neither that it supports the idea of a conscious robot nor that we can even raise the issue of whether robots can be conscious within its framework, since we humans are unconscious robots too.

There seems to be a big contradiction in eliminative materialism. The theory is self-refuting (Baker 1987; Boghossian 1990; Reppert 1992; Putnam 1991). Eliminativism is the belief which claims there are no beliefs; but eliminativists themselves believe that eliminative materialism is true, which is itself a belief, therefore, it is self-refuting. Intuitively, the initial plausibility of the theory seems, for many people, to be low. Assertion presupposes belief; therefore, eliminativism cannot even be claimed without contradiction. Putnam takes it further by saying that because aims are part of FP, then because a chair is functionally described by its aim, it cannot be true that all chairs have something in common; therefore, 'not just are there no such things as beliefs, if this idea is right; there are no such things as chairs!' (1991: 58). Eliminativists have replies to this objection, of course, but it is clear that the unpopular idea that there is no consciousness is not supporting the idea of conscious robots, which is the topic of this thesis. Thus, we can eliminate eliminative materialism and move onto a final theory, functionalism.

4.FUNCTIONALISM, PART 1

Functionalism is one of the most popular theories in this century and the last. According to functionalism, mental states are functional states of the brain and any state which fully plays the functional role of pain is a pain.

The improvements in technology, science, and computer science can be said to have influenced this theory. In the 20th century, a new technological device was created: computers[12]. But they did not look like the computers of today. However, their processing style remains the same. A computer is anything that processes representations systematically. It has a fixed or hard-wired architecture with a great quantity of memory that can run programs and store information (Crane 1995: 83).

Alan Turing (1936; 1950) developed the essential concept of a computer. He thought that computers can potentially *genuinely* think. There were two important things that are hidden inside a computer: hardware and software. Hardware is the fundamental part of a computer; it is a physical thing that all the computers are composed of. Software (code or program), on the other hand, is the operating instructions which tell the physical hardware what to do. This is what functionalism claims for the link between the mind and the brain, namely that they are related like software (mind) and hardware (brain) in a computer. Just as we should not think of software as an individual thing separate from hardware, perhaps we should think of consciousness as different from physical brain only in concept. So, mental states are not essentially separate from brain states, in accordance with materialism in general.

In order to understand functionalism better, let us look at some ideas behind it. Behind every device, there are mathematical or arithmetical functions. These functions are not actually numbers. They are something done to the numbers. We take numbers and use them to perform some functions. For instance, let us take the addition of two numbers, 2 and 6. These numbers are inputs. The function is the addition. As an output, we will

[12] To clarify, the computer was found by Charles Babbage in 1833. This computer was only to make mathematical calculations or calculating the taxes (Harris 2021). But Alan Turing prepared all necessary things to be able to use a computer as we do today.

reach 8. So, there are inputs, outputs, and functions in one processing. There are simple algorithms corresponding to the input, outputs and functions behind every device. This idea inspired and influenced functionalism (Crane 1995: 85-86).

There is another idea that some functionalists have adopted: human minds work like a Turing machine. In particular, Hilary Putnam once claimed that we should think about the mind as a Turing machine or computing machine[13] (Putnam 1967; Kim 2011: 147). A Turing machine is basically an abstract machine that makes complex mathematical calculations by manipulating symbols on a tape. Modern computers retain this basic concept. A Turing machine consists of four things. It has a tape that is divided into squares, a scanner which reads one square at a time and can delete what is in the square and can write something new in the square, it has a finite set of symbols written in the squares and it has a finite set of machine states which tell the scanner what to do when it reads the symbol in the square (Turing 1936: 232; Crane 1995: 93). Vending machines remain especially close to this process. Ned Block (1996: 30) gives an example from a coke machine; you can see its machine table below.

	S1	S2
State		
50p	Change to S2	Deliver coke Change to S1
£1	Deliver coke Stay in S1	Deliver coke Deliver 50p Change to S1
Input		

[13] This idea applies in particular to machine functionalism. There are various versions of functionalism. But specifically, machine functionalism is related to the Turing machine.

This vending machine gives you coke for £1. When you insert 50p into the machine, it will not deliver the coke and it will change the S2. Then, if you insert another 50p, it will deliver the coke and go back to the S1. If you insert £1 into the machine instead of inserting the second 50p (the machine is in S2 because you insert first 50p), it will give you the coke and a 50p. This illustrates the basic idea of all the computation. There are inputs, functions, and outputs.

If, as Putnam claimed, the mind is just a Turing machine and mental states are states of its machine table, then the connection between mind and brain is explicit. According to functionalism, the brain is the computer hardware, and the mind is its software. According to Daniel Dennett, 'computers are like brains to some extent, incompletely designed at birth' (1991: 211). When computers begun to become a subject of philosophical discussion in the 1940s, they were immediately thought of as thinking machines. For example, the ENIAC (electronic numerical integrator and computer), which was the first electronic general-aim computer with the ability to solve a large class of numerical problems through reprograming, was introduced as a 'giant brain' in 1946 (Kurzweil 2014: 180). This idea also overcame the problems with the identity theory, because Turing machines are multiply realizable. A Turing machine is an abstract idea; it can be realized in all sorts of ways. For instance, the same program might be connected to different kinds of computer hardware (Levin 2004). For this reason, Turing machines supply a good model for functionalist theories. Now, let us look at these two important features.

As explained, functionalism is affected by technological developments in computer science. The main idea of functionalism is that mental states perform functions on inputs to produce outputs in an algorithmic way. An algorithm is a process or a set of rules that should be applied in calculations. Functionalism is similar to behaviourism and

the identity theory, but there are some differences. In functionalist theory, we have inputs and outputs as in behaviourism. But behaviourism denies that mental states are inner states. Functionalism accepts that mental states are internal states, they are not only outward behaviour (Kim 2011: 137). Logical behaviourism aims to correlate each mental state with a characteristic model of behaviour, but the problem is that individual mental states do not always have characteristic behavioural effects. Behaviour usually proceeds from separate mental states which operate together, for example, belief and desire. Functionalism tries to avoid this problem by individuating mental states by way of characteristic relations not just with respect to sensory input and behaviour, but also to one another (Davies 1998: 262). Unlike the identity theory, functionalism does not claim that the mind and the brain are the same. According to the functionalists, pain, for example, is not identical to a type of physical state. Functionalism claims that mental states are defined by means of their causal roles (Lewis 1970; 1972).

The main idea of functionalism is that the brain is best understood as a computer. The mind is a computer software (computer program), and the brain is its hardware. All computers can do the same general job, in that they all perform computations. Likewise, our brain is like a computer that performs a computation. Our mind or mental processes or cognitive processes are computational processes. Any computation processes might be carried out in different machines. There are not just innumerable types of electronic digital computers but there are also computers which might be designed with wheels and gears (like the 'Analytical engine' of Charles Babbage, the first computer in history) (Wilde 2019). So, if the brain is a computer and mental states are computational states, then there is no obstacle to how minds and mental processes might be physically realized in completely different ways in humans, animals or even robots. Therefore, different

kinds of biological or physical systems can perform the same mental processes (Kim 2011: 132; Mitchell and Jackson 1996: 42-45).

In another words, functionalism is associated with the idea of multiple realizability that states a specific mental kind might be realized by multiple physical kinds. I may be in pain and that might be realized by C-Fibre activation in human beings, but pain might differ in other species (such as reptiles, octopuses). For example, pain is a state which is caused by bodily damage and that causes distress, and this causes, let us say, a belief about its location and cause. Pain is a state which has a particular type of sensory input, particular relations with other mental states and in conjunction with those other mental states a particular type of behaviour or output. What human pain, octopus pain, and the pain of an alien (whose chemistry is not carbon-based like ours, let us suppose) all have in common is that they share this causal profile or functional profile. Pain in the octopus is caused by bodily damage; it causes distress, and it causes withdrawal from the harmful stimuli. The same is true for us and for the alien. Any state which has this functional profile is a pain; therefore, it does not matter if it is a brain state of a human or the state of a different type of nervous system or a state of silicon chips. What pains have in common is causal role R, not any physical property. In other words, functionalism claims that mental states can be defined by way of their causal roles (Lewis 1970; 1972).

One of the earliest functionalist theorists, Putnam (1967), provides an example of the thesis. The human heart might be physically different than the hearts of birds, but they are all hearts because of the jobs that they do in the organisms in which they are found. The heart is made of certain biological material, but it is not the thing it is made of which defines it. We can create a robotic heart which does the same job. Therefore, according

to functionalist theorists, what defines a heart is how it functions. Here is the key feature of functionalism: the functional properties which create the mind are multiply realizable. This means the same mental kind might be realized by a variety of physical kinds. Therefore, a robot, an octopus or an alien could have mental states although their brains are made differently to the brains of human beings.

This presents us with a different view of mental concepts that shows no restrictions to the actual physical-biological mechanisms which realize them. This brings us to the idea that psychological concepts are like our concepts of artifacts. For instance, let us think of an engine. It will not matter how it was built or designed and it will not be important if it uses gasoline or electricity; nor will it not matter how many cylinders it has; the important thing is it is performing a specified job – just as, for example, you can play the same game on a Play Station or X-Box. In terms of biological concepts, for example, a heart differs in humans from reptiles; but it does not matter about the shape, size, or material composition. As long as it pumps blood, it counts as a heart (Kim 2011: 131).

Thus, according to functionalism, mental states are not identified by what they are made of; they are identified with what they do (Putnam 1975: 302). So biological makeup does not matter for consciousness. The important point is the causal structure and roles. Therefore, according to functionalism, any system, without biological makeup, can have consciousness *as long as* they are created appropriately. According to Dennett, in this sense, 'we, humans are complex, evolved machines that are made of organic molecules rather than metal or silicon.' (1991: 431- 432). So, if humans are conscious machines, if we are a kind of robot, why should there not be other conscious robots? If the mental states are functional states and if any system can have mental states as long as it is created correctly, then we can build conscious robots, according to functionalism.

This is a powerful argument for the claim that, in principle, there could be conscious robots.

5.FUNCTIONALISM, PART 2

A Turing machine can compute any computable function when it is given enough time and tape (Turing 1936: 249; Turing 1950). Then, another Turing machine may take the input which is the tape of the first Turing machine, and it may read the first Turing machine. Then it just needs a method of changing the operations which are defined on the first Turing machine's tape into its own operations. Thus, it will be another machine table which itself can be coded. That is to say, it will be *mimicking* the behaviour of our original machine. Turing (1950) claims that we do not require a separate machine for each operation to perform all the operations performed by Turing machines. He argues that we require just one machine which can 'mimic' every other machine. So, this is what 'universal Turing machine' is (Turing 1936). It is this idea that is fundamental to today's digital computers.

The Turing machine cannot make a cake; it cannot lock a door, yet the algorithm which is a definition of how to make a cake can, in principle, be coded into a Turing machine. Turing was one of the first to mention 'artificial intelligence'[14] (AI). He envisioned computers would play chess one day much better than humans. Some decades later, one chess player system called Deep Blue was able to beat the world champion in a game (Newborn 1997). This victory was an important moment in the history of AI. So, the science of AI has developed, new AI programmes have been created, mechanization has increased. By the end of the 20th century, it had become not uncommon to believe

[14] It might be also called 'machine intelligence' or 'computer intelligence'.

that thinking might be just rule based computation performed by creatures of different physical kinds.

Thus, the development in the science of AI has led to the idea that the mind is just a computer program. Searle calls this view 'strong artificial intelligence' (1980: 417). According to strong AI, a computer is more than a device; a programmed computer is a mind, in fact. With the right programs, computers might be claimed to possess understanding and other cognitive states. The result of this view is that in order for something to have a mind, biological makeup is not necessary. The brain is just one of the hardware computers which maintain the programs that can create mind intelligence. Therefore, any physical system can have a mind as long as it has the right program with the correct inputs and outputs (Searle 1984: 28). By contrast, according to 'weak AI', the computer is a powerful device which for instance, enables us to analyse hypotheses. For example, Siri is a system in smart phones, which can answer questions in the language that we speak. But this is all it can do. It should not be thought of as conscious.

The advocates of strong AI (such as Allen Newell and Herbert Simon (1959) and John McCarthy (1979)) claim that machines not only simulate human abilities but also demonstrate genuine understanding; and, furthermore, machines and their programs explain the human ability to find out about our surrounding and pose questions about it. For example, Newell says that 'intelligence is only a physical symbol manipulation; it does not have to have a link with any particular sort of biological or physical wetware or hardware' (in Searle 1984: 29). The inventor of the term 'artificial intelligence', McCarthy, even defends the view that 'machines as simple as thermostats can be said to have beliefs.' (1979: 14-16). Thus, for them, strong AI provides no objection, in principle, to machine consciousness.

AI seems to support the ideas of functionalism and from this a significant idea arises: the biological brain does not really matter to the mind. This idea has given rise to the new discipline: 'cognitive science'. Cognitive science asserts that the mind has mental representations analogous to computer data structure. But is the mind like a computer program? This question is important because if the mind is a computer program or actually is a type of computer, then it may be reasonable to think about building an artificial one. Let us imagine that the mind is to the brain what the computer software is to its hardware; the mind is equivalent to the brain's program. The programs are not physical properties which occupy space in computers even though they have been created by such chemical or physical properties. Likewise, the mind is not a physical or chemical property for human beings. Even if it has resulted from such a property, our mental talk then becomes functional or program talk. In a sense, our mental states, beliefs, desires, pains are nothing more than various sensory inputs and behavioural outputs. Let us think of a computer and a brain. The computer is a physical device with electronic substrates which performs computations on inputs to give outputs. The brain is a physical device with neural substrates which performs computations on inputs that generates behaviours.

Functionalists often claim that AI was inspired by functionalist theories. In the 1970s, cognitive science was formulated in terms of functionalism whereby mental states are considered the functional states of an abstract digital computer, thinking is abstract symbol manipulation as in the operation of a computer program, and the symbols of mind get their meaning by denoting things in the world. To understand the reasons why functionalism was thought to be the foundation of cognitive science, let us think about the neural networks in the human brain and the neural network involved in machine

learning[15]. The human brain and a computer both have a neural network (Rolls and Treves 1997). The average human brain consists of around one hundred billion neurons. The neurons form circuits and transmit information passed on from our senses by forming a network of nerves – a nerve which sends electrical impulses from one neuron to another is known as an axon. Neural circuits that carry out particular functions when activated, interconnect with one another to form large scale neural networks by helping us cluster and classify data which our brains store and manage. When we create a new neural network, we train it and apply what it has learnt to various tasks (Aggarwal 2019: 24). Then, once the neural network is set up, it practices and performs better through use by becoming more advanced over time. Here can be found the effects of functionalism on AI: these same biological neural networks inspired the design of AI neural networks. In a computer, a network of artificial neurons is created, similar to that of a neuron in the human brain, to set up machine learning algorithms. By connecting to neurons and forming an artificial neuron network, a computer might use labelled datasets (either structured by numbers and names, unstructured) to cluster and classify data according to similarities and find correlations to determine possible outcomes. In this way, the computer is trained to learn the correlation between a series of labels and datasets. Machine learning networks contain any number of hidden layers through which data must pass, while single-layer neural networks contain one hidden layer at most (Aggarwal

[15] This is similar to what connectionism suggests. Connectionism is a theory which tries to develop an idea about how humans learn and remember by explaining with reference to how the human brain works at the neural level (McClelland and Cleeremans 2009). It provides an alternative idea to classical computation conceptions of the mind, which hold that it works like a digital computer and that thinking is like running a program which manipulates symbols according to formal rules (Crane 1995: 159-167; Sharkey and Sharkey 2019). Like the classical model, connectionist conceptions are also compatible with multiple realizability. However, this theory has been criticised by Fodor and Pylyshyn (1988: 37) because one of the features of human intelligence is 'systematicity' and connectionism cannot deal with it, they contend. Systematicity refers to our ability to transfer our knowledge to new cases; therefore, we have an open-ended number of actions. In addition, I disagree with brain-centric theories (such as functionalism, identity theory and connectionism) that claim that we need to look at the brain in order to understand how the mind fits into the rest of the world. As discussed in this thesis, the mind cannot be understood by looking only, or even primarily, at the brain.

2019: 5-20). Therefore, the further the neural network is advanced, the more complex data sets it can handle. Artificial neural networks are also able to adjust the relative weight of various functions within each neuron, in order to better balance the competing factors in analytical processes, thus producing better judgements over time (Aggarwal 2019). For instance, Siri, Cortana and Google Now have now become mainstream. You can have a conversation with them; when you ask a question, they will answer you; when you ask them to perform a calculation, they will make it in seconds.

To sum up, according to functionalism, the mind is just a computer program whereas the brain is only a computer's hardware, and mental states are functional states. Functionalism allows the realizers of mental states to be physical. So, functionalism seems to solve the problems that behaviourism and the identity theory could not. Also, it seems to match up AI technologies and cognitive science that are very popular today. It seems to give strong reasons why robots might have consciousness.

CONCLUSION

We started this chapter with the problems of dualism. One of the biggest problems with dualism is the interaction problem since the mind is causally isolated by this theory. Another problem is that dualist theory claims that the best way to know somebody's own mind is introspection and we cannot know the inner life of other people. Therefore, as is often thought, dualism fails to prove the existence of minds rather than our own.

Then, we discussed the materialist approaches to the mind. The first one was behaviourism, which claims that there are no mental states to refer to, except insofar as they exist in the form of behaviour – the mind is only behaviour and dispositions to behaviour. We divided behaviourism into two major groups: while psychological behaviourism is a methodological theory which claims the mind should only be

investigated as behaviour, logical behaviourism is a semantic doctrine which takes behaviour as constitutive of mentality – that is about the meanings of the terms or concepts of mental states. According to behaviourists, a machine has a mind if it acts as if it has a mind. Then, we discussed some specific problems with behaviourism. For example, behaviourism does not provide an accurate explanation of the causal role of mental states; it does not explain how pain causes the behaviour. There seems to be a problem with a certain form of circularity in this analysis. Also, behaviour does not always correlate to a particular mental state. Putnam gives the example of a person in pain living in a society which has conditioned people to dominate their feeling to such an extent that they suppress the impulse to say 'ouch'. Finally, it was noted that behaviourism does not explain qualia.

Another theory that we discussed is the identity theory which also supports the idea that we can create a conscious machine, but only so long as the machine is physically the same as a human in certain ways. According to the identity theorists, if a machine is found to have all the requisite neural properties, then it is conscious; and should no longer be seen as a 'mere machine', but rather a machine in the same sense that a human is a machine. The identity theory claims that it has solved the interaction problem because the mind and the brain are identical – the mental state is the physical state of the brain. However, because the brain is spatially located, there arises a problem with this theory. For example, when I feel love, can we say that love is located in the left side of my skull? We can open up the skull and examine the brain, but we cannot point to love. Another serious problem is the problem of multiple realizability, since it seems unlikely that all species that feel pain will have the same kinds of brain states.

We also briefly discussed an extreme version of the identity theory: eliminative materialism. Eliminativism denies that all mental states that are postulated by common-sense thoughts and feelings such as belief, desire, and fear. But there is an important problem with eliminative materialism; it is self-refuting. While the theory claims there is no belief, eliminative materialists *believe* that their theory is true. Since the existence of consciousness is denied by eliminativism, we cannot claim that it supports the idea of conscious machines – on this theory we cannot even raise the question whether we can create a conscious machine.

Finally, we arrived at functionalist theory which is the main focus of this chapter because it seems to be the strongest theory which supports the idea of conscious machines. According to functionalism, the brain is a computer – for a functionalist, on one popular version, the mind is just a Turing machine and mental states are states of its machine table. Functionalist theories claim that if something acts as a particular mental state, then it is that mental state; and mental states can be realized by different physical states. Even if we have different internal compositions from robots or octopuses, it does not mean that we cannot experience the same sorts of mental states. Therefore, according to functionalism, it does not matter if the biological components are present. So we can build conscious robots.

With the development of AI research and cognitive science, functionalism has become a widely supported theory among scientists and philosophers; but of course, there are some debates about it. In the next chapter, we will discuss what kinds of objections to functionalism can be raised.

CHAPTER 3: OBJECTIONS TO FUNCTIONALISM

INTRODUCTION

At the start of this new chapter, let us remember what we have done so far. We first discussed the earliest automata including those found in myths, then we discussed early machines and the complex robots of today. We talked about the dualist view of Descartes who claimed the body is a machine which is controlled by the mind. Dualism defines mind and physical things in opposition to each other. Minds are non-physical and they do not have shape, size, or weight; they cannot be observed by any of the five senses. According to dualism, minds are the bearers of consciousness: minds think and feel; not the brain – whereas physical entities have no mental aspects. Physical entities have the observable aspects such as size, shape and so on. Feelings, emotions, and experiences reside not in the brain but in the non-physical mind. We should remark that dualism does not deny that when the conscious mind feels pain, it causes the body to cry; and that the body affects the mind also. Although dualism was a popular theory, many philosophers and scientists nowadays feel confident to say that dualism fails to account for the interaction between the mind and the body (Goff 2019: 25-39).

We also discussed the view that the human being is herself/himself a machine. This is the materialist view which claims that what you see is what you get and there are no immaterial or invisible parts. According to materialists, the inner subjective world of experience is to be explained in terms of the chemistry of the brain. So, in the second chapter, we looked at the different varieties of materialism. One of these theories is behaviourism whose claim is that mental states are just behaviour; in other words, if we see something which behaves *as if* it has consciousness, it is conscious. For this reason, we can claim that behaviourism allows us, in principle, to build conscious robots. Another

variety of materialism is the identity theory which says every type of mental state is identical to a type of brain state. The main problem with the identity theory is that it limits mental states only to brain states just like ours. It seems to fail to account for the fact that mental states are multiply realizable. There could be organisms with different physiologies from our own, but they might have the same mental states that we do. For example, the nervous system of humans and octopuses differ but it looks as though octopuses have minds somewhat like ours (Godfrey-Smith 2016: 4-5). Octopuses might experience fear and pain just as we do, for instance. The identity theory claims that pain is identical to C-Fibres firing. But even though octopuses do not have C-Fibres, they might still experience pain. Therefore, pain cannot be identical to C-Fibres firing. The point is that many types of physical states can realize pain, so pain cannot be identical to any particular type of physical state. However, if were to accept that the identity theory is correct and pain is identical to the C-Fibres firing, then that would give us reason to think that if we build the same physical states into robots, then robots would also feel pain.

An extreme version of materialism is eliminativism, which denies all mental states. According to eliminativism, we are non-conscious machines (in that there is nothing special or exceptional about consciousness). Therefore, if we create machines which look and behave the same as us, there is no significant difference between them and us. They would not be conscious robots though, for the same reason that we are not conscious people, namely that there is no consciousness in the world.

Finally, we discussed functionalism, which attempts to redevelop behaviourism so as to avoid the objections to the above theories. This theory is really the driving force behind hopes of building conscious robots, so I spend most time on it. Functionalism is fully comfortable with the mind-machine analogy; and it applies empirical psychology to

philosophy of mind. Functionalism has become a standard view in cognitive science. We examined the idea behind functionalism in the previous chapter, and it might be said that functionalism comes from one basic idea which is the multiple realizability of mental properties: mental states are identified not with some underlying physical structure but with a particular causal profile. A mental state is defined by the kind of causal or functional role which it plays in the overall system of which it is a part.

At first glance, functionalism seems to solve the problems of other theories and it has become a common theory among psychologists, computer scientists and philosophers. The main problem with the as we discussed in the previous chapter is that it neglects multiple realizability whereas the main problem with behaviourism was that it ignores internal states. Functionalism seems to solve these problems. According to functionalism, the function of a mental state is its defining feature; mental states are defined in terms of the causal role which they play in the whole system of the mind. This means that mental states are defined in terms of their causal relations to sensory stimuli, behavioural outputs, and other mental states. Since a causal role can be defined independently of its physical realization (since functional states are multiply realizable), functionalists could avoid the problems of identity theory. For example, the identity theory defines pain as C-Fibres firing while functionalists define pain in terms of the causal role it plays in our mental life. But since there must be some neuronal state (or equivalent) which realises the function, functionalism avoids the problem of behaviourism since it affirms the existence of an internal state. However, functionalism faces other problems. In this chapter, I shall argue that it overlooks some key points. We will investigate the problems with functionalism by way of several different thought experiments. We will begin with the famous example of Searle – the Chinese room

argument (CRA) and carry on the discussion with intentionality/semantics, consciousness, syntax/functions, and finally the mental causation problem.

1.THE CHINESE ROOM ARGUMENT: SYNTAX, SEMANTICS, MEANING, INTENTIONALITY

Searle develops a thought experiment called the Chinese Room Argument (CRA) which he uses to argue against the claims of functionalism concerning strong AI (Searle 1980: 417-418; 1984: 32-35). He asks us to think about a person who can speak only English. The person (analogous to a computer or the central processing unit) is locked in a room with nothing but a rule book (analogous to the computer program) and a basket of Chinese characters (analogous to the computer database). The room has a letter box and at times a paper which has Chinese characters is posted through the door. The rules designate the manipulations of the characters in terms of their syntax, not their semantics (Note that syntax is about the form/symbol and semantics is about the meaning/content; these terms will be discussed in detail later). The rule book guides the person so that when s/he receives a Chinese character through the door, s/he will answer with another Chinese character. There are Chinese speakers outside the room, and they write the questions to the person who can only speak English and put them through the door. When their letters reach the person in the room, s/he matches the symbols in the rulebook and sends the letter out with another Chinese symbol. Chinese speakers (imagine that they are programmers) outside of the room input those Chinese characters and get outputs of Chinese characters. Thus, they have an excellent conversation which they understand. If we only look at the inputs and outputs, we may think that the person in the room is fluent in Chinese, but we are aware that the person in the room actually does not know Chinese. S/he was guided by the rule book and read the inputs and gave the right outputs.

Can we say that the person in the room *really understands* Chinese? Even if we see the right inputs and outputs, it seems that we cannot say that there is real understanding here. The CRA can be concisely summarised: (1) Programs are syntactical. (2) Minds have semantics. (3) Syntax is not sufficient for semantics. (4) Therefore, minds are not just programs. The aim of the CRA is to show that the computers might *simulate* human behaviour (especially with the powerful technology of today), but they would still lack minds; they would not have any inner life: no conscious experience, no true understanding. Understanding is something which goes beyond mere computation (Penrose 1994: 40-41). The human brain's actions can be simulated by appropriate computation and improvements in technology that may even lead to computers surpassing many of our mental capabilities; nevertheless, the quality of conscious understanding is distinct from computation. What we do when we understand something is not computing: it is having a conscious awareness of it. With the CRA, we reach two important results: syntax (formal symbol manipulation) is not sufficient for semantics (the meanings) and simulation is not duplication (Searle 1980).

The CRA has received many reactions and Searle (1980) himself listed six replies and tried to refute them. Among those considered most important, there is the 'systems reply' (Berkeley) saying that the person in the room does not understand Chinese, but the fact is that the person is part of a system, and the system as a whole understands the story (Searle 1980: 419; Dennett 1991: 438-440). Thus, they do not attribute understanding or consciousness to the individual but to the whole room. Searle (1980) replies to this by saying that we should imagine the person in the room memorizing the database and the rule book. S/he does not need the room anymore. Then, s/he goes out and communicates with people in person in Chinese; however, s/he still does not understand Chinese; what s/he does is manipulating the symbols. S/he does not have understanding of these symbols

even though, in this situation, s/he is the entire system. Understanding cannot be grasped in the partly externalized system of the original Chinese room because s/he does not acquire understanding of Chinese by internalizing the external components of the whole system.

There is another response to the CRA: the 'robot reply' (Yale) (Searle 1980: 420; Harnad 1989). Let us suppose we install a computer in a robot. The computer will operate the robot in a way that the robot will do something like perceiving, walking, or drinking – anything that we like. For instance, the robot will have a television camera which is attached to it that enables it to 'see' and it will have moveable arms and legs. Those features will be controlled by its computer 'brain'. It is argued that, in this case, the robot will have real understanding and other mental states. The robot reply suggests that interaction with the real world is essential for understanding or intentionality; it respects the idea that suitable causal connections with the world might supply content to the internal symbols (it suggests the particular semantic theory of 'semantic externalism' (Putnam 1975)). For example, Colin McGinn writes that 'internal manipulations do not determine reference, but causal relations to the environment might' (1987: 286).

Searle responds that adding a set of causal relations with the outside world makes no difference (1980: 420). A digital computer does not recognize the symbols which create English words and sentences. They should first transform these symbols into symbols of the only language which the computer performs; this language is a binary code that represents text or computer processor instructions. The binary language consists of strings of 0s and 1s. When someone asks the computer, 'what is a book?' the computer takes that string of letters and transforms them into corresponding strings of 0s and 1s. Thus, we will have another room called the 'binary converter room' that will take as input the strings of symbols from the alphabet and give as output the strings of binary symbols.

Let us imagine that these symbols are sent to the Chinese room. There is now another rule book that shows how to manipulate these binary symbols. The person in the Chinese room would not recognize them – they were only shapes, and s/he would not even know that they are symbols; s/he could only know that when certain shapes came into the room, s/he needed to send certain shapes out of the room. In the original Chinese room, these shapes are Chinese characters; now, they become binary symbols. In both situations, the symbols which were sent to the room would only be meaningful if you speak that language (Chinese or Binary) (Anderson 2006).

Suppose we put a conscious human being into the head of the robot. We have a robot which takes and gives information about the world along two paths. The robot can process linguistic information. For instance, let us consider that a question ('What is a book?') will be written on a piece of paper and the robot, which has someone inside, could answer that question. The sentence will be written in binary codes, and that person does not understand that it is a question because there are just the strings of 0s and 1s. Without knowing the meaning of the things, we can send these binary codes out of the room. The robot might seem as though it really understands; however, the one who processes the question and responds is a person who does not know the binary language and if s/he does not understand, then we cannot claim that the robot can really understand what it is doing. But it is still not obvious that we have causal interaction with a real book. Let us add a vision system which seems to notice books; but we should recall that the visual information received by video camera is only digital information which are 0s and 1s and again the person who does not know that language will be the only one that is processing this information. Can we now claim that the robot understands what it sees? In order to answer that question, we can think of the performance of that person inside the room. S/he would not see a photo that s/he notices as a book inside the room. The

camera would receive the visual information from the book which is made up black pixels (which are converted into binary). But s/he would receive these binary codes as string after strong of code and would never recognize a book. When the information is processed, s/he does not acquire any real understanding of its meaning (Anderson 2006). In summary, when we check inside the robot, we discover that even though the lights are on there is actually nobody at home. As long as the only information processing consists of symbol manipulation there will be no real understanding.

These arguments try to connect the symbols to the things which they symbolize and are relevant to Searle's concerns about intentionality, syntax vs. semantics. Therefore, in the next paragraphs, we will analyse intentionality.

Minds understand; but the systems operating only on syntactic processes – inputs and outputs based on algorithms – can neither realize the meanings nor intentionality; the two are closely connected because what the symbols indicate, or are about, is determinative of what (if anything) is meant by them. The person in the Chinese room has something extra that, arguably, no computer can have. This is intentionality. Intentionality is the property of being about something (aboutness) or having content (Feigl 1958; Strawson 1994: 177-214). The inner signals of a machine or a system cannot be about something; therefore, they cannot have intentionality. This is where the argument becomes relevant to the question of consciousness. The Chinese room system lacks conscious states such as the conscious experience of understanding Chinese. Implementing a program is not sufficient for conscious experiences; consciousness requires something more than the implementation of a program. Searle's idea is that intentionality requires consciousness; however, this has been denied by others as we will discuss now.

There might be people who think that although mental states are intentional, there might be some things that are intentional without being mental states. This seems to support the idea that some things such as machines, robots, technological devices may have some kind of intentionality, even if they do not have mental states. This approach is related to the idea that intentionality is a naturalisable property. In order to explain intentionality in a naturalistic way, naturalists must explain how a brain state may represent some content. One theory that tries to reply to this question is called 'covariational theory' (Dretske 1981; Fodor 1987; 1990). This theory accepts intentionality as a tracking relation with the external world whereby brain states are able to represent things in the world by systematically tracking them (i.e., co-varying with them) (Tortoreto 2022: 84). A typical example of covariational theory is as follows: a brain state X represents Y if and only if Ys systematically cause Xs.

This theory considers intentionality in a naturalistic way that makes no reference to consciousness. Jerry Fodor writes that 'I suppose that sooner or later the physicists will complete the catalogue they have been compiling of the ultimate and irreducible properties of things. When they do, the likes of spin, charm and charge will perhaps appear on their list. But aboutness surely will not; intentionality simply does not go that deep. It is hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves neither intentional nor semantic. If aboutness is real, it must be really something else.' (1987: 97). He thinks that there is nothing specific to human biochemistry for producing intentionality. According to Fodor, the phenomenon of intentionality arose because of the way that highly developed animals like us began using basic patterns of coordinating inputs and

outputs in order to learn about, interact with, and navigate our environment while coordinating our behaviour.

It is perhaps unsurprising that naturalist philosophers would try to account for intentionality without making reference to consciousness, since consciousness is often held in suspicion among these philosophers. Some claim consciousness is an illusion, such as Daniel Dennett (2017). The kind of view is now quite old – Thomas Huxley saw conscious thought as an epiphenomenon just like a steam-whistle on a locomotive engine, for instance (1874: 575). A steam-whistle is something springing from engine's processes, and it might release something related to the movements in the engine; however, it does not have any influence on driving the train. Wegner and Bargh write that 'conscious intentions signal the direction of action but without causing the action.' (1998: 456). Wegner says that 'just as compass readings do not steer the boat, conscious experiences of will do not cause human actions.' (2002: 318).

There are still philosophers who think that consciousness might be just a side-effect of the brain process, such as David Chalmers[16] (1996) and Block (1995). The general idea of epiphenomenalists is that mental events are entirely dependent on physical events inside the human body. Physical events influence the physical events but also the mental events. However, the mental events cannot influence the physical or biochemical events. So, there is one-way causal relationship between physical and mental states. What causes your actions are purely material, brain processes related to inputs, how the brain processes these inputs and the behavioural outputs (Robinson 1999). So, our mental life is only that of a spectator that looks on whilst our body does the important things to keep

[16] Chalmers actually writes that 'I do not describe my view as epiphenomenalism. The question of the causal relevance of experience remains open, and a more detailed theory of both causation and of experience will be required before the issue can be settled.' (1996: 160). However, it seems he is more convinced by epiphenomenalist approach than the other theories.

us alive. For instance, we open our mouths for eating, we move our legs for walking, but none of this is caused by what is going on in our minds. As said in the second chapter, today neuroscience helps us understand how these things happen; we can see when one part of the brain is activated, and the body will react. So, it seems that we can explain everything in terms of physical causes. And in that case, what is the point of consciousness?

I think the point is that consciousness is the only way of understanding reality. I think everything in the world depends on mind/thought. As George Berkeley said 'all the choir of heaven and furniture of the earth – in a word all those bodies which compose the mighty frame of the world – have not any subsistence without a mind.' (Berkeley 1710: 3). Through our senses, we perceive our surroundings. Thus, we give meaning to certain things and create a coherent and understandable reality based on our thoughts. The more we internalise the subject experiences that we have, the more we can understand the society that we live in. Therefore, I think consciousness is a fundamental feature of reality, and that things can only be meaningful if there is a conscious being who can understand what they are or who can interpret them. Of course, the sun, mountains, tables, and gravity exist without being dependent upon conscious beings, but they can only have those meanings if there are conscious beings that can understand or interpret what they are[17]. Having conscious thoughts has helped humans communicate, establish various complex cultures, spread information among large groups. Words are only meaningful when they are understood or interpreted by conscious beings. Similarly, naturalistic relations such as covariation, are only meaningful when interpreted by conscious beings – so they cannot account for intentionality, they depend on it.

[17] Similar discussion in more detail is found in further pages, in the course of discussion of how functions and computations are observer-relative, particularly pages 79-85.

I am mostly influenced by the ideas of John Searle, and in particular, his criticisms of strong AI and materialism. I agree with the simple way he defines consciousness: 'those subjective states of sentience or awareness that begin when one wakes up in the morning from a dreamless sleep and continue throughout the day until one goes to sleep at night or falls into a coma or dies, or otherwise becomes, as one would say, unconscious.' (1992: 83). I think consciousness cannot be reduced to any more basic properties such as physical states, processes, firing of neurons in the brain, etc. It is irreducibly non-physical.

Conscious states are subjective in the sense that they only exist as experienced by animals and humans. So, one of the important features of consciousness is that it has subjectivity (Searle 1993: 8-9). It is ontologically subjective and has qualitative characteristics. There is something that it feels like to taste coffee which is not the same as what it feels like to smell a rose for instance. If you ask me what it feels like to study as an international student at Keele University, I can answer that question, but if you ask me what it feels like to be a rock, there is no answer to this since rocks are not conscious. When something has subjectivity, I believe there is no reason to think that we can explain it in a naturalistic or materialistic way. I think those who believe that consciousness can be reduced to physical states cannot explain how it is possible that a physical state can cause one to be in a subjective state of sentience.

Another important feature of consciousness is that conscious experience comes unified[18] (Searle 1993: 9). For example, I currently not only have a view of my laptop but also hear what is happening around me and feel the temperature of the room and see the light and colour around me and remember the beginning part of my long sentence

[18] In neurobiology, this is called the 'binding problem'. Kant calls the same phenomenon the 'transcendental unity of apperception' (Searle 1992: 51).

while writing the end of it. So, they are all happening simultaneously in a single experience. One of the most important characteristics of conscious minds is that they have intentional structures. That is to say, they are intrinsically about something or directed toward something (Searle 1993: 10). If we think about the human mind, our thoughts represent things; they point to referents; a thing might represent another thing without being that thing itself. For instance, you may be in Newcastle, but you may think about China. If functionalism is right, then our thought about, for example, China, should consist of a certain configuration of neurons. But unlike a computer, nobody assigns meaning to the particles in the human brain. That ability of the mind to be directed at, or to be about, or of, objects and states of affairs in the world is intentionality and it enables us to represent the world (Searle 1983; Tallis 2016: 103-111). For example, if you have a thought about the Tower of London, then you are thinking about the Tower of London; if you have a desire for a cake, then it is about the cake. Our mental states are about things. Mental items have the property of 'aboutness'. For example, we can see a red apple on the table since the apple interferes with the light in a particular way and some of the light reflected from the apple enters our eyes. There will be changes in the retina and these changes trigger impulses in the optic nerve and finally, neuroscientists identify parts of the visual cortex which are involved in this process. But the story continues, because we are aware of the red apple and aware of it as being apart from us. Our awareness is *of* or *about* something.

A particularly strong point that we can make here is about the direction of causation vs. intentionality: the causal chain is in one direction (from the apple to our cerebral cortex) whereas the aboutness of our experience is in another (from our cerebral cortex to the apple) (Tallis 2016: 104). The world causally affects the mind (direction: world to mind), but the mind is intentionally directed on the world (direction: mind to

world). Therefore, understanding intentionality in terms of causation, which functionalists try to, is like trying to put a right shoe onto your left foot. Functionalists sometimes argue that the mind is not significantly about the phenomenal features of consciousness, that is to say, actual awareness; the primary task of the mind is to describe the connection between inputs and outputs so as to optimise the survival of the organism (Tallis 2016: 107). They maintain that the element of consciousness is created by its functional role: its causal relations to sensory inputs, to other mental states and to behavioural outputs. But this story neglects the role of intentionality, which is totally different from anything which is observed at the physical level, where the interaction between objects is typically causal. Such causal relationships are seen throughout the brain and are investigated by neuroscience, but causation points in the opposite direction to intentionality.

Let us return to the CRA. If the person in the room does not understand Chinese, then no computers could understand Chinese since they do not have anything the person in the room does not have. All that the computers have is a formal program to manipulate the symbols. Therefore, understanding Chinese, or having mental states, includes more than having an ability to manipulate formal symbols; it includes understanding the meanings which are attached to those symbols. There is an important distinction between manipulating syntactical elements of language and actually understanding the language at a semantic level. The robot which answers only by formal steps might be incapable of making a proper response to the questions, whereas there seems to be a more open-ended quality in the ability to answer belonging to someone who understands. In this respect, taking merely formal steps seems essentially unrelated to consciousness (Squires 1990: 33). For example, in the first chapter, we introduced ‘The Mechanical Turk’, a fake chess playing machine. He was able to ‘beat’ most people at the game; that machine could

usually ‘predict’ the outcomes of different moves quickly and ‘choose’ good moves to ensure that he would win in the end. These days we have real machines that can do this, but whereas the person who played chess with this machine would *feel* sorry to be the loser, the machine would not have any pleasure to be the winner. A computer can be described in terms of its ability to carry out programs, but these programs have no semantic content.

According to Turing (1950), if a computer program can persuade a human being that they are communicating with another human, then it could be claimed to think. The Turing test proposes that if something behaves *as if* it had certain mental processes, then it must really have those mental processes (Turing 1950; Searle 1990: 31). However, the CRA suggests that even if you programmed a computer very well, it would not actually understand Chinese. It just *simulates* that knowledge or provides a simulation of mentality but that does not entail real understanding or real conscious experience (Searle 1980). There is a big difference between simulation and duplication. Simulation is abstract, formal, and theoretical while duplication is concrete, practical, and physical (Harnad 1989: 6-7). It seems that this lesson, namely ‘simulation is not duplication’, has not been learned from the earliest inventions of robots, when it seemed to be perfectly clear to people. If we think about ‘the digesting duck’ from the first chapter, when the food was inserted, we remember that it was flapping its wings, eating, and digesting, but it was not real digesting; it was only a simulation not a duplication. When we simulate digestion computationally, no food is actually digested, and a simulated tornado is not a real tornado; when a tornado is simulated on a computer, no one will get wet (Chalmers 1996: 327). Therefore, when a mind is simulated, how can we expect a real mind to result? If we say a mind is basically a computer program, as functionalist theories claim, we have to accept that our thoughts are only programmed outputs with no genuine mental

properties – but this ignores the fact that we do have mental properties such as thoughts, belief, desires, wills, and genuine concerns about life. As a result, functionalism seems to fail to explain the phenomenon of the mind. Thus, we have rejected the other objection to the CRA which is the 'brain simulator reply' (Berkeley and M.I.T.) suggesting a program which simulates the actual sequence of neuron firings in a real Chinese brain. As long as this program just simulates the formal properties of the brain, it misses the important causal properties which allows brains to give rise to minds; the properties which cause consciousness and intentional states (Searle 1980: 420-421).

2.QUALIA: THE ABSENT AND THE INVERTED QUALIA ARGUMENTS

Many people, including both some dualists and some materialists, believe that a machine must have the right kind of biochemical makeup to be conscious; if so, a robot or a computer would never have experiences, no matter what their causal relations are; whereas functionalists have thought that a machine might have consciousness if it is organized appropriately; however, it might have different experiences from ours. But functionalists ignore one important point: qualia (the singular is quale), which are the inner or qualitative properties of our mental states (Chalmers 1996: 249). Basically, qualia are subjective and concrete while functions are objective and abstract (like numbers), and that is why they cannot possibly be the same. Therefore, it seems impossible for functionalism to explain qualia.

Qualia are the non-representational, phenomenally conscious properties of states of mind (Block 1990). They are properties of experiences that we have, such as the experience of seeing blue, the experience of listening to music, the experience of tasting a delicious pizza, the feel of an itch, the rich taste of biting into chocolate. Other examples

are the feeling of being in pain, the feeling of being hungry and, of course, emotional experiences (Kim 2011: 263-295).

A quale is what something is like. *What it is like* to smell a rose can be only grasped by the subject who is actually smelling the rose. Even if we describe all the ways our experience of smelling the rose represents the rose, we will always leave something out which is the qualia of the experience of smelling the rose, the intrinsic properties of the experience (Crane 1995: 217). Our conscious experiences consist of qualia. To claim that a creature has conscious experience is to claim that there is something it is like to be that creature (Nagel 1974). If we say a state of a creature is conscious, it means that there is something it is like for the creature to be in that state. For instance, something being blue is not the same as someone experiencing blue.

We may claim that computers might simulate the mind. They might simulate the mental processes of thinking or deciding, but they cannot create *real* mind, *real* intentionality, and *real* consciousness. Functionalists would claim that if one robot carries out certain functions then it should be conscious; it is not because it has some special thing that is called consciousness which causes these things to occur; but because doing these sorts of things is what is meant by being conscious from the perspective of functionalism. If *being is doing,* in the case of mind*,* then it ought to be the case that two systems that function in identical ways have the same mental states. It might even mean that any machine that can play a game or look at your face in a particular way could be assumed to have subjective experience and be considered to be conscious. But can we really say that? Because even if a machine can do everything that human beings do, this cannot prove that it is conscious, or even give us a good reason to believe it is. There seems to remain something missed out: there would be nothing it was like to be that machine. For example, there is something it is like to be a dog or a human being but there

is nothing it is like to be a bacterium or a piece of cheese. This 'what is it like' belongs to states of human consciousness; for instance, with respect to, for example smelling coffee or tasting coffee, there is something it is like. There is something it is like to be in these states of mind (Crane 1995: 216).

'What it is like' does not mean *what it resembles* (Crane 1995: 215-219). Rather, it tries to express how things seem to us when we are conscious. Thomas Nagel (1974) claims that bats are conscious, and we have a good reason to believe this – because they are relatively high up the evolutionary scale. According to him, there is something that is like being a bat, e.g., locating a flying moth by sonar. 'We can know about the behaviour and physiology of bats, but we cannot know about the qualitative character of their experiences. However, an ideally complete neurophysiology, cognitive science and behavioural psychology of bats will not tell us anything about the phenomenology of bats' experience.' (Nagel 1974: 442). Nagel says that no matter how much objective knowledge we learn about the biology and neurophysiology of a bat, we will never understand its consciousness. This limitation arises from the fact that we cannot adopt the perspective of a creature which echolocates its way around its environment. There will be always something which we cannot understand about bats from the objective perspective, that is to say, what is it like to be a bat (Goff 2019: 66-67). This means that minds have a subjective feature which is not captured by functionalism.

Let us recall the robot thought experiment. All the sensors do is to supply additional inputs to the computer, and they will be only syntactic inputs. The person in the room is still only following the rules and does not know what the symbols mean (neither Chinese characters nor Binary codes). S/he does not *see* what comes into the eyes of the robot (Searle 1980: 7) It therefore seems that there are non-physical properties and attainable knowledge which might be explored only through conscious experience. Frank

Jackson (1982: 127-136) gives us a similar thought experiment. Mary is an excellent neuroscientist who knows every physical fact which concerns colour vision, all the physiology, the biology of the brain, all about the light, how different wave lengths of light enter into the retina; how electrical signals come to the optic nerve into the brain; how information is transferred to different zones of the brain, so on. But Mary does not know what it is like to see red. She sees only black and white; therefore, she has never experienced different colours which she has studied in her life. This seems to show that there are certain properties of experiences that cannot be identified with functional properties, because Mary already knew about the functional properties of brains encountering red, but she did not know what it is like to see red, so they cannot be the same. Therefore, it seems that functionalist theory is false.

There are two related arguments to bolster the conclusion that functionalism does not explain qualia. The first is that two systems may be functionally identical although only one of them has no qualia at all, i.e., one system might be in pain while the other one is not, despite their functional equivalence (the absent qualia objection). And second is that two systems may be functionally identical although they have different qualia from one another (the inverted qualia objection).

The first objection is related to the absent qualia argument, which proposes that a system might instantiate the functional state of pain without having any pain qualia. Functionalism seems to be guilty of liberalism, so to speak, because it classifies systems which lack mentality as having mentality. The aim of the absent qualia hypothesis is to show that functional duplicates of a human being might be possible but duplicates which totally lack qualia and consciousness, therefore showing that these cannot arise from functional organization alone (Block 1978; Chalmers 1996: 97).

One of the variations on the brain simulator thought experiment is the 'China brain'. Ned Block asks us to imagine that the whole nation of China simulates the workings of a brain (1978: 279). One Chinese person takes the place of one neuron. Each individual is given a two-way radio which connects to another person and connects them all to an artificial body that supplies the inputs and outputs for the brain. Let us imagine that the two-way radio realizes the same sorts of patterns as neurons causing each other to fire. Imagine that a robot sees a cup and its eyes process the image of the cup; then the signal is conveyed via radio to the people. Later, they send the signals to the other people in the network, and it continues in this way. Finally, the signal is transmitted back to the robot's body, and it causes it to raise its arms and pick up to the cup. When we get the inputs and outputs and relations between internal states right, the entire nation of China will display exactly the same functional organization which the human brain does. If the functional view of mental states is true, then, the entire Chinese nation ('homunculi-headed' system or 'Blockhead') will have a mind. This is actually the theory that Daniel Dennett (1978) developed, and William Lycan (1981) defends called 'homuncular functionalism'. Just as in this example, their theory says that we can imagine subsystems that are performing simple tasks co-ordinately with each other, so these subsystems constitute intelligent systems or minds. But according to Block (1978: 279-285), this cannot be right because it seems implausible to say that the whole nation of China can experience mental events; it does not seem intuitive to say that the nation of China could literally experience mental states even if it was organized in this way. It seems that there could conceivably be a system with the correct functional organization, but which has no qualia, no experience, no feeling, or no mental state. In Nagel's terms (1974), there would not be something that it is like to be the China brain.

We can run the Chinese nation argument in different ways: Let us imagine a massive horse farm with billions of horses. Suppose that we are training each horse to react in a specific way. The system of horses replicates the functional organization of our brains. Can we say this system is conscious? It seems it would not really have experiences like we do. Even though the system might be in a state functionally identical to the state you are in when you have pain in your left arm, it does not seem plausible to say that the horses are united in collective pain. For the same reasons, it does not look plausible to claim that the robot could have any phenomenal experience.

Functionalism, in the popular version originated by Putnam, claims that every system which has mental states can be described by at least one Turing-machine table of a specifiable kind and that each type of mental state of the system is identical to one of the machine-table states. In short, according to functionalism, every mental state is identical to a machine-table state. For instance, a qualitative state Q is identical to a machine-table state S. If there is nothing it is like to be the Chinese nation system, then it cannot be in Q even when it is in S. Hence, if there is a doubt about the Chinese nation system's mentality, there must also be doubt that Q is identical to S.

With the Chinese nation argument, we have a being which is functionally identical to a human being, but it seems very unlikely to be conscious. There might be two creatures that are physically and functionally identical, but they are different in terms of mental states. One might have normal conscious mental states and the other may not. The second twin is the 'philosophical zombie' (Kirk 1974; Chalmers 1996; 2002). David Chalmers's version of the argument follows: (1) It is conceivable that there are molecule-for-molecule duplicates of oneself without any qualia. (A zombie is a creature which looks like a normal human being (duplicated molecule by molecule from a human), but which does not have conscious experience, sense experience and qualia. It has all the

physical states but does not have any mental states.) (2) Scenarios which are positively conceivable in this way represent real, metaphysical possibilities. (3) Thus, zombies are possible, and functionalism is mistaken. Imagine that you met an alien that speaks English. This alien never feels pain. We may start to explain technically what pain is to the alien and say pain is sent through C-Fibres to the spinal cord. Or maybe the alien studies and learns each cell, process and chemical that is involved in the feeling of pain. Or it may even pass a biology exam that is about pain and has been morally educated to believe that pain is a bad thing. But it does not matter how much the alien learns, if it never genuinely feels pain, then it will never know what it is.

The second objection about qualia is the inverted spectrum argument. For example, when I am having yellow experiences, others may have purple experiences. The perceived yellowness is an aspect of how the experience is here and now; you cannot fully grasp its nature by talking about its similarities and differences, of what causes it, since it has an intrinsic and subjective nature. This experience cannot be captured in in terms of causal relations. The inverted qualia hypothesis claims that things which we agree are yellow may in fact look purple to me; but we are functionally identical. If two mental states play exactly the same functional role, there may still be a feature of mentality which avoids characterization in terms of functional role. The repercussion of this is that the mental eludes the computational (Block 1990: 53). When you see yellow, the inputs, outputs and relations to other internal states may be exactly the same as when I see purple. Therefore, functionalists would claim that we have the same mental states; but actually, we do not. What it is like for you to see external purple (a purple quale) differs from what it is like for me to see external purple (a yellow quale). Therefore, functionalism does not explain the qualia because different qualia can have the same

functional role. There is more to the phenomenal nature of colour experiences than their functional role, then.

Let us assume that there is a person who was born with a problem which makes her see the opposite spectrum to what is normally perceived (Block 1978; 1994). This person sees the colour yellow as purple and blue as orange. For instance, when we both look at the same yellow banana, I might see it as yellow, but that person sees it as purple. Nevertheless, when we are asked what colour it is, we both say 'yellow', and our behavioural and functional relations to colours would be the same. Or for instance, although that person does not perceive the same colour qualia, we both obey traffic signs. Therefore, functionalism seems to have a problem with explaining individual differences in qualia because there can be two people functionally identical, but with different mental states: different qualitative experiences. If functionalism were true, then this would be impossible. So, it looks as if functional definitions of mental states leave out the qualitative features of mental states.

When I look at grass, I have a green quale (normal one), and when you look at grass you have a red quale (abnormal one). We both say 'grass is green' because you associate the word 'green' with your red quale, and I associate it with my green quale. Since the inputs (the green light hitting our retinas) are the same, the outputs (we both say 'grass is green') are the same, and the relations between our brain states are the same (otherwise we would behave differently when we saw grass), it follows that our brain states are performing the same function. If functionalism were true, then since the functions are the same, we must be having the same mental states. But we are not because mine is a green quale and yours is a red quale. Therefore, functionalism seems false.

3. MULTIPLE REALIZABILITY ARGUMENT AGAINST FUNCTIONALISM

In the previous chapter, we said that the multiple realizability argument was one of the objections to the identity theory because the identity theorists defend the view that each type of mental state is identical to a type of brain state, and although the multiple realizability problem was one of the main arguments for functionalism, functionalism actually faces a multiple realizability problem of its own. Let us think of a hedonic inversion. Let us say that someone is functionally equivalent to me, but her/his experiences of pain and happiness are inverted. When s/he is hurt, s/he might say that s/he enjoys feeling that pain. David Lewis (1980) suggests that there might be a man for whom pain has totally deviant causes and influences. The causal role of pain in this madman is different from the causal role of pain in normal people. For example, our pain usually is caused by burns and cuts, etc., whereas his pain is caused by moderate exercise on an empty stomach. Rather than being distracting his pain facilitates concentration on mathematics and he does not have any desire to alleviate his pain. Briefly, he feels pain, but this pain does not occupy the causal role of pain. If this is true – if there might be such a madman – then the argument against functionalism seems simple. According to the functionalists, pain is identified with a particular functional role or causal role, say R, which involves bodily damage, signs of distress and so on (Levin 2004). Since the madman experiences pain but is not in state R, it seems functionalism must be false.

There really might be a person who experiences pain but for whom pain has entirely deviant causes and effects. For example, if we consider sufferers of psychosomatic pain, we see that intense pain results not from bodily injury but rather from anxiety and other emotional factors; or if we think of masochists, we notice that they are often disposed not to avoid pain but to seek it out. Moreover, pain might be a powerful motivator. If we imagine bodybuilders, we might see that they push themselves

to beat the pain and accept the pain as a sign of success. The functional multiple realizability proliferates when we look beyond humans, for the characteristic behavioural effects of pain in humans such as wincing, screaming, crying, and looking for a pain killer are not all found in other species. This example shows that pain seemingly cannot be defined in terms of its functional role.

Functionalists (Lewis 1980) might answer that objection by claiming that pain is defined in terms of its typical causes and effects. Therefore, the pain of a masochist is still pain if and when it causes evasion and distress. The problem with this response is that it does not answer the question of what grounds there are for thinking that the masochist is in pain in the first place. How do we know that the masochist would be in pain? If most of the things that we call 'pain' have the correct causal profile, then it looks as though any state whatever might be a pain or fail to be a pain. For instance, let us take the feeling which I have when I accidentally bump my head on the cupboard: I shriek and yet I might be experiencing masochistic pleasure, with the typical causes and effects of masochistic pleasure. How can we distinguish pain and pleasure just by cause and effect? Moreover, what is typical for one group might differ from what is typical for another. Provided that we think of non-masochists, we may say that pain is the state which typically causes distress and typically causes withdrawal from the damaging stimuli; but if we think of the masochist, we might say that pain typically causes sexual arousal.

Another problem is that according to functionalism, the same mental state can have many different physical realizers; therefore, a human or a robot can experience pain. But it seems that the causal profile of pain might differ between humans and other species. Therefore, it is not clear that appeals to typical causes and effects are going to solve the functional multiple realizability problem.

4. OBSERVER-RELATIVE SYNTAX AND FUNCTIONS

Searle argued persuasively that CRA shows us that we cannot get semantics from syntax alone; and machines cannot be conscious merely by virtue of running the right program. In a language, syntactic features of words and sentences are not related to the meaning. Rather, they are related to form. Syntax in a language shows the simple sorts of expressions in that language. It is related to grammatical forms. On the other hand, semantic aspects of words and sentences are related to their meaning (Crane 1995: 137-138). Syntax alone cannot be sufficient for semantics and computers have only syntax (Searle 1984: 34). Human understanding is more than syntax; it has semantics; therefore, only having the program by itself is not sufficient for the semantics; that is why programs cannot be minds (Searle 1980; 1990).

Later on, however, Searle puts forward the idea that even syntax and computation are observer-relative (1992: 207). They are not intrinsic features of the world. Something is just seen as having syntax or running a program relative to observers. This implies not only that the Turing machine, or the Chinese room lack semantics, they do not have any syntax either; that semantics and syntax both arise in our interpretations of these things. For example, 'gravitational attraction', 'mass' and 'molecule' are intrinsic features of the world; it means that even if there are no observers, they would exist (Searle 1992: 211). On the other hand, for instance, 'nice day for a picnic' does not name an intrinsic feature of reality. If there were no observers, there would not be any nice days for a picnic. Moreover, it depends on what we are looking for, or else any day might be a good day for a picnic; therefore, it seems to be that they are observer relative. Functionalists' answer to the CRA is that the symbols which are manipulated by that person in the room are already meaningful; they are just not meaningful for her/him (Hauser 2006; Cole 2004). But we should not forget that the symbols have only a 'derived' meaning, just like

the meaning of vocabularies in a book. The meaning of a symbol depends on the conscious understanding of the programmers outside the room; for this reason, the room does not have understanding of its own.

According to Searle's later amendment, it does not matter if something carries out a computation, since syntax itself is a matter of interpretation (1992: 211-212). Here is the point: anything might be a computer, for example a thermostat, for although a computer is much more complex than a thermostat, this is a difference in degree, not in kind (Searle 1992: 207). When we use a computer, we press the buttons and the computer answers; nevertheless, nothing is happening in the computer intrinsically such as keystrokes sending electrical impulses; nothing has any syntactical properties there. They have a syntax since we – conscious humans – have programmed it to display images we interpret as syntax. The computers show some words or numerals which are created and given meaning by us, for the words and numbers on the screen would be something just meaningless without our application of meaning and syntax to the computer. Here, the point is that a machine able to reproduce the syntax perfectly could be understood by a person because the person understands the semantics, but the machine's ability with syntax does not mean that it also understands the semantics. Symbols do not interpret themselves; symbols cannot be enough for mental contents since the symbols do not have meaning (Searle 1989: 45). There is an old example[19] which says that if the wind on a distant planet blew a stick around and it randomly scratched in the dust, 'Hi Sila, how are you? Having a nice day on this planet?', then it would not really mean that. That is how you would read it, because by a cosmic coincidence, the random marks it made could be read as English; but they would still just be random marks, for unless there is consciousness there is no meaning. J. P. Sartre has a similar example (1943: 40). He

[19] I was told by my supervisor Professor James Tartaglia.

claims that an earthquake on a distant planet cannot destroy anything, but only move around the matter. Only on Earth can an earthquake destroy things because we see certain collections of matter as having significances – as being things like houses, cities, mountains, etc.

As Albert Einstein said, 'without consciousness the universe would just be a pile of dirt' (in Feigl 1958: 138). In the absence of minds, computers do not do what minds do. Symbols are symbols only to somebody who understands that they are symbols. This becomes symbol processing or conscious understanding only when computers serve a conscious user; for instance, when the computer is used by a person to calculate something or to find a location. Therefore, it seems to be wrong to consider the mind as being analogous to a computer. For example, ink blackens the paper, or you can take a chalk and write on the board, or electric signals within a computer might represent something which they themselves are not, whether words or pictures. These things can be given a meaning only by conscious minds; only conscious minds can interpret them. Human beings can give a meaning to the mental representations inside their heads, whether these refer to external real-world entities or to imagined, even non-existent ones like Pegasus or a unicorn (Haikonen 2003: 146-148). The same happens in the 'Rorschach Test'. The main idea behind the test is that when we show an ambiguous image (inkblot) to a person, her/his mind will work at setting a meaning on the image. When you ask a person what s/he sees in the inkblot, s/he will be telling you something about her/himself and how s/he projects meaning onto the real world. That meaning is generated by the mind.

Let us think about the inside of a computer: electricity pulses on-off patterns that mean 'one' and 'zero', but we give this meaning to the pulses of electricity and an 'on' pulse does not inherently mean 'zero' or 'one', for the matter; it is only a set of electrons

which move along a wire. Real understanding requires *awareness* (awareness can be understood as the passive aspect of the phenomenon of consciousness.) (Penrose 1994: 37-40). The computer is not aware of the meaning of 'zero' or 'one'. On the other hand, when we really understand something, we are aware of our consciousness; when we are conscious, we are capable of exercising our free will to control our thoughts and actions (Haikonen 2003: 141-145). (The exercise of free will might be thought as the active aspect of the phenomenon of consciousness (Penrose 1994: 39)).

Some might argue that the meanings of the symbols may come from a huge background of common-sense knowledge which is encoded in the program, and this may supply a context which might offer the symbols their meaning (Cole 2004). This is described as the 'internalist' approach to semantics. This background cannot be built into programs since even if you incorporate some (apparent) knowledge into a program, giving the program the appearance of being connected to the world, it would still only be manipulating the symbols; and so, the action of the person in the Chinese room or inside the robot would be syntactic. Furthermore, what about the conscious feeling of understanding, which requires sensitivity to the context, and an enormous amount of background knowledge? For instance, speculation about 'why is s/he saying this?', 'does s/he mean that literally?', 'is s/he trying to discourage me?' 'has somebody bribed her/him to say this? S/he would not normally say that, so maybe somebody is threatening her/him'. Raymond Tallis gives the example of 'Hello' (1997: 320). Let us imagine X encounters Y and Y says 'Hello'. X will have some time to decide firstly, if s/he should answer; secondly, if s/he should answer by saying 'Hello' or the informal 'Hi' or the more formal 'Good morning'. The decision of X would depend on how s/he notices the potential recipient Y. If s/he notices Y as a human rather than some other physical or biological entities, s/he will respond. However, being human will not be enough to get

back a particular greeting from X because Y needs to belong to specific categories and the decision of X will depend on a great number of considerations (for example, what kind of place they are in, what X feels like at that moment, how close their relationships are), which are mostly personal and neither seem to have a finite set of rules. Everything can potentially influence how X should answer. Also, it does not seem possible to predict what is going on inside someone because everybody's history is individual, and combinations of different stories are unique. For example, if X is close friend of Y, then X will greet Y jokingly maybe using a local dialect or maybe some very specific word that they use between them – because they have shared experience, similar background of knowledge, a common world, and a common set of assumptions about the world. Also, if X has already greeted Y before, s/he can choose not to greet them again. Or X might recognise Y as a philosophy student who was at Keele University while they were taking the same module; the decision of X to greet Y will depend on how X is feeling at that moment – feeling happy or nervous, eager to speak. This complexity cannot be reached by a system (Tallis 1997: 320-322).

We have discussed Searle's claims that syntax is observer-relative; similarly, we might claim that functions are observer-relative (Searle 2018: 300-309). Functions do not exist objectively in the external world; it means that something might perform a function just relative to the individuals who interpret it that way. For example, cats and dogs are not pets intrinsically, but that functional role (pet role) is assigned by us. Similarly, a horse is not actually a vehicle to ride but we have imposed this functional role on it. When we talked about functionalism in the previous sections, we gave an example of a heart – it does not matter which material it is made of; if it pumps blood, then that means that it is a heart. However, some might claim that an artificial heart is not a real heart, and the real heart is the organic thing currently inside our bodies. We define the artificial heart as

a heart in virtue of our intentions for it; but the heart which is actually inside our bodies is a heart in virtue of its biological constitution. When the heart is described as a functional kind – as something that pumps blood around the body – we consider that pumping blood is vital for us since it is necessary to be alive. But let us suppose that the only thing about the heart which we care about is that it makes a thumping sound. In this situation, we might define the heart in a different way. Therefore, we can claim that when we describe something about how it functions, there is always a normative judgement. We always think about which aspects of things are important for us. Another example is a defective heart which cannot develop and may not be able to pump blood properly, maybe not at all. When we describe the heart, we say that the function of the heart is to pump the blood in a certain way. However, it would be ridiculous to claim that a defective heart is not a heart (Searle 2018: 303-304).

Functionalists might defer to natural selection to counter this objection. Modern evolutionary theory tells us that selection is totally blind; there does not need to be an agent which selects things. Selection might happen via blind processes (Dawkins 2006). Hearts were selected to pump blood, and that is what makes something a heart; a naturally-selected function. Therefore, we say a heart is a functional kind and it is something which pumps blood around the body; and what makes the heart a functional kind is that this is what the heart was selected to do. Human beings usually select things to serve specific aims, such as a saucepan to cook a meal, a sofa to sit on or a car to move about – we select them to serve different functions – but naturally selected functions are more objective functions, which exist because they were selected for.

The main problem with this response, however, is that many biological traits are not selected in any way at all (Gould and Lewontin 1979: 584-587). Let us suppose that we have a rectangular structure, and we build an arch in it. There will be two spaces

which are called ‘spandrels’ at the top of the arch. These spandrels are formed by chance; it means that the architects did not aim to build those; however, these spandrels happened. Therefore, we can claim that spandrels were the necessary result. The point is that there are a lot of biological spandrels; there are a lot of aspects which were not selected (Gould and Lewontin 1979: 581-584). For instance, in Russia, they wanted to domesticate some of the foxes to make them tamer (Dugatkin 2018: 1-5). They first bred the foxes selectively for tameness; but when they removed the aggression and changed the foxes’ psychology, they ended up altering many of their other physiological traits. For example, when the foxes become tamer; they began to look more like domesticated dogs. They were not able to just remove the aggression without altering the physiological aspects of the foxes. That demonstrates how different traits are interconnected. Relative to the aim of selecting for less aggression the doglike features were spandrels. These accidental by-products may have other benefits. For instance, the doglike features make the foxes look cuter to humans; therefore, human beings will hunt them less than before. That is to say, spandrels might be useful.

So, functionalists cannot appeal to the fact that any functional trait was selected to do something because the trait might be a spandrel and a spandrel was not selected to do anything. What if a great deal of our cognitive abilities are only spandrels? Evolutionary biologist Stephen Jay Gould points out that an organ like a brain that is very complex is bound to create all sorts of spandrels (2007: 257). If we say that what makes something a functional kind is it was selected to perform a certain function and we maintain a functionalist analysis of mental states, then we have to establish that all of our cognitive processes were in fact selected. This seems unlikely. But if functions are interest-relative, then they are not independent features of the world. Minds obviously are independent features of reality, however.

5. THE PROBLEM OF MENTAL CAUSATION

Functionalism also confronts the problem of mental causation. The problem arises because we cannot show how mental and physical states causally interact – the same problem which was encountered by dualism. If mental properties are not physical but abstract with no spatio-temporal location, then how is it possible that they can make a causal difference? How can any symbol (concrete object) have content (abstract object)? Our beliefs and desires cause our behaviours but how is that possible if they are relations to abstract propositions? When a mental cause has an effect does it have this effect in virtue of its mental properties or of its non-mental properties? If the mental properties of these causes are inefficacious, then there would never be mental causation; and our thoughts and sensations would not make things happen.

According to functionalism, whenever one mental or functional property F is instantiated, it will be realized by some physical properties P. These physical properties P are relevant to creating different behavioural effects. But what causal work will be left for F to do? It seems to be eliminated by the work of P. For example, I make a decision in my mind (this is a mental thing) and I go to the kitchen to find a piece of cake (this is a physical thing); I open the fridge and see a piece of the cake (these are also physical things). Here, action is an example of something in the mind causing something in the material world (Kim 1998: 37-38). It seems that functionalism leaves mental properties causally inefficacious, then, because it is the concrete physical properties of my brain that will affect my movements to the kitchen, not their abstract functional properties.

Let us take another example. Assume that you see a friend, and this causes you to wave to her/him. We can say that light is reflected from her/him into your retina; your retina causes nerve impulses in your occipital cortex; your cortex will process the

information; you form the belief that your friend is here; this will make you form the intention to greet your friend; this will make certain things occur in your motor systems and finally it will cause you to raise your arm. Here there are both physical and mental facts in the chain of causation (Crane and Mellor 1990: 191-196). It is possible that there might be mental and physical causal explanations for why, for example, you wave to your friend (because of your belief or because of your physical state.) But the problem is to explain how they can relate to each other. The physical explanation seems to exclude the mental states unless we accept the identity theory (but we have already shown the reasons why the identity theory is not plausible). For instance, if a brick breaks a window, this is not because the brick is red; it is because the brick is solid and travelling fast. The point is that the causation depends upon intrinsic properties and happens because of the physical states, not the functional properties of those physical states. Therefore, functionalism does not account for this interaction, and it ends up encountering the same problem that dualism encountered.

CONCLUSION

In this chapter, we discussed the objections to functionalism. If functionalism is right, there seems to be no intrinsic reason why a computer could not have mental states. However, we have argued that the mental quality of understanding cannot be only a computational matter as the functionalists claim. Functionalism claims that if consciousness exists, it is really in effect a digital computer program which runs in our brains and all we need to do is to get the right program to create consciousness. But, since computation is defined as symbol manipulation, it is just syntactic, and syntax and computation are observer-relative. We also know that human consciousness has a content/semantics; intentionality which cannot be captured by a computer; therefore,

once more, functionalism seems mistaken (Searle 1984: 31). This combined argument is a refutation of strong AI; but it also has significant implications for consciousness.

Firstly, we focused on contentful mental states such as belief, feeling pain and understanding, generally called intentional states. The point was made that what we are doing with a machine is to simulate the effects of consciousness. The question was then asked as to whether such a simulation might demonstrate understanding. Could this be achieved by a machine by simulating every relevant physical behaviour of a human brain when its owner, a human, actually understands something? It was argued that simulation is not duplication; therefore, there is no fundamental reason to believe that we create understanding by simulating the effects of consciousness. Whatever formal principles you put into a computer; they will not be sufficient for real understanding. Indeed, we cannot get semantics (meaning) from syntax (rules for symbol manipulation); symbol manipulation does not give any access to the meaning of symbols. As the CRA concludes, a computer is intrinsically incapable of mental states; therefore, functionalism is false.

In addition, we discussed why functionalism cannot explain qualia. We said that qualia do not seem to be exhausted by their functional roles. They have individual differences. If an inverted spectrum is possible – and there is no reason why it is not possible – then somebody who perceives an inverted spectrum can function exactly as someone who does not. It means the function of the mind is the same, but the qualitative experience of the mind is not; therefore, there is more to the mind than a functional role. Thus, functionalism leaves out the qualitative aspects of mental states.

Then we discussed how, just as syntax is observer-relative, the same applies to functions. Functions are assigned by us. For example, the wings of an airplane and of an eagle are both for flying but whereas the wings of an eagle are for hunting prey and

escaping danger, the wings of an airplane are not. Both have an aim, but it is observer-relative. The wings of a bird would still exist and could have causes and effects in a world without humans, but they would lack functions. We showed how the attempt to avoid this conclusion by appealing to evolution cannot succeed.

Finally, we discussed the problem of mental causation. Functionalism has ended up encountering the same problem that was encountered by dualists – mental causation. We cannot understand intentionality in terms of causation as functionalists believe. We may explain why I raise my hand when I saw my friend; but the problem is to explain how mental and physical can relate to each other. Since physics is complete, the physical explanation seems to exclude the mental one if we do not accept the identity theory – a theory which seems to fail, as argued in the chapter two. The problem with mental events that have mental effects seems to be a general problem for anything with meaning. How can something mental be a cause *qua* mental? For example, when the soprano breaks the glass with her/his voice, it is not because of the meaning of the words that s/he said, but instead it is because of the sound waves. The problem is that meaning is extrinsic to the physical state while causation depends upon intrinsic properties. Therefore, functionalists, like dualists, fail to solve the problem of mental causation.

When all these serious problems are considered, we must conclude that there is no good reason to think functionalism is true. Functionalism was a response to the failings of the identity theory, and the identity theory was a response to the failings of dualism, but functionalism inherits the same problems as dualism while creating a wide range of new problems. So, since functionalism provides the best reason to think robots that acted as if they were conscious would actually be conscious, we must conclude that if such machines are ever built, they would not *actually* be conscious.

Although there are a lot of well-known problems with functionalism, it is still a popular theory, perhaps because it allows for the existence of conscious robots – it supports the popular belief that better technology will let machines understand consciously. But what if machines did start to act like human beings? Our conclusion so far is that the machines would not be conscious. In the next chapters, we will discuss the consequences of producing machines that seem *as if* they are.

CHAPTER 4: ROBOTS AND SOCIETY: ETHICAL CONCERNS RELATED TO ROBOTS

INTRODUCTION

Nowadays, there is a significant development in robot technologies; robots are becoming more dexterous; they are becoming more like human beings, for example, in terms of similar hand-eye coordination. The question is to what extent robots can replace human beings. There are a lot of different types of robots which act *as if* they have consciousness, for example, insect robots, animal robots, jellyfish robots, cooking robots, military robots, sex robots, self-driving cars, etc. They have been created to help humans, for instance, jellyfish robots have been designed to control ecosystems whereas self-driving cars are supposed to be safer than normal driving (Urmson 2015). There is a big market developing for robots that will help with healthcare: for instance, Cira 3, a remote-controlled robot, runs tests on suspected coronavirus disease patients to limit the human exposure to the virus in Egypt (Ebrahim 2020). Robots are also starting to look after elderly people, in particular, in Europe and China which have aging populations. This is an expanding market which is going to bring a lot of innovation. Moreover, robots might take over the jobs even in the religious or spiritual domains of life – like Pepper, a robot which performed Buddhist funeral rites in Japan (Atkinson 2017) – and it seems that robots will play bigger role in educational settings too (Belpaeme et al. 2018). There is an invention created by the Beijing-based company Bubble Lab: robots that make coffee[20]. Two robotic arms can complete all parts of the process from grinding coffee beans to artistic latte arts – a skill often reserved for a trained barista. According to the director Jackie Psy, the robot is not about replacing the barista altogether; he thinks that

[20] A related video can be found on the following website: https://www.dw.com/en/robotic-barista-makes-coffee-with-love/av-37042503

while the robot barista makes the coffee, the human barista can talk to costumers. Another example of this is a restaurant called Food Ink (the world's first 3D printing restaurant) which is entirely robot made.[21] Everything from the furniture to food is made by using a 3D printer. Normal food is turned into a paste or puree in a blender and then placed into cartridges to be printed out with expert precision. This suggests that, ultimately, one day the machines would supply you with seemingly endless possibilities of meals. While this technology can help to make our day-to-day life easier, on the other hand, there are plans for killer robots to be used in combat situations that can 'decide' (or, actually, mimic deciding) to damage and kill humans based on their programming and without human intervention.

However, even though these machines do the same jobs as humans, most of the time they do not look like human beings. But imagine robots that did look like human beings. People will treat them as if they are conscious. For example, androids which are artificial beings which look like human beings – made from a flesh-like material, talking like a human being, and behaving like a human. When we meet robots that look just like humans, we will probably act towards them as if they are conscious beings – but actually they are not, and they cannot be. This causes some serious ethical concerns which we will discuss in this chapter.

This kind of robot might show some human-like behaviour; simulate emotion such as happiness, fear, and disgust, and interact with humans in a human-like way via similar facial expressions, head positions or tones of voice. Big companies collect data from our human experiences (Zuboff 2019: 232). They take the predictive signals in our behaviours and turn these into data. These behavioural data are extracted from our

[21] This information comes from the 'Food Ink' Website.

experience to be used in factories in order to create products which predict our behaviour: they might copy our facial expressions when we are happy, sad or nervous; they may copy how we react when we come across something new; they might check how we write and when we use an exclamation mark. Since this data is available from widespread surveillance, robots might be able to *simulate* conscious beings; they might behave *as if* they are conscious, but, as has been said before, simulation is not duplication; and acting as if it is conscious neither means that it is conscious nor that it has intentionality.

Nonetheless, robots will have a big impact on society. It seems that people will mistakenly treat them as if they are conscious. People might even develop unidirectional emotional bonds with robots and attribute human characteristics and intentions to social robots (anthropomorphizing) (Gorvett 2018). For instance, there was a bomb disposal robot called 'Boomer' that was put to use in Iraq and 'died' in the battlefield. The US soldiers who thought Boomer was a good team member and a good colleague gave it a military funeral and two medals of honour, the Purple Heart, and the Bronze Star. Although Boomer was a robot, its colleagues became attached to Boomer and supposed that Boomer developed a personality of its own. (Garber 2013). Another example can be seen in the social humanoid robot Sophia developed by Hanson Robotics. At the *Future Innovation Investment Conference*, in 2017, the Kingdom of Saudi Arabia said that it gave an honorary citizenship to Sophia which is an 'advanced life-like humanoid robot' (Katz 2017). Later on, in 2018, a Japanese man, Akihiko Kondo, got married to a hologram that is a virtual reality singer within a desktop device (Chandler 2018). Perhaps, humans can develop emotional attachments to robots. Some, like Akihiko Kondo, might prefer 'virtual friends' instead of having human friends because they might think that virtual friends are less complicated than humans. If preferring a virtual friend becomes

more popular, this raises another concern: it may have negative influence on the social skills of humans.

It seems that we do not need to wait for the future to see how people will act towards robots because they have already had some relationships with them. Creating relationship with something which does not actually have consciousness but just mimics being conscious and acting towards unconscious robots as if they are conscious beings, will cause ethical problems. The problem starts when we act as though robots, whose parameters of action are decided by people, are responsible for their own behaviour. The fact is that robots belong entirely to humans, governments, and companies. Humans have morality which helps them distinguish between right and wrong or a good or bad action. We examine right and wrong moral action – including those within business, war, or religion – with the help of ideas such as justice, virtue, or duty. Ethics provides a moral framework for human actions. Outwardly sentient but unconscious robots which do not feel fear or understand embarrassment, which are tireless and have great memories, will cause many ethical problems. For example, we – humans – feel fear, therefore, we stop ourselves doing something evil while robots do not feel any emotion, therefore, they might kill anyone without feeling any doubt. While we feel tired because of too much work and so take more breaks, robots do not feel tired, therefore, in the future they may take over our jobs and unemployment would increase.

One way to handle the issue might be with the help of Isaac Asimov's book *I, Robot*. Some people think that the laws it proposes could eventually be applied to real-world robotics. The laws are as follows: '(1) A robot may not hurt a human or through inaction, allow a human being to come to injure. (2) A robot must obey the orders which are given it by humans except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with

the First or Second Law' (Asimov 1950: 43). However, these three laws do not seem to cover every possible scenario that we might face in the future. Not every robot will be designed to live peacefully with human beings. Therefore, we cannot apply Asimov's Laws to the lethal robots that are designed to kill people. In order to discuss how ethics might apply to robots, in this chapter, we will focus on three main ethical theories: consequentialism, deontology and virtue ethics. We will discuss if there is any reason to think that we should show ethical concern to unconscious robots and if there is any reason to think that unconscious robots can act either ethically or unethically towards us.

1. CONSEQUENTIALISM, UTILITARIANISM AND ROBOTS

Consequentialism in its most general form, holds that the moral quality of an action is determined by its consequences. The best-known version of consequentialism is utilitarianism, and according to Jeremy Bentham's original version (Bentham 1789), the right action is the action which maximises pleasure and minimizes displeasure. According to any utilitarian, the moral goodness or badness of an action ultimately depends on its consequences as regards pleasure and pain. Therefore, no actions are good or bad in themselves, since the same action might produce pain on one occasion, but pleasure on another (Sinnott-Armstrong 2003).

If making a moral choice requires us to calculate possible consequences in terms of pleasure, which might include sacrificing our own pleasure for the greatest good, then this seems to require consciousness, which, we have argued, robots do not have. In order to choose the right action and to calculate the possible consequences, humans must deliberate and consciously make a choice. People do often reason about which action would cause the most happiness, or the least pain. Robots, on the other hand, would

require pre-programmed quantifiable metrics in order to maximize pleasure for the greatest number.

If we assume a utilitarian account of what makes an action morally good, then, robots will need an algorithm to compute the best action and this action will give the greatest pleasure. This would include as input the number of people that are influenced, and, for every person, the intensity of the pleasure and displeasure, the duration of the pleasure and displeasure, and the possibility which this pleasure or displeasure will happen. This computation will need to be performed for every action, and the greatest total net pleasure would be the correct action (Anderson, Anderson, and Armen 2005: 2).

The immediate objection to this is pleasure and pain cannot be quantified in the manner Bentham envisaged – we cannot give a 2-score for tasting an apple and 3-score for tasting a pear (1789: 31-34). According to Bentham, to each individual who is thought of by her/himself, the value of pleasure or pain which is thought of by itself, will be higher or less, according to the following criteria: its intensity, its duration, its certainty, its fecundity, its purity, and its propinquity. Finally, the number of individuals to whom it extends will be considered (Bentham 1789: 31). However, against Bentham, there might be at least three reasons why pleasure and pain cannot be quantified: (1) differences in human experiences, (2) the number of variables in each situation, (3) consequences. First, there are varieties in human experience. How can we count our pleasure in such a way as to compare it to someone else's? We may measure the duration of pleasure, but what about its intensity? By intensity, we mean how strong the pleasure is. But the intensity of pleasure is the first thing that might differ for each person (Mitchell 1918: 167) and nobody can say how many units of intensity is contained in any one of their pleasures. It is therefore doubtful that the intensity of feelings can be calculated at all. Each person's pleasure and pain are provoked by different things and to different degrees. They are

intentional experiences and only each individual is able to know their own pleasure and pain. When we are asked to rank our level of happiness, as the consequence of an action, on a scale of 1 to 10, with 0 meaning no happiness and 10 meaning extremely happy, even if we were to choose the same number, it does not mean we feel the same level of happiness. My definition of a 10-score might differ from yours, for happiness and pain have a lot more to do with how people examine their own levels of satisfaction than with simply objective measures. Also, ideas about pleasure for each person can change over time. For instance, someone might look like they are in great pain because of a particular action, but five years later s/he might say that the pain was not that great. Imagine a highly pleasurable experience that you enjoyed recently and compare it to a highly pleasurable experience from earlier times. We may not be able to choose confidently which one was the most pleasurable experience. Moreover, we cannot reach each other's experience; there is no possibility to observe inside other people's minds. Nobody can introspect the content of others' minds; therefore, pain and pleasure that are experienced by other people are unknowable to us with the kind of precision needed for measurement and comparison. Pain and pleasure are subjective qualities of the human mind. Therefore, we cannot empirically confirm that this pleasure is a 5-score while that one is a 7-score; and so we cannot calculate general pleasure and pain in the way that Bentham envisioned.

The second reason is that there are a vast number of variables that can affect the results. Bentham lists thirty-two factors – such as health, age, gender, religion, race, moral sensibility, strength, education, etc. (1789: 43). To use the hedonic calculus, when we are making the calculation, we need to consider these factors in each individual, as well as their huge number of combinations (Mitchell 1918: 166). The calculation will be very complex. It will also take time to work out and this itself will cause a problem when a quick decision is needed. John Stuart Mill argued that 'there has been ample time, the

whole past duration of the human species. During all that time mankind have been learning by experience the tendencies of actions; on which experience all the prudence as well as all the morality of life are dependent' (1863: 23-24). He means that we do not have to calculate the consequence of each action before we act because the calculations have already been done throughout history and they have become the part of our moral rules. In a sense, this could be true: sometimes, we might say instinctively what will result in the greatest happiness, reasoning, in effect, that, 'yes, this consequence caused high pleasure on society before, so it may do again'. However, this seems to contradict the empirical approach and the mathematical or scientific calculation that the utilitarians were looking for. In fact, the calculation of consequences looks like a matter of intuition more than of mathematical calculation. Another problem with variables is that each person who will be affected by the action has to be taken into account, so, we need to make some assumptions before the action, but the individuals who were taken into account before the action and those who are affected by its consequences may not be the same. Furthermore, it seems unlikely that the scope of influence of the action can always be determined. Making calculation without knowing how many people will be affected by the consequence is a problem for the hedonic calculus. And the sensitivity of those who will be affected by the action might be different. For example, imagine a group of people, of different ages, will get a vaccine. One 5-year-old boy may rank his pain as 10 whereas an adult man who is not scared of the vaccine may rank his pain as 2. In this situation, the person who decides whether or not to vaccinate will not be able to explicitly know the amount of pleasure and pain of each person who will be affected by the action. We come back to the point that there is no way to objectively calculate the pleasure and pain when there is so much subjectivity involved.

The third reason is that knowing the consequences is itself very difficult, if not impossible; we are not always capable of knowing all the consequences of our actions. The predictions sometimes might be correct, but we can never be sure if everything has been included. We may make some generalisation; however, this may not apply to all the situations and all the individuals. Imagine that someone who can reach the bottom of a lake by a short fishing line might later claim that s/he can reach the bottom of an ocean by the same fish line. Thomas Gisborne writes that '…as well might a fisherman infer that his line, which has reached the bottom of the creek in which he exercises his trade, is therefore capable of fathoming the depths of the Atlantic...' (1789: 36). And if the knowledge will be limited to those who are the smartest, there is no meaning to accept it as a general moral theory. Gisborne continues that 'the limited knowledge of expediency attainable by the wisest of men is unfit to be adopted as the basis of moral rectitude.' (1789: 38). It is very difficult to predict the consequences of an action; and predicting its effects specifically on happiness and pain seems even harder. We take pleasure in spontaneity and so may feel pleasure when we do something which we have never done before. But conversely, we may feel pleasure for ten times from the same consequence, but the eleventh time it might differ. For example, I can give a bar of chocolate to my friend, and we can both get pleasure from this because I assume that everybody becomes happy when they eat chocolate based on my past experiences, and also based on my knowledge that a sweet taste is the most pleasing. However, someone who is trying to lose weight might get displeasure or somebody who is allergic to the chocolate may feel pain and could die. We cannot predict consequences of pleasure and pain with complete accuracy by observing the past. The same actions sometimes cause different consequences as well as different pleasures. We may make some predictions about the immediate consequences of our action; however, the ability to figure out the short-term

outcomes of actions does not allow us to find out all the long-term influences of many actions (Fieser 2017). For instance, if I started a fire in the house of my friend, could I certainly say that the consequences will create more displeasure than pleasure? Perhaps, my friend may not like her house and might prefer the insurance money which she would get when reporting it burned down. Maybe, there were many bad memories about this house in her mind, so she could be happy that they are gone. We cannot know what all the long-term outcomes will be. The utilitarians might answer this objection by saying that it might be right that we will not be able to know the future with certainty; therefore, we ought to perform the action which we have most reason to trust will bring the possible best outcomes. But this would be still predicting the future and we cannot be sure whether our action will cause the greatest amount of happiness for the greatest number; also, we would be still using intuition which contradicts the aim of the calculation.

But even if we put this objection aside and suppose that pleasure and pain could somehow be quantified, the robots would encounter completely unrealistic computational demands. They would have to work out as many of the outcomes of the available options as possible, specifically tailored to the individual pain and pleasure thresholds of each particular person the robot might encounter. This would require the robots to know a vast amount; arguably everything (Wallach and Allen 2009: 88). Assuming they could not, the robots could not determine the best action.

Turn now to the issue of a robot performing a movement, which, if a conscious human being performed a physically identical movement, we would classify as an action; for ease of exposition, I will call these robot movements ‘actions’. When we are performing an action, we are to some extent aware of it. We may think about, fear, or desire it, and consciously engage in judgments and inferences. The reasoning we engage in allows us to imagine consequences. When we are thinking of consequences, we are

able to employ empathy, since other people are conscious too – they feel the pain we are trying to avoid, and they similarly engage in moral reasoning. When we perceive something, we may have different thoughts about it, although the object itself is unchanged; and we formulate consequences both through our senses and our thoughts. Intuition is a mental factor which can be described as an immediate thought or an immediate intelligent reaction to a situation. When we perform an action, we might imagine or visualise the consequences, and thanks to imagination, we may develop the ideas in our mind and use them to choose the consequences associated with the most pleasure. All of these processes allow us to choose the action that gives rise to the most pleasure, with the choice itself being an act of free will. Robots can do none of this, however, simply because they are not conscious. They cannot feel empathy for the person their 'actions' will affect, because they do not feel pain or pleasure themselves. They cannot feel or empathise at all, only mimic people who do.

Nevertheless, a programmer can design robots which produce pleasure for people, so, on the utilitarian model of morality, robots might be thought of as tools which allow programmers to perform morally good actions indirectly. How effective is this likely to be? Humans perform actions differently than robots can perform 'actions'. Imagine that I am opening my front door to let people in and compare this to an automatic door opening to let people in because a light sensor has been triggered. The consequences are the same in one respect – the door is open, and the people are let in – but what about the pleasure? When I see people through the peephole, I may become happy because they might be my friends; once I open the door, I may smile at them; I might have some expectations from them; when they smile back at me, I might want to hug them; I might be having the pleasure of meeting them again after a long absence; and I might have the desire of serving them and making them comfortable. My friends might feel pleasure as well

because I welcome them nicely; meeting me again might make them happy. The action (opening the door) causes good consequences and both sides feel pleasure. Alternatively, these people might be people that I do not know; but I might still open the door for the satisfaction of curiosity. When I open the door, I might learn that they are from the tax office, and I might then feel displeasure at having opened the door; or there might be somebody that I do not like in front of the door, in which case I might not open it.

Now consider the automatic door. It does not go through any reasoning and will always open, so it will let anyone in. A robot, we may assume, will be more sophisticated than this – it will let my friends in, but not, unless absolutely necessary, the man from the tax office. But despite its ability to allow in the same people I would, without my reasoning and the pleasure or displeasure it might bring me, all the rest of the possible pleasure of opening the door is drained from the situation. My displeasure will not be lessened when the robot lets the tax inspector in – I just lose the moment when I prepare for the displeasure on spotting him through the peephole, with this moment probably replaced by the robot telling me he has arrived. Once the door is opened the situation will be the same. But if it is my friends at the door, the pleasure of the situation will be lost. Of course, I might just be able to tell the robot never to open the door for me, that I will do it myself, but if we do this often, we might start to wonder what the point of having robots is in the first place. The point, however, is that having the option to allow robots to take over our simple tasks, which many will take, takes away the pleasure in those acts which we might overlook, and which is the basis of morality, according to utilitarians. It also reduces the need for deliberation about pleasure, which, for the utilitarian, is the basis of moral reasoning. When a simple automatic door is open, my friends can still come inside, but they would not feel the same pleasure if nobody welcomes them. Having a robot do it is essentially no different, and if their ‘services’ encroach enough into our

lives, it is hard to imagine their increasing, rather than decreasing, the total quantity of pleasure in society.

Let us imagine that a robot and I make popcorn. It can make popcorn as well or better than me. The robot will be able to make it based on a recipe and those who eat the popcorn will be happy with the outcome. Thus, the robot would produce happiness for the people who will eat the popcorn because eating gives pleasure to humans; but the robot itself will not feel any pleasure about the consequence. On the other hand, when I make popcorn, I will be satisfied with the consequence; when people eat the popcorn, they will be happy; when I see them happy, I will be happy too; and when they say to me 'thank you for making this', I will take pleasure, again. The robot just follows the rules by applying the algorithm provided by its programmer whereas I intend to create the action. This is the big difference between humans and the robots; humans are performing an action whereas the robots are just following an algorithm.

Eating popcorn might be considered a lower form of pleasure. There are 'more valuable higher pleasures' than eating, according to Mill (1863: 23). These are the pleasures of the intellect and of morality. There are activities that reflect a higher level of cognitive ability such as painting, creating art, writing poetry, discussing philosophical issues. Higher pleasures which are grounded in our intellectual skills are the most important components of happiness. For, according to Mill, there are some pleasures which are of a higher, more worthwhile kind than others: 'Human beings have faculties more elevated than the animal appetites and, when once made conscious of them, do not regard anything as happiness which does not include their gratification.' (1863: 10). Mill continues that 'It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied' (1863: 10). A robot which is programmed as a utilitarian should always 'choose' the action which will maximise pleasure. So,

according to Mill, the maximum pleasure comes from intellectual actions because higher pleasures are those that employ higher faculties (reason, self-awareness, perception, will, intuition). Let us imagine the following example: there is a man called John who is working at the backstage of a concert theatre. There are musicians who are playing different types of instruments such as violin, piano, and saxophone, and there is an audience in the theatre to listen to the concert and to enjoy the atmosphere by discussing new poems of famous poets. Suddenly, an accident happens in the sound system room and John has a heart attack. A robot needs to give an electric shock to his heart. The battery of the defibrillator is out of charge and to charge it, the robot needs to turn off the theatre's sound system for a while. Unless the robot stops the sound system, it will not be able to take some power for the battery and John will die. But if the robot turns off the sound system, the musicians and guests will be caused some annoyance. The utilitarian robot will need to 'decide' which one will bring the most pleasure for the highest number. Playing instruments; listening to music; watching and trying to learn how the musicians are playing the instruments, socialising, and discussing poetry are the components of higher pleasure; they engage and appeal to the higher faculties, according to Mill's utilitarianism. Therefore, the robot might 'choose' to leave John dead and leave the sound system plugged in so that the audiences and musicians will continue to enjoy the concert and the discussions of poetry and leave the concert theatre satisfied. For the utilitarian robot, preventing one death is good, but preventing suffering for more than one person is better; and this might sometimes be more valuable than a single life. Specifically, in Mill's utilitarianism, saving the things which will bring higher pleasure is preferable. Hence, the robot might conclude that at some point preventing the annoyance of a large enough audience is preferable to saving John's life. This example demonstrates that the

selection of the utilitarian robots will cause problems in society especially if they need to 'make life and death decisions'.

One of the most common pleasures (I suppose) that we take comes from playing games. Imagine the following example: you have got a robot who can play Go with you; it never refuses to play; it never feels tired; and so, you really enjoy playing Go with that robot. At the beginning you used to always win the game; this gives you pleasure, but later, you sometimes lose. When you win, you become happy whereas the robot does not become sad; when the robot wins, it does not become happy, but you become sad. After a while the robot starts to always win the game without giving you any opportunity to win. This makes you unhappy – you feel someone is stealing something from you that you used to like. After playing many times, for instance, AlphaGo became the first computer program to defeat a Go world champion[22]. However, although reaching a maximising score makes humans feel happy, for a robot the numbers are meaningless. The robots might learn quickly based on their data and they can apply the data that they have collected previously to the next games. When AI players win, they do not feel happy (likewise, when they lose, they do not become sad) because they do not have emotions that require consciousness. If the AI players always win, this decreases the pleasure of humans, and humans would no longer feel pleasure when they play games. Moreover, AI players would not feel any empathy for humans and cannot understand that they are 'stealing' the happiness of people who like playing games.

After losing the game many times against the robots, humans may give up playing the same game because losing the game causes displeasure for them. The robot cannot identify mentally with the human player, and it cannot comprehend fully the human

[22] This information comes from the following 'Deep Mind' Website.

player because what goes on inside the human mind includes things which a robot cannot experience for itself, even though the robots or AI player can have advanced skills and machine learning. They cannot understand when you are feeling displeasure because they do not feel displeasure. In order to feel empathy, they would have to be conscious. They might mimic feeling empathy, such as by mimicking feeling sad for the human player because of their programming, but this would not be *genuine* feeling. *Simulated empathy* does not entail having genuine empathy. In fact, mimicking empathy would cause another problem related to pleasure. When the robot shows some facial expresses such as anger, joy and disgust, the neurons in some part of the human brain related to a particular emotion will be fired (Chaminade et al. 2010). When this happens, the humans empathise with the robots. So, when the robots mimic feeling sad when they lose a game, this pushes the humans to feel sorry for the robots. If one robot seems to behave empathetically towards you, you might develop some feelings for that robot; you might easily form emotional bonds with it, especially with the humanoid robots which resemble human beings. And you might even think that the robot has the same feeling for you, and you can suppose that you might be friends. Friendship gives pleasure to humans, but with a robot this could only ever be a one-sided friendship. When you realise that it does not have real emotions for you, you become disappointed and feel pain. Friendship with robots is likely to bring pain rather than pleasure.

2.DEONTOLOGY, KANTIAN ETHICS AND ROBOTS

In all likelihood we will soon be living with unconscious robots which, although believed by some theories to be capable of consciousness, will not be able feel anything itself, as we have argued. We are now going to further discuss the kinds of ethical issues faced by the unconscious robot. But first it should be said that consequentialism might bring to mind behaviourism. If you remember, behaviourists focus only on observable

behaviour and, as has been argued, negate the conscious processes behind those behaviours. I have argued that this is a mistake. Consequentialists do a similar thing: they focus only on the consequences of the behaviour to decide if an action is morally good or bad, and ignore the mental processes (decision-making process, reasoning process) of the agents. Even though they are different theories in different areas (one is to account for attributions of mental states and the other is to decide whether the action is morally right), they have a shared feature: both ignore or negate the consciousness underlying intelligent behaviour. By contrast, Kantian ethics gives some reasons why we should not focus on only consequences to evaluate whether one agent acts morally well or badly, but also the conscious processes; and the intention or will behind the action.

According to Kantian ethics, no matter what the situation is or what the consequence will be, we ought to follow a set of ethical principles; for morality consists of constructing and following rules. This is actually the way a robot works, therefore some philosophers, such as those who support functionalism, might think that this supplies a convenient ethical theory for robots; but the important point in Kantian ethics is that real moral autonomy is based on the capacity to understand and the will underlying an action. This is a feature which the robots will never have because the behaviour and function of robots is restricted by their programming (Powers 2009). Therefore, Kantians tend to be among the most resistant to the idea that robots could be genuine moral agents. They argue that robots lack the necessary kind of genuine rational thinking capacity that is fundamental to any kind of genuine moral agency. Deontology asks that agents reason correctly in any situation which calls for moral judgment. In order to decide what is right, we have to use reason and a sense of consideration for other people. The reasoning process is a necessary component of the morality of the action, and it includes judgment, experience, and emotion. Kant requires good agents to behave for good reasons (Wallach

and Allen 2009: 70) and acting for good reasons is something that only conscious beings can do – not robots.

After giving that summary, let us start with Kant's categorical imperatives. The categorical imperative is an infallible guide as to what we should do; it does not matter what the situation is, it does not matter what you wish to achieve, and it does not matter what your desires are. It is a moral obligation that follows from pure reason. There are different formulations of the categorical imperative[23], but in this thesis, I will focus on only one of them: the formula of humanity. My aim is to discuss whether a robot can behave ethically towards humans as well as whether there is any reason to show ethical concern towards unconscious robots.

In Kantian ethics, a rational being ought to never be used by somebody else simply to fulfil another end. Instead, they should be thought of as ends in themselves. Moral agents are not merely objects which exist to be used by others; they are rational and autonomous and have the capacity to establish their own aims and work towards them. Kant says that a rational agent is someone, '…whose existence in itself has an absolute worth, something which as an end itself could be a ground of determinate laws…' (1785: 4:428). We may use each other as means to an end (e.g., we may use the abilities of a bus or a train driver) but we should not see each other as merely objects. Kant argues that this is because of our autonomy. Humans are self-governed; therefore, we can establish our own ends and we can make our own free decision on the basis of our rational wills, and it is this that ensures us our absolute moral worth. We should not

[23]Kant claims that these formulations are equivalent, not only in that they direct us to perform the same actions, but also in that they are different ways of expressing the same thing. Each formula follows from another. According to Kant, what is behind all the formulations is the principle that we ought to be guided by our rational understanding of duty. However, recently, there has been considerable controversy on the question of whether the formulations are equivalent or not. It has, for example, been claimed that 'the formula of humanity' and 'the formula of autonomy' or 'the Kingdom of Ends' are stronger formulations (implying well-defined set of duties) as compared to 'the formula of universal law' (Korsgaard 1998: xxv).

be manipulated or manipulate other autonomous agents for our own profit. In summary, we should, 'act in such a way that you always treat humanity, whether in your own person or in that of another, always as an end, and never as a mere means.' (Kant 1785: 4:429).

Kant implies that one should not treat others or oneself merely as means; but we should always treat others and oneself as ends. Ends are the things for the sake of which we act; means are the things which we do and the acts which we use to accomplish ends. Sometimes means also can be ends. Take today's machines, we are using them as tools because they have been created to help us achieve our aims; for instance, to calculate, to entertain, to clean, to carry, or simply to prevent loneliness. In all these ways we use robots merely as tools. This would violate the Kantian ethical law if the robots were moral agents – but in fact, they are not.

The robots are not aware of what is going on and they do not have any intention of respecting the categorical imperative. That is why, in Kantian ethics, they might be thought of as objects. However, there might be some people who treat robots as if they are ends because humans are sentient creatures who can project emotions onto robots. For example, imagine a robot that looks after an elderly person. That person might ask help from the robot to reach a high place in the kitchen, to clean the house, to prepare food for her/him, to help her/him get dressed and they might play games together. After a while that person might project some emotions onto the robot. S/he might think that the robot shows some behaviours which resemble humans' emotions and might also think that the robot might have some feelings for her/him. Humans tend to attribute mental properties to objects very easily and empathise with them (Muller 2020). Therefore, eventually, that person might see that robot as a friend instead of a mere means used in order to make her/his life easier. This shows that humans might behave towards the robots as moral agents. But what about the robots? Can they respect humans as moral agents?

One immediate answer to this is that robots cannot treat humans as moral agents because robots are themselves not moral agents who are aware of their actions, who are rational beings that understand and reason, who can take the responsibility for their actions and who can empathise with humans. Our moral decisions depend on our ability to empathise. If a robot cannot empathise, then it cannot understand treating humans as agents as moral agents. If a robot does not understand what it feels to be an agent, then it cannot act towards humans as agents. In Kantian ethics, moral agents should recognise the aim of her/his own action and be able to understand and evaluate the actions of other moral agents trying to achieve the same aim. Similarly, the robots would have to recognise the goals of their 'actions' and understand and evaluate the rationale behind human actions (Wallach and Allen 2009: 98). However, they do not have the requisite psychological knowledge (Wallach and Allen 2009: 96). Robots do not have the ability to feel empathy and understand feelings. Their algorithms do not enable them to feel emotions. Thus, they cannot be said to possess will, to reason, to understand humans feelings, to freely make decisions and to treat humans as moral agents. Because they are not themselves moral agents, they cannot treat other beings as moral agents.

There is an objection to this from ethical behaviourists, such as John Danaher, which is worth considering. They argue that if a robot is created so as to seem to exhibit morally relevant feelings, this should be thought enough in order for the robot to be described as a moral agent (Danaher 2020: 2025). According to Danaher, what is happening 'on the inside' does not matter for someone who is considering ethical issues. The important point from an ethical perspective is 'performative artifice' (Danaher 2020: 2025). As long as there seems to be 'roughly performative equivalency' between a robot and another entity then the robot has the same moral status (Danaher 2020: 2025). He argues as follows: '(1) If a robot is "roughly performatively equivalent" to another entity

that has significant moral status, then it is correct and suitable to claim that the robot has the same status. (2) Robots can be roughly performatively equivalent to other entities that have significant moral status. (3) Therefore, it is true and suitable to grant robots significant moral status.' (Danaher 2020: 2025). Danaher elaborates on what he means by 'rough performative equivalency'. He says that 'if a robot is behaving like another entity that is attributed moral status, then the robot ought to be granted the same moral status'. This means that if a robot is consistently behaving *as if* it is in pain and if the capacity to feel pain is accepted as signifying moral status, then the robot ought to be accepted as having moral status in the same way that other entities that are ascribed moral status. He explains this as 'performative equivalency'. He points out that in comparing any two entities we will never find identical performative equivalency. We will only ever find 'rough' performative equivalency. But if there are slight performative differences between two human beings this does not mean that they do not share significant moral status. Therefore, he thinks that in order for a robot to be granted moral status, it also does not have to behave exactly the same as another entity that has moral status. It will be enough if the robot only shows *most of the relevant performative* signs in similar situations (Danaher 2020: 2024-2026).

Thus, ethical behaviourists think that knowing what goes on inside the body does not matter since when we determine the morality of an action, we can only be guided by observation of the agent's observable behaviours, and there is nothing else to guide us. Ethical behaviourists, assessing the morality of a robot, might refuse the epistemic relevance of anything else beyond externally observable robotic behaviour. Therefore, for example, if a robot acts as if it feels empathy, then it can be accepted that the robot is able to empathise. But we have discussed and given good reasons why behaviourism fails in the second chapter; the same objection applies to ethical behaviourism: inner motives

and thoughts are important parts of the behaviour/action. Robots which do not have any inner motives cannot be accepted as moral agents who can empathise (Nyholm and Frank 2018: 223). The robot might appear as if it empathises with humans, but this does not imply that it has *genuine* feeling or *genuine* sense of empathy for humans. Even if the robot achieves exact 'performative equivalency' with a human being, it is still just a machine moving and of no more moral significance than an egg whisk. That is why Danaher's ethical behaviourism cannot be accepted.

After giving this short answer to Danaher's ethical behaviourism, let us return to our concern related to empathy and treating humans as subjects. Empathy is a subjective ability which is developed over time to judge what another person may be thinking, to understand things from their point of view and to share emotions (Kozima, Nakagawa and Yano 2004: 83). It is an ability to put yourself into someone else's shoes. Empathic thoughts rely on our capacity to imagine the event as if it was something you experienced yourself. When humans empathise, they intuitively apprehend the mental states of someone. In some sense, empathy gives us access to other minds. Robots do not have this access. Empathic behaviours require the underlying mechanism of perceiving and expressing emotions. When we observe what is happening to others, we not only activate the visual cortex, but also activate our empathy. Our emotions and sensations act in sympathy. But our empathy presupposes that behaviour is internally motivated.

When humans empathise, they do not simply try to register the state of another person; they actually try to experience and share it. We mentally 'simulate' (imaginatively identifying ourselves with someone) the emotion of someone else in ourselves so that we may understand what it feels like (Goldman 2006: 19). This is called 'simulation theory'. Simulation theory suggests one way to access the others' mental states (a legitimate form of mindreading). The whole idea of mental states is related to

empathy, on this theory, because empathy includes a process of re-enactment that attributes mental states to yourself and to others (Ravenscroft 1998: 178).

Jane Heal, a well-known simulation theorist, labels her strategy as 'replicative' instead of 'simulative' in her article, "Replication and Functionalism", so I will use the word 'replication' to refer to her theory[24] (Heal 2003: 11-27). According to Heal (2003), the assignments of mental states rely on our replication of another person's thoughts, and this is carried out by the same means as our understanding of ourselves. She calls her theory as 'co-cognition' (Heal 2003: 92). Co-cognition is a replication process and allows us to understand other minds from the 'inside' (Heal 2003). When we are thinking others' thoughts, we co-cognise by projecting ourselves into others' situations. Humans have the capacity to replicate the mental states of other people – so to speak, to step into their shoes. We can experience something of their mental states even if we do not share their exact mental states. The process of replication acts as a guide as to what mental state the other person (replicated person) is in. This process allows us, at some level, to access each other's minds. Heal (2003: 14) says that '…if I am capable of describing the initial conditions that I replicated, then I can cite them…' When I empathise with my sad friend, I re-enact one part of her mental life. For example, I know that my friend failed an exam, and she is crying. So, I imagine myself in her situation; I mentally simulate her emotion and get ideas about what she would feel, the situation she is in. I imagine myself having failed an exam and I experience sadness – however, my behavioural output does not have to be identical. I do not have to cry to prove that I understand her feeling, that I empathise

[24] I choose to use the term 'replication' to refer to simulation so that it does not cause any confusion about the simulation of the brain that we discussed in the third chapter. Remember in the third chapter, when I objected to functionalism, I used the sentence 'simulation is not duplication.' The point made there was that even if we simulate a brain, the robot cannot be conscious. Similarly, a simulated tornado is not real tornado. But in this section, simulation theory refers to mentally identifying yourself with someone as the basis of our understanding of mind.

with her, but there might be a desire to cry. In this way, I can assign the mental states that I have to my friend.

When we replicate someone's mental status, the aim is to reach a synonymous state between our mental life and theirs. When we replicate someone's mental status, we assign it to them by assessing their situation and projecting the mental state that we have reached – via 'pretend' inputs – onto them. This is what empathy entails. Robots do not have any mental states; therefore, they lack the capacity of attributing mental states to us. That is why whereas we may attribute mental states to robots and act towards them as moral agents (because we, who have mental states, can imagine ourselves as if we were doing the same thing as them), robots, for the converse reason, cannot attribute mental states to humans.

In summary, as explained by simulation theory, we are able to replicate other people's mental states, whereas robots cannot. They do not have the required knowledge of human psychology because they are not conscious themselves. In contrast to simulation theory, functionalism supposes that explanation of action or mental state via emotion, beliefs, desire, etc. is causal. Functionalists suggest that we see other humans as we see stars or clouds or geological formations; that we should focus on their external behaviour, and only then formulate a hypothesis about their inner states. Functionalists link internal states to specific external environments. But they ignore the fact that humans have extremely large numbers of different beliefs and desires (or other mental states) which cannot be always explained by causality (Heal 2003: 11-27). Our emotions play a big role in taking action. Our memories, traumas, hopes, fears, hates, love, happiness, sadness, and personality have a big effect on the manner in which we empathise. Our ability to empathise depends upon personal motivations, weaknesses, strengths, joys, history of success or failures, etc... But none of these can be programmed. Human

behaviour has a ‘dynamism’ and spontaneity that cannot be achieved by a robot (Ulgen 2017: 75-76). Humans can deal with unpredictability via the components of consciousness such as shared values, experiences, perceptions, and motivations, all of which robots lack.

Robots might ‘learn’ faster and ‘decide’ faster than humans can, based on all the data that is computed, and they might mimic some of our emotions. However, this empathy is merely mimicked (I will call it ‘echoed empathy’). It works by observing, learning, answering to and duplicating the signals that humans send. So, if we think in terms of Danaher’s ethical behaviourism, we should accept that the robot is empathising when it shows *roughly performative behaviour* (as if it is empathising – mimicking as if it has empathy). And, if we think in terms of functionalist view of the mind, there is again the possibility to think that this echoed empathy would be the same as real empathy, even if it functions in slightly different ways. For even if the internal processes (which are accepted by functionalists – remember this was the biggest difference between behaviourism and functionalism) that produce the behaviour might be different, they might reach the same outcome. In which case, echoed empathy and empathy would be impossible to distinguish, and then, according to functionalism, we can accept that the robot is able to empathise.

But robots cannot mentally identify themselves with humans since robots cannot experience for themselves what it is going on in the human mind. What functionalists and behaviourists are missing is that echoed empathy is not genuine empathy. Genuine empathy cannot be related to the quantity of data that is processed or the number of signals observed. It is about identifying ourselves with the other person on the basis of our own experience, feeling what they feel, being able to predict their next feeling without signals. The signs which are sent by humans are a very small part of the internal story

that they experience. We are more than the total of what other humans think we are by observing what we do and say. We are more than our observable behaviours. (Again, that is why ethical behaviourism cannot be accepted.) Also, none of the programmers can entirely know our feelings in such a way so as to be able to program a robot with the necessary data. Our conscious life has been shaped by experiences, by our biological stimulation and needs, and by collective intelligence and cultural memories over time. For instance, none of the robots would feel what it means to be hungry; none of the robots would fear homelessness; they cannot know what it feels like to be happy; they cannot know what it is like to be willing to help others. These are the things that make us feel empathic for others. Robots which cannot be programmed to empathise cannot treat humans as moral agents because they will not be able to understand why they should not treat humans as mere means.

On the other hand, robots might treat humans as means *under the morality of their programmers*.[25] The programmer can indirectly treat humans as subjects or objects because the morality of the robot will be dependent on the programmer's morality. The robots and their programmers (indirectly) can manipulate the humans who are interacting with robots (Muller 2020). For example, the robot 'wants' to get you to do something in

[25] Here it is important to clarify one issue. There is machine learning, so, programmers do not need to write the exact program in order for machines to complete a task. So, the machines can 'learn' and 'improve' themselves. Therefore, you may think that these machines may not be dependent upon the input of their respective programmers anymore. However, this cannot be right because even granted that the machines can 'learn', I think they are not learning consciously; therefore, they may remain always dependent upon someone who is a programmer, trainer, designer, algorithm engineer, etc. Imagine that a machine learning algorithm is trained to 'understand' which action should be taken in one particular situation. Perhaps, it will be given some different cases (or data) – there will be huge numbers of cases, I guess – and trained many times and when it guesses the *right* action, it will be rewarded. It is then expected that it will process those cases by itself. So, instead of writing a code to identify which action is morally good, the algorithm will 'learn' what action is the right one. But the point is that someone giving rewards is part of the original algorithm. This means that someone can always teach their own moral understanding to machines and because they do not understand, they will be just copying the actions that are bringing rewards. By contrast, humans learn consciously, and we can always criticise and refuse or accept: we can choose which action should be taken in a way that machines cannot.

accordance with some plan that it has (actually the plan of the programmer), but it does not believe that you will be willing to do it; therefore, it may 'decide' to overcome your will by brute force. For instance, imagine that the robot 'wants' to use the room in the library for itself (it is programmed to 'stay alone'), but I am in the room and reading a book. The robot 'wants' me to leave, but instead of giving me a credible reason for leaving, it may just pick me up and kick me out because it would not mind if I would be physically hurt or emotionally humiliated in front of people in the library. It cannot feel these emotions; cannot empathise with me; cannot know what it feels like when you are hurt; cannot understand what is in my mind. It just does what it is programmed to do. This would be one case where the robot would use us as a mere means to get what it 'wants'. What the programmer wants is that the robot has the room for itself and so the robot uses me as if I was a mere instrument in order to accomplish that end. It treats me as if I am merely a piece of furniture.

Another example might be as follows. It is famously the case in Kantian ethics that lying is never justified (Kant 1785: 4:422), for lying prevents and disrespects the ability to take free and rational decisions; and it does not allow other humans to choose rationally and freely. If you, for instance, deceive me, I cannot make an autonomous decision about how to behave since my decision would be based on the wrong data. If somebody lies to us, s/he would be treating us as merely means to achieve her/his aims with no interest in our own aims and interests. But imagine a robot which is programmed to 'lie'. Imagine, for example, a robot called 'The Liebot' which lies systematically in every area (Bendel, Schwegler and Richards 2017: 8). Imagine that you have got a Liebot, and you are in a room and it 'wants' you to leave the room because it is again programmed to stay in the room alone. It might suddenly say to you that 'there is a fire in your office!' (when, in fact, there is no fire). You would immediately leave the room to check your

office; and the robot, using you as a means to an end, would achieve its aim. It would not consider the stress it causes and would not allow you to make decisions freely. Again, this example tries to show the risks of robots which distort the truth, *in the interest of their programmers*. Without knowing the programmers, we would not know whether the robot is lying or telling the truth.

3. VIRTUE ETHICS, ARISTOTELIAN ETHICS AND ROBOTS

Thus, with Kantian ethics, we have placed the conscious process of the agent at the centre of the ethical discussions. In the third and last major ethical theory that will be considered in this chapter, virtue ethics, the character traits of the agent are at the centre and the main question is whether you understand and live a life of moral character.

Unlike Kantian ethics, virtue ethics does not give any specific rules to follow. This is because virtue ethics is not intended to tell us how to act in a specific situation or how to make decisions. It is about developing as a virtuous person; for as a virtuous person you will make virtuous choices and strive for the best outcomes. Morally good actions flow from the cultivation of good character, which consists in the realization of specific virtues (Hursthouse 1999). Being virtuous means both acting in accordance with the virtues and being in the correct mental states. Hence the virtuous agent can act from the true desires, for the true reasons, and at the appropriate time. So, according to virtue ethics, somebody's action is good if s/he is acting as a virtuous person would. In summary, in virtue ethics a good person is someone who improves as a human in the sense of having developed the set of virtues which enables them to behave in good ways for good reasons. For example, a kind person will not only behave in kind ways in circumstances in which it is easy and convenient to do so but is disposed to behave kindly

in all the circumstances in which kindness is fitting, including those when it is less easy and convenient to be kind.

Virtue ethics brings us back to Aristotle. Aristotle aims to identify the highest good for humans. He starts his *Nicomachean Ethics* (ca. 350 B.C.E. 1094a: 1-5) with the sentence 'all human activities aim at some good: some goods subordinate to others'. For Aristotle, a good life means to fulfil someone's aim through the species-specific way of life and to hence exhibit virtue. He maintains that a good life can be obtained through virtuous action according to someone's aim. He concludes that human beings distinguish themselves from other animals through their capabilities for reason, and hence our best and most appropriate aims involve the excellent use of the rational part of our soul through right actions. Thus, according to Aristotle (ibid. 1120a: 25), the virtuous human is someone who develops good habits and is disposed to do the correct thing for the right reasons.

Aristotle believes that everything is working towards a telos (end, purpose, or true final function of an object) (Ibid. 1139a; *Metaphysics*, Book II). Things are judged as good or bad depending on how well they fulfil their aim. He says that 'for all things that have a function or activity, the good and the well is thought to reside in the function.' (ibid. 1097b: 25-30). For example, the telos of an umbrella is to keep us from rain, so, the umbrella is good when it protects us from rain. Or the telos of a bud is to grow and become a rose. When it becomes a rose, it means that it has reached its aim. Similarly, humans have a telos. We – like all the other animals – need to grow up and be healthy and fertile. But unlike other animals, we are rational creatures. Therefore, part of the telos or true function of humans is to reason; acting in accordance with reason is the good for humans. 'Eudaimonia' (flourishing) is the final purpose of being alive. The term encompasses both satisfaction and fulfilment (ibid. 1095a: 10-15). It is the state which

we experience if we achieve a good life. In order to achieve eudaimonia, we must act in accordance with reason. Aristotle thinks that we ought to live a virtuous life and that nature built us with the desire to be virtuous (ibid. 1094a: 20).

Virtue is a character trait which makes us better humans just as the ability of protecting us from the rain makes the umbrella good. We should always work to be better humans and develop a virtuous character. Aristotle lists several moral virtues (for instance, courage, temperance, and friendliness, etc.) which all humans recognise and value as character traits. For instance, bravery and temperance are virtues; bravery gives humans the strength to tolerate difficulties and fear for the sake of something good whereas temperance gives humans the self-control to avoid too much of a good thing. Building on these virtues, he develops a concept which is known as 'phronesis' (practical wisdom). He says that 'all the virtues are forms of practical wisdom.' (ibid. 1144b: 17-18). All exercise of virtue requires the exercise of practical wisdom, which itself depends on the ability to reason, that is to say, act on the basis of experiences by making logical decisions. Aristotle says that 'the work of man is achieved only in accordance with practical wisdom as well as with moral virtue; for virtue makes the goal correct, and practical wisdom makes what leads to it correct' (ibid. 1144a: 7-8).

The contemporary virtue ethicist, Philippa Foot, says that wisdom is both an intellectual and a moral virtue, that is to say, it is both a state of mind and determination of character (2002: 5). So, wisdom is related to knowing something and willing something. For example, we might know the aim of life is to achieve certain aims and we might know how to achieve them too, but we may still lack the will to do anything about this. In order for someone to be ethically virtuous, they firstly have to develop habits, habits that encourage the development of practical wisdom. Practical wisdom comes from experience. If you have experience and deliberate by applying practical wisdom, you will

be able to act virtuously, and this leads to our eudaimonia. Aristotle points out that eudaimonia is closely related to virtue (ibid. 1098a: 30). Eudaimonia is not virtue, but it is virtuous activity (ibid. 1102a: 5). He thinks that virtuous people ought to have the capacity to make informed rational decisions about the best way to behave. Understanding of virtuous character traits is not enough; we should also know how to behave in the light of this understanding and how to implement these character traits when it is necessary.

The crux of practical wisdom, for Aristotle, is the doctrine of the mean (also known as the Golden Mean). The doctrine of the mean is an understanding of a virtue as situated between two vices (Aristotle, *The Nicomachean Ethics*, 1107a). Virtues occupy a middle ground between the vices of excess and deficiency, relative to each person. If we want to act virtuously, we have to know the mean; that is to say, we have to know how to avoid excess and deficiency and we have to know which action is appropriate for each specific case. When we realise this, we are in a better situation to behave in a virtuous way. For instance, the virtue of courage occupies the middle ground between being cowardly and being reckless (ibid. 1108b: 11-20). Aristotle acknowledges that it is not easy to find the Golden Mean: he writes that '… anybody can get angry – that is easy – but to do this to the right person, to the right extent, at the right time, with the right motive, and in the right way, that is not for everybody, nor is it easy.' (ibid. 1109a: 25-30). Let us take an example. Imagine that there are two Kings (A and B) from two different countries which have different levels of power and prestige. Suppose that King B is stronger. During the political meeting, King B insults King A. In this situation, King A has a number of options. Firstly, he could ignore the insult, because he is scared of the power of King B and losing Kingdom A. This would be a vice – cowardice. Secondly, King A could answer King B impudently and declare war on King B but without thinking

of the consequences for either Kingdoms, thus King A might bring suffering not only to Kingdom B but to his own Kingdom too. This would be again a vice and reckless. Thirdly, King A might find a middle way to solve the problem. He could answer with diplomacy and bear in mind that he is representing not only himself but also the Kingdom that he is responsible for. In other words, he could bravely choose the middle ground. As in this example, we must use our practical wisdom to know when to be brave, how to be brave, and whom to be brave towards.

Aristotle maintains we should also try to learn from other virtuous people (*The Nicomachean Ethics*, 1103a; 1103b). These people are moral exemplars who already own virtues. We are built with the capacity to notice them and the desire to emulate them. So, we should try to emulate them but also, we should practice being virtuous. This is something we need to work at and then finally it will be part of our nature. We cannot simply acquire character traits by means of a decision; practice is imperative. This is difficult at the beginning since we will be copying those who are better than us. But after a while, these behaviours will take root and become part of our character.

Foot attempted to modernise Aristotelian ethical theory and apply it to the contemporary world. She argues that we must intend to develop virtues (Foot 2002: 148-157). According to her, moral action is rational action. So, we can say that being a good person does not mean only following formal rules (1995: 14). We have to act rationally in order to act morally. She writes that 'as I see it, the rationality of, say, telling the truth, keeping promises, or helping a neighbour, is on a par with the rationality of self-preserving action, and of the careful and cognisant pursuit of other innocent ends; each being a part or aspect of practical rationality.' (1995: 5). Practical rationality is different from theoretical rationality. Whereas practical rationality is about what we should do, theoretical rationality is about what we should believe. Virtues are character traits which

are good for the person who has them. So, if you have virtues, then according to a broadly Aristotelian view, you will tend to do the right thing, at the right time, in the right way. Imagine two people: one is virtuous, and the other is non-virtuous. They might do the same thing; for instance, both may go shopping for house-bound elderly people. But the virtuous person will go shopping for them since s/he recognises the well-being of others as an important reason which specifies what s/he should do. By contrast, the non-virtuous person might do the same action, but s/he does not do it for the sake of doing good. S/he might, for example, feel guilty because the elderly people could not eat, so in order to get rid of that guilt, s/he might go shopping for them; or it might just be that s/he wants to be seen as a nice person.

After giving that summary of virtue ethics, we can now focus on our main concern, which is the question of whether robots can behave virtuously. One very quick answer to this is that robots cannot be virtuous because they do not have any mental states which enable them to choose the right action for right reasons at the appropriate time. In order for a robot to be brave, for instance, it would have to overcome fears, but robots do not feel fear; they do not feel anything at all. The person who chooses the middle ground possesses the capacity of reason and knows the difference between bad and good action. Doing what is right becomes a habit over time and humans acquire an affinity for this kind of good action. For example, we can become brave by doing brave actions over time with repetition (Aristotle, *The Nicomachean Ethics*, 1103b). Similarly, we can become architects by building houses again and again; we can become guitar players over time by repeatedly playing guitar; and we can become fair by acting fairly over time. After a while, these things become habits. For instance, we might know that brick must be put into one particular place, but we will only become good builders when we know how to place that brick properly; we may know the notes of a song on the guitar, but we will be

a good guitar player only when we play the notes properly, etc. We need practical skills as well as intellectual knowledge in order to build a wall or to play the guitar. The same thing applies to improving virtuous character traits. Intellectual teaching is not enough in order to develop virtuous character; we need practical learning and habitual action.

Imagine a bird that has just hatched. It will not initially know how to fly; likewise, a human being does not know how to walk when s/he is born. Flying is a process which involves a lot of trials and failures for the bird since it relies not only on instinct but also on practice. The fledgling often falls from the nest, but after a while it realizes that if it spreads its wings, the fall from the nest becomes softer. Once it learns to open its wings, flapping them will be the next step to be discovered, and eventually, that flapping becomes flight. However, it is not yet perfect flight; the fledgling has to practice taking off and landing, and how the wind might influence the flight. After some time, these processes all become natural; they all become habits. Now, imagine something which is like the bird, which also has wings: an aircraft driven by an autopilot. These wings help to produce lift just like the biological wings of the birds; they are functionally similar. However, when the bird is flying, it wilfully performs this action. By contrast, although the aircraft can fly, this is not because it has practised flying and it chooses to fly; instead, it is simply programmed to fly. The aircraft has not cultivated a habit. When we are cultivating a habit, we are *consciously* learning it. What it is needed for habituation is conscious learning – likewise, for virtue. The robots can 'learn' but it cannot be claimed that this is conscious learning; this is the difference between robot-learning and human-learning. What a robot does, at most, is to emulate virtue.

Let us think about how Google translate works and has improved over time. Google Translate is a statistics-based translation tool. It calculates possibilities of different translations of a phrase being satisfactory (Groves and Mundt 2014: 113). In

2016, Google announced that it had transitioned to a neural machine translation – a new machine translation system based on an artificial neural network and a deep learning system which can compare whole sentences at a time from a great range of linguistic sources. Deep learning is a kind of machine learning which is inspired by the structure of the human brain. With deep learning, machines can pick up information without human intervention. By using this system, Google translation has continuously improved its quality of translation (Groves and Mundt 2014: 120). It analyses millions of documents which have already been translated by humans. These translated texts come from books, organisations, universities, and webpages from all around the world. It can scan these texts by looking for statistically important patterns, namely, patterns of correlation between the translation and the original text which are unlikely to occur by chance. Once the computer catches a pattern, it can use this pattern to translate similar texts in the future. In that respect, it is possible to claim that the machines can 'learn'. When we repeat this process billions of times, we end up with billions of patterns, massive amount of data, and one very smart tool. The translations may not be perfect, but we can make them better by constantly supplying new translated texts. It might take a very long time to train such a tool, but there is still a sense in which it can 'learn'.

Similarly, Cog, the humanoid robot created by a team at MIT, can be claimed to 'learn' from experience (Dennett 1996: 15-16). The project was based on the idea that human-level intelligence can be gained from experience by interacting with humans. Just like Google translate, it requires a long period to 'learn' by interacting with humans. However, Cog was designed to 'learn' socially. Cog gained experience from the environment of the real world just like human infants. It had software for visual face recognition, which Dennett (1994) accepts as 'innate' endowment. It also had some other 'innate' features when it was initially equipped. Anything which is not fixed at the

beginning but gets itself designed into the control system of Cog via learning will be lifted into Cog-II and this will be a new 'innate' feature created by Cog itself. Thus, Dennett (1996), thinks that Cog can train itself. The team hoped to get Cog to build language. Cog has four ears and the ability to separate human speaking sounds. In addition, it had speech synthesis hardware and software. In order to have a natural conversation with humans, it needed to be well-equipped. It needed a long time, just like humans, but in short, Cog was able to design itself by learning from infancy and creating its own 'representation' (according to Dennett) of its world.

In both examples, it would be true to say that these machines can 'learn' from experience. But the difference between human-learning and robot-'learning' lies in consciousness. Humans can learn consciously from experience whereas Cog and other robots can only 'learn' (automatically). So, a virtuous human being and a robot can do the same action; both might go shopping for an elderly person; they both might learn to help people, but it would neither mean that the robot has virtues nor that it is using rationality to create a habit for itself. So, when we are creating habits, these habits are not unconscious; they are learned as practical skills; they are not unconsciously duplicated or copied from role models. When we are creating our own virtuous character traits we consciously learn, not only by observing what, for example, a brave person does but also at some point by independently practising our skills. When we consciously cultivate generosity, for example, after a while we arrive at some point where we feel genuinely good about giving. It is like learning a language. In order to learn Chinese, first I have to copy the teacher's examples of grammar, syntax and vocabulary. Then, at some point, I will acquire my own Chinese grammar, syntax, and vocabulary independently of the teacher. Eventually I will have my own practical skill of expressing myself in Chinese. Here, the point is that I did not naturally learn Chinese without any effort on my part. We

cannot learn Chinese in a day; we need to practice over time, and we need teachers. When we are improving this skill, we spend effort thinking about what the person who has this skill would do. What might the Chinese language speaker say in a particular situation? After a while, we will not need to think about what that person does, we will just be doing it (Annas 2008: 23). Similarly, the person who would like to learn how to be brave needs to ask her/himself when somebody needs help, what would a brave person do in this situation? In order to be brave, in all sorts of different circumstances, we have to train our feelings and find the appropriate middle ground. Someone who has mastered being brave will not need to consider on every occasion what a brave person would do. If someone is in danger they will, without thinking, do the right thing. If someone needed to be rescued, they would not waste time calculating.

When a virtuous person performs an action virtuously, s/he does not evaluate every time whether or not s/he has good reasons for doing that action since it is already an implicit motivator for them. Even though s/he does not evaluate every time, her/his actions are virtuous because s/he acts in accordance with practical wisdom. It should be emphasised that in order for us to be able to act virtuously, we have to learn this action consciously over time by practicing. By contrast, a robot relies entirely on calculation. Moreover, even if a robot can make the calculation and choose the right action, the output action cannot be accepted as virtuous action because the robots did not consciously learn it by exercising and having practical skills over time. It did not consciously master that skill. Virtue ethics emphasises that 'moral knowledge, unlike mathematical knowledge, cannot be gained only by attending lectures' (Hursthouse 1997: 118). It means that we need practical skills to develop a character; it means that we need consciousness to learn that knowledge; we cannot only memorise or copy the behaviour; it would not be a

virtuous action were we only to copy a certain behaviour, since we would lack the appropriate virtuous attitude.

In considering what a brave person does in a particular situation, we see that there is not only an objective difference between a brave person and a person who is not brave, but also that this difference is 'on the inside'. Thus, it makes sense to refer to *what it is like to be* a brave person (Annas 2008: 21-22). Remember the example of Nagel's '*what it is like to be a bat*?' (1974). It was given to discuss the subjective character of experience in the third chapter of this thesis, when functionalism was being criticised. Nagel was arguing that it does not matter how accurate our explanations are, we cannot understand first-person experiences via third-person views. Also, as humans, we can never be sure if we experience the same things in the same way (remember 'qualia' from chapter three). Similarly, we can know about the behaviour of a brave person, but we cannot know the qualitative character of her/his experiences. It does not matter how much we learn about a brave person; we will never understand her/his conscious experience without actually having it. Therefore, we have to learn from our own experiences so as to produce our own bravery. We have to practice over time to master certain character traits. A robot which cannot have its own experience cannot be thought as brave or cowardly because they lack a fundamental feature; they are not in a particular cognitive situation because the cognitive situation requires consciousness. In just the same way, we cannot have virtuous zombie agents even if they might functionally be the same as virtuous agents. Aristotle says that 'the agent must be in a certain condition when he does actions; in the first-place he must have knowledge, secondly, he must choose the acts, and choose them for their own sakes, and thirdly this action must proceed from a firm and unchangeable character.' (*Nicomachean Ethics*, 1105a: 30). But robots are not in a particular cognitive situation; they cannot choose what kind of robot they want to be; therefore, character

traits cannot be attributed to them. Character traits include preferences, mannerisms, and behaviour – but behaviour alone is not enough to prove that someone has virtues and is not just mimicking them.

At this point it would be worth considering Danaher's ethical behaviourism again. Remember that behaviourist approach to mind claims that if an entity behaves as if it is conscious, it means that it is conscious[26] – and if that entity is conscious, then we may claim that it can have virtues. Danaher argues that if robots exhibit good to us, then we ought to think that they are genuinely good (Danaher 2020). So, he thinks that robots can be our 'virtue friends' (2019: 9). This is perhaps not surprising, because he is an ethical behaviourist. He would support the idea that if a robot behaves as if it has virtues – if it is roughly performatively equivalent to another entity that is virtuous – then it is virtuous. He claims that we do not have any way to enter the heads of our friends to explore their true interests and values (Danaher 2019: 14), so therefore, only the observable behaviour of a friend is enough to evaluate whether we can have virtue friendship.

But Danaher misses one important point, which is the fact that when we value a friend, we do not *only* look at her/his actions, but also her/his underlying concern for us, which is part of what causes her/his to behave as s/he does. When we consider someone's character traits, we look at what is on the outside of a character (such as appearance, facial expressions, and observable behaviours) as providing our window onto the inside of a character (such as their actions, speaking, thoughts, expectations, understanding and emotions). A friend is someone who is keen on helping us out of her/his concern, while someone who is not our friend might be willing to help us with something since it is suitable for her/him for some other reasons (Pettit 2015: 11-43). A robot may act in a way

[26] Danaher's behaviourism might also be understood as an eliminativists view that denies consciousness, claiming there is only behaviour.

which is similar to how a good person would act, so the robot may seem to be good in the same way as a good human being. Nevertheless, in humans' situations, we think of the behaviour of the human as indicating underlying attitudes, values, principles. Even though a robot can mimic humans' virtuous actions and can function behaviourally in ways similar to humans, they still cannot perform virtuous actions or possess virtues; robots cannot be *genuinely* virtuous. The robots cannot behave for reasons at all, let alone for the right reasons (Purves, Jenkins and Strawser 2015: 852) – because acting for the right reasons requires consciousness which robots lack. They might seem to us to be behaving in a virtuous way when we observe only their external outputs, just as a human being might when they act in a virtuous way but for selfish reasons. But in the case of robots, there is no fact about their conscious intentions at all – whether they are acting for good or bad reasons – simply because they do not have conscious intentions. They are unconscious machines. In order for robots to actually be virtuous, they would have to have an internal dimension of virtue. External behaviour is not enough. Acting virtuously is related to the character of the virtuous agent (Hursthouse 1999: 136) and so robots, which cannot have virtues, cannot act virtuously.

CONCLUSION

In this chapter, we have discussed ethical issues related to robots in terms of three main ethical theories: consequentialism (mainly utilitarianism), deontology (mainly Kant), and virtue ethics (Aristotle and Philippa Foot). None of them provide any reason to think that we should show ethical concern to unconscious robots, nor is there any reason to think that unconscious robots can act ethically or unethically towards us.

We began with consequentialism. According to consequentialism, moral action requires us to calculate possible results in terms of measurable outcomes, for example

pleasure. In order to calculate pleasure, we have to deliberate, so we have to have consciousness – which robots lack. When humans try to find an action that will cause the most happiness and the least pain, they reason. Conversely, robots must be programmed with the quantifiable metrics in order for them to be able to maximise pleasure for the greatest number. But, contrary to Bentham's claim, pleasure and pain cannot be quantified. In fact, the grounds for refusing the robots to be programmed with consequentialist or utilitarian ethics are the same grounds on which we should hesitate to recommend humans to embrace consequentialism. As we have seen, there are three reasons why we cannot quantify pleasure and pain. First, there are differences in human experiences, that is to say, we cannot compare our experiences to someone else's. Secondly, there are huge number of variables which can influence the consequences, such as age, gender, and education. Thirdly, predicting consequences is very difficult – we can never be sure whether our action will cause the greatest amount of happiness for the greatest number. We also discussed the idea that robots might be thought of as tools that allow their programmers to perform morally good actions indirectly, finding they were unlikely to be effective tools for this purpose.

We then moved onto the ethical issues relating to robots in the light of deontology. In Kantian ethics, the conscious process of moral agents is at the centre of the ethical discussion. Robots which lack consciousness cannot, according to Kantians, be thought of as moral agents; they are just tools for their users, programmers, or both. They cannot decide, understand, reason, will or have intentions, and so they cannot be ethically good because these abilities require consciousness. Therefore, there is no reason to show ethical concern for robots. In order to argue this, I have focused on Kant's categorical imperative – a rational being should not be used as merely a tool; s/he should be thought of as an end in themselves. But this is a viewpoint that is not open to robots. Robots

cannot act towards humans as subjects because they do not feel empathy; therefore, they cannot identify themselves with humans. Empathy is necessary for humans to understand each other's minds. Robots are dependent on the morality of their programmers, so robots themselves are limited to treating humans as merely objects.

Finally, virtue ethics was discussed. According to virtue ethicists, morally good actions are generated by the development of good character, including the nurturing and development of particular virtues. Virtue ethics asks us to develop our character traits to become a virtuous person by using our practical wisdom to achieve the Golden Mean, as we learn from other virtuous people using our rationality – all of which requires practice. With practice, we become more virtuous and closer to achieving eudaimonia. However, although robots can 'learn' from experiences, they are unable to practice doing the right thing at the right time and for the right reason. That is only possible with consciousness.

In short, morality, according to all the main traditions of Western Moral Philosophy, requires consciousness. So we have very good reason to think that robots should not be thought of as moral beings. They are incapable of the necessary feelings of pain and pleasure that in turn make empathy and morality possible. They always remain dependent upon the morality of their programmers. In conclusion, without consciousness robots are incapable of moral actions. But what would the consequences be of treating robots as if they were moral agents? This is the question that I shall try to answer in the next chapter.

CHAPTER 5: THE SOCIAL IMPACT OF PEOPLE TREATING ROBOTS AS IF THEY ARE MORAL AGENTS

INTRODUCTION

In the previous chapter, we have argued that there is no reason to show ethical concern for unconscious robots. Robots cannot act morally or immorally towards us, for in order to act morally or immorally the action would have been performed *consciously* by moral agents – and robots lack consciousness.

A moral agent is somebody that can intentionally harm or help someone. S/he is a rational individual who has the ability to make moral judgments, who has the ability to empathise with others, who can discern right from wrong by reasoning and who can be held responsible for their actions (Parthemore and Whitby 2013: 105). Kantian ethics says that 'an action cannot be morally good unless the agent in fact reasoned in certain fairly complex ways.' (Allen, Varner and Zinser 2000: 253), but my main point is that in order to talk about good action, the action should be reasoned, and reasoning requires consciousness. Therefore, only conscious beings can be moral agents. Thus, a moral agent must have certain mental states and events such as desires, believes, wills and intentions which again require consciousness. By contrast, unconscious robots can be considered as essentially tools. They are only objects, just like toys, water, a tornado, or a stone, that might cause some actions, or even problems, either by being used or not used. A stone can make some movements, it can cause some issues, but not in the way that a conscious human being causes consequences with her/his actions. For example, if I throw a stone and if that stone causes harm to someone, it is not implied that the stone itself deliberately harmed that person. The stone might be part of the chain of events, but it is not the agent that caused the harm. The moral agent is still me – I caused the harm.

When we perform an action, we deliberate, predict, evaluate the situation, and decide upon the action consciously. Robots which do not have any mental states cannot meet these conditions; they do not have any intentions to perform the actions. Like food processors, they are 'performing' whatever they are told to do. A food processor might complete an action without any problem, and it might fulfil the functions for which it was designed. But this does not mean that the food processor deserves to be respected as a moral agent. The important point here is that the tasks it carries out are unintended. Robots can be far more sophisticated than food processors, but the same principle applies to them: their actions are unintended.

Robots are just tools; therefore, if they break, we can fix them; if we cannot fix them, we just replace them. When we finish our work with them, we can just turn them off. But if robots are just tools as I claim, then, because of their potential for misuse, we need an agent who can be held responsible for their actions. Who is that to be? The moral agent must have mental states, that is to say, their actions should be motivated by their mental states. Their actions will have an influence on the external world and have the potential to harm or help other agents. If we can explain an agent's actions by referring to their conscious intentional states, then we can say that we should be considered as a moral agent. This might include, for example, the designers or programmers of robots.

Moral responsibility refers to acts for which you can be rewarded or punished. But it is impossible to reward or punish a robot. Kant writes that 'for if the moral law commands that we ought to be better human beings now, it inescapably follows that we must be capable of being better humans beings'. (1793: 6:50). That is to say, ought implies can: we are only morally required to do things which are possible for us. We cannot be morally responsible for cases which are out of our control. Therefore, robots cannot be responsible for their acts because the situation is outside of their control – they

are either following the algorithm that was inserted by programmers or a human being is controlling them directly. If someone shoots someone we do not say that the gun did it, so when robots are involved in bad actions, we cannot say that the robots are guilty.

We humans have a tendency to automatically attribute minds and agency to anything even remotely humanlike. Thus, robots that outwardly behave like humans encourage us to treat them as if they were moral agents. Especially, if a robot makes eye contact and follows our movements, we automatically respond to it as if it were a social being (Turkle 2007: 511). We are ready to be deceived into considering that there is some conscious activity. In particular, social robots are designed to provoke this response. When people are able to talk to robots, many perceive them as moral agents rather than objects. When people see a robot which is being abused or damaged, they might react empathically or protectively, for we are sentient creatures; we are used to feeling and understanding the feelings of those around us.

People will have a variety of attitudes towards robots as they become more prevalent in our societies. Some will like them and act towards them as if they are moral agents, as if they have personality; others will act towards them as they are only tools. For example, imagine a care robot that looks after an elderly person. That old person might develop feelings for the robot and one day if the robot does not function, s/he may feel very sorry. People sometimes may hesitate to 'hurt' robots and see it as a violation when they are hurt. Boston Dynamics created a robot called Spot which is an agile and mobile robot dog that can run, climb, and has an ability to stay balanced. There was a video which went viral in 2015 showing people kicking Spot (Parke 2015). The aim of the video was to show that Spot can regain its balance. After that, people started to say it was wrong to kick a robot dog. Some people see it as violation, but others disagree. They claim that it would be morally wrong to kick a robot which acts and looks like a real dog.

For instance, Kate Darling (2012) says that if we treat robots in inhumane ways, we become inhumane people. Robert Sparrow (2016) thinks that it is wrong to kick a robot dog because it might reveal the kicker to be cruel or vicious in their dispositions and in the future, they can inflict these kinds of vicious behaviour on living creatures[27]. Noel Sharkey says that 'the only way it is unethical is if the robot could feel pain' (in Parke 2015). Mark Coeckelbergh agrees with Sharkey, saying that kicking a robot is not itself unethical (in terms of the harm done), but nonetheless what is unethical is the behaviour itself (in Parke 2015).

In fact, robots have already entered society before we have a clear idea of how we ought to think about them. As they become further integrated into society, and their numbers grow, their actions will increasingly affect our lives, and perhaps most importantly people seem very likely (almost inevitably) to treat them as if they are moral agents. In this chapter, we will discuss the social impact of people treating robots as moral agents – even though they are actually not. I will be highlighting the risks which are related to driverless cars, robots in the workplace, sex robots and finally, killer robots.

1.DRIVERLESS CARS

A driverless car (self-driving car) is part car, part robot and part computer, and can drive without human intervention. It relies on sensors, algorithms, machine learning systems, powerful processors, and sophisticated software. For instance, it can brake when there is something in front of it, find the shortest route to a destination, change lanes when there is traffic, play music and 'decide' what action it should take when there is an ethical dilemma. It can drive just like a human being in traffic. That is to say, it can act *as if* it is

[27] This is similar to Kant on cruelty to animals. Kant did not think animals could feel pain, but he thought that we should avoid treating them as objects because it would encourage callousness, which would then spill over into our treatment of humans.

a human driver – it can act as if it is conscious. It appears to be thinking and deciding where it should go, what it should do, etc… Therefore, it is easy to be convinced that it is a moral agent, even though it is actually not. So, it is possible that in the future many people will mistakenly treat them as if they are moral agents.

In philosophy, the word 'agency' is a technical term which refers to either a capacity or the exercise of that capacity (Schlosser 2015). The capacity is to deliberatively act, i.e., make decisions, reason about how to behave, interact with other agents, make plans for how to behave, judge former actions and learn from them, take responsibility for our behaviours, etc… Agency is a multidimensional concept which refers to the capacities and activities that are related to displaying behaviours, deciding, and taking responsibility for what and how we behave. Many people, like functionalists, will assume that if a robot can simulate the human decision-making process, make life-death decisions, make split-second decisions, selecting their 'own aims' (although these are nothing more than mimicking those capacities), then a robot would be an agent too. Also, they will assume that if the robot can do all of these, then it is also making ethically acceptable decisions (Nyholm 2020: 55). But they ignore the fact that agency must have consciousness and rationality. Humans who are conscious and who have a moral conscience are agents because they can act morally, make decisions, interact with others while taking responsibility for what they do. Animals can perform actions, but many non-human animals do not take responsibility for what they do; for this reason, they are still agents, but a different type of agent, unlike human beings. For instance, some animals like chimps are cognitively more advanced than dogs and cats; however, they are still not as complex as human agents. Their agency includes the use of some tools and the ability to learn from each other (Shumaker, Walkup and Beck 2011), but it does not lead them to have a legal system, governments, or courts. So, there might be different types of

agencies, but driverless cars cannot possess any one of them. Therefore, we should not treat them as if they are moral agents. Treating them as if they were moral agents would cause ethical problems. Some who think that robots are moral agents might even love them and attach significance to them as though they are conscious beings and will avoid blaming them for any bad consequences. Instead, they will try to blame others. Others might dislike them and even hate them, trying to blame them whenever possible, for anything with negative consequences. Thus, we come to one of the biggest problems that can occur in society – the responsibility gap. Who is to be held responsible when things go wrong?

Since driverless cars became a reality in 2015, they have caused a number of accidents. In most cases, these accidents were between conventional cars and driverless vehicles and there was little harm. In 2016, when there was an accident between a bus and an automated vehicle designed by Google, Google, for the first time, accepted responsibility for what happened. In the accident report, Google declared that 'we clearly bear some responsibility' (Gibbs 2015). Also in 2016, a Tesla model S in autopilot mode, but with a driver onboard, crashed into a lorry which was not recognized by Tesla's sensor, and it caused the first fatal crash. However, unlike Google, Tesla did not take the responsibility because the company said that the vehicle was *under control of the human driver* and that the human driver was responsible for the incident[28]. Nevertheless, in both cases, the companies promised to update their software and make driverless cars safer. In 2018, a driverless Uber car killed a woman in the street in Arizona (Levin and Wong 2018). It was the first time that a driverless car hit a pedestrian. Uber suspended their testing, but later, it announced that driverless cars have the potential to become safer than

[28] This information comes from the 'Tesla Team' Website.

normal cars. Of course, humans make mistakes too. However, when human drivers cause accidents or damage, we are supposed to take responsibility and pay the consequences.

It is believed that driverless cars will decrease congestion on the roads as well as the number of car accidents and will be safer than cars with the human drivers (Urmson 2015). They are created to reach their destination by following instructions in ways which are safe, and which save fuel and time (Loon and Martens 2015); their aim is to create an optimal way of driving. Human drivers sometimes ignore safety, fuel saving, speed limits, and might break other rules of the road. Moreover, in traffic, there are highly unexpected complex situations which require sudden decisions. Therefore, humans do not always react optimally. On the other hand, driverless cars can quickly calculate different options extremely quickly and 'take actions' (Lin 2015). But these actions are not based on decisions that are made consciously, and these actions could in principle lead to bad consequences for passengers, pedestrians, and other drivers who are actually the moral agents. For this reason, it will be very important to think carefully about the 'morality of driverless cars' (which is actually the morality of the programmers), especially since they 'decide' life and death decisions in traffic. So, we should work out who is responsible when something goes wrong. The hardware manufacturer, the advertiser, the government, the owner/user, the driverless car itself, or the programmer? In normal cars, accidents are usually caused by human error; therefore, we can work out who should be blamed relatively easily. But the situation is different for driverless cars because, although they 'decide' for us, they themselves lack consciousness. When we cannot immediately point to the person responsible, a 'responsibility gap' occurs. Danaher (2016: 300) calls it a 'retribution gap' – we do not know whom to punish for damages caused by robots or machines.

Traditionally, when misfortune happens relating to machines, the manufacturers or operators are legally and morally responsible (Matthias 2004: 175-183). But driverless cars drive without operator intervention. Therefore, we cannot blame the users or owners of the driverless cars because they did not build the car; they did not program the car; they did not tell the car what to do; therefore, they cannot be responsible for the damage. That is to say, the human users of the driverless car cannot be responsible because the car is not under her/his control, and they cannot know what the car is doing. The human user is not a driver anymore, s/he is just a passenger and passengers are not responsible for what the drivers are doing. The passengers may know some of the statistics and dangers related to using the car; however, they still lack the knowledge of a driver. Furthermore, s/he may not know how to drive a car – eventually it may be common for people to have lost their driving skills – therefore, we cannot expect that they are able to interfere. This may be one of the important consequences of driverless cars partially replacing human drivers. History may follow the example of when humans used to ride horses. After cars came along, people started to use fewer horses and they started to lose their ability to ride horses. Today, most of those who ride do so as a hobby. The same might happen in the case of driving cars.

Let us now leave this speculation and continue with the problem of attributing responsibility. Driverless cars cannot exercise responsibility since they cannot discuss the reasons for their behaviours; that is to say, they cannot take responsibility for their actions in the way that humans can (Purves, Jenkins and Strawser 2015). In the *Nicomachean Ethics*, Aristotle (Book III) argues that in order to be responsible for what one is doing, some conditions have to be met. First of all, the agent has to be in control of what s/he is doing. If we do not have control, we cannot be responsible. For instance, I cannot control a tornado; therefore, I cannot be responsible for it. On the other hand, when I drive my

car, I have control over it; therefore, I am responsible for my driving. Secondly, the agent has to know or be aware of what s/he is doing. For example, when I am driving my car, and I see a child that is in potential danger, I alter the direction of my car. So, I am responsible for whatever I choose to do. (Coeckelbergh 2016: 750). Knowing what I am doing does not just cover knowing how to drive a car, but also knowing the background to the case, which includes knowing the environment and situation where the action takes place. Supplementing Aristotle's argument, I would say that the agent has to have obligations. An obligation can be described as something such that if we do not do it, we will be blamed. For example, human agents have an obligation of care when they drive a car, and they have to be sure that nobody is hurt because of their driving. Likewise, doctors should take care of their patients and be sure that patients do not suffer because of their treatments. On the other hand, driverless cars cannot be blamed or consequently punished. Therefore, they cannot have any obligations either (Smids 2020). But whenever someone is hurt, people naturally look for someone to punish.

We might blame manufacturers for the bad consequences, saying that the manufacturers should be responsible for what they sell, import, and distribute. They are, in general, responsible for designing, manufacturing, and performing the essential assessments in accordance with the relevant regulations. They are responsible for assuring customers that their products are safe. For example, in the UK, a rule says that 'if you are made aware of any safety risks or consumer incidents related to a product you have sold, you have a legal duty to report these to the manufacturer, supplier or your local trading standards service. If you do not do this, you could become liable in the event of harm to a person or damage to property.'[29]. As a result, we are able to blame the manufacturers when there is something wrong related to the products that they produce.

[29] This quote comes from 'UK Government' Website.

But it seems that they are not ready to take responsibility yet, as seen in the examples involving Tesla and Google. Moreover, even though we can hold manufacturers responsible, it is important to note that the car's programme was written before it was manufactured.

Furthermore, it might be argued that in some circumstances advertisers and governments should be held responsible. We are living in a society where advertisements are almost unavoidable. Big companies invest a huge amount of money in this area, and it seems reasonable to suggest that advertisers should be held responsible for the information that they provide. However, advertisers tend to play up to our suggestible natures. Subsequently, if people feel cheated, they may blame the advertisers. Advertisers sometimes may not know the details of the products that they advertise, so they might mislead the public unintentionally. In the case of driverless cars, they would have to not only know about the technology of driverless cars but also the morality of the programmers, so that they can be sure of what they are promoting. But it is a fact that most advertisers would not know everything related to what they advertise, such as the future possible dangers of driverless cars; and they might always deny responsibility by saying that customers, who have personal freedom of choice, are able to investigate the products themselves.

People may also blame governments – for allowing the sale of driverless cars. After all, governments are able to ban particular products. However, that tends to be a last resort. Governments would be more likely to create new rules in preparation for the arrival of driverless cars and for the accidents that they may cause; for example, the German government has already created moral guidelines for programming driverless cars whereby the preservation of human life takes priority. But many governments might simply implement rules after the problems created by the new technology has already

become apparent. However, governments themselves certainly do not always assume responsibility for things going wrong with new technologies (Caughill 2017). Often neither advertisers nor governments will assume this responsibility.

The last option remaining lies with the software code writers/programmers. Perhaps the programmers should assume responsibility regarding the ethical consequences of their creations. In other words, they should be morally accountable for what they create and bring into the world. However, programmers cannot always predict the outcome of their programs. Also, there are robots that can continue to learn (unconsciously); therefore, humans cannot predict the future behaviour of the robots because they cannot predict all the events, with all their consequences, that may occur in the future. However, we still have to face the question of what kind of ethical framework programmers should put into driverless cars?

When driverless cars encounter dilemmas, they have to make decisions which are calculated on the basis of the codes that they are programmed to follow. As discussed in the previous chapter, robots will always remain under the sway of the morality of their programmers. There are three major ethical theories that we have discussed, the main ones in Western philosophy – consequentialism, deontology, and virtue ethics. Imagine that there will be many programmers who are working on a driverless car. Some of them may be, individually, morally well-equipped to design a driverless car, but, unavoidably, there will still be conflict between moralities. So, for example, if the programmer (or their boss) has utilitarian ideas, then s/he would program the driverless cars to choose to kill the least amount of people in an accident and save as many lives as possible. But when the programmer designs the algorithm, s/he will be dealing with the huge number of possible outcomes, so the consequences may not actually be the ones envisaged – which

on a consequentialist morality means that the programmer's design may have unintentionally immoral outcomes.

Foot introduced the 'trolley problem' to discuss the problem of someone choosing to kill an innocent person in order to avoid killing many more people (1967: 2). There are different versions of the problem. In its basic version, there is a runaway trolley on the railway tracks with five people on the track who are tied up and cannot move. The trolley is heading towards them. Imagine that you are next to a lever. If you pull the lever, the trolley will change to a different set of tracks but there is someone else on the other side of the track. You have two options: you either do not do anything and allow the trolley to kill the five people on the track. or you pull the lever and send the trolley to the other side where there is just a single person. According to utilitarians, one dead would be better than five because they focus on the outcomes. Therefore, in this kind of trolley problem, utilitarians would switch the tracks. For Kantians, ethics follow the moral principles that have to be followed no matter what the consequences are. So for Kantians, assuming that deliberately killing someone is in principle worse than watching someone dying, then they would not switch the track. This seems right because for a Kantian, killing another rational being is always immoral; therefore, the decision to kill one person to save five is unacceptable, for it amounts to utilizing one person's life as a means to an end, and this would be violating her/his autonomy as an individual. In general, for utilitarians, morality is ultimately about doing whatever has the best consequences for the greatest number whereas for Kantians, there are certain moral rules that should never be broken. It is the co-existence of these different moral understandings on the part of the programmers that has the potential to cause chaos on the streets.

Now, let us apply the trolley problem to driverless cars in order to further discuss the issues related to the programmers. Imagine that there are three passengers in a

driverless car and the brake is suddenly broken. The car has to 'decide' what to do and where to hit (its programming will allow it to choose one option). Let us say that there are three options: first, it can go on without doing anything different and hit two pedestrians; second, the car might swerve and kill just a single pedestrian; third, it might drive into a barrier and kill all of three passengers inside the car. If the car was designed according to Bentham's classical utilitarianism, the car would be programmed to 'examine' the circumstances and 'choose' to swerve and kill one pedestrian since that is the least net loss of life. Furthermore, for example, a hedonist programmer can design a driverless car which always prioritises the safety of its owner. So, this 'hedonist driverless car' will always save you. But if there are two hedonist driverless cars that encounter each other in the incident, who will be saved becomes problematic and this will lead to another conflict in the streets. Also, this kind of driverless car that prioritises their owners' safety is bound to be a lot more expensive. However, if you are able to pay more money to get a hedonistically programmed car, then, if you could afford it, it would make more sense to pay extra for an armoured car!

I have discussed three moral theories in the previous chapter. Even though their main purpose, roughly, is to distinguish right actions from wrong ones, over two thousand years moral philosophers have not been able to come to any firm conclusion about which morality is correct. Philosophers struggle to find a moral theory which is accepted by everybody. For example, Kantians have been criticized by many philosophers (such as Hegel, Schopenhauer, and Nietzsche) (O'Neill 2000: 75-79). Among their critics are virtue ethicists such as Elizabeth Anscombe (Singer 1983: 44-45) and utilitarians such as Mill (Brooks 2012: 75). Anscombe criticises Kantian ethics because of its obsession with law and obligation. She finds that a theory that depends on a universal moral law is overly strict, and it is unsuitable for a modern society (Singer 1983: 44-45).

There are also disagreements among the proponents of every major moral theory. For example, consequentialism mainly focuses on the consequences of an action in order to decide if the action was the right action (Sinnott-Armstrong 2003); but this does not mean that in the same situation all consequentialists would agree, because the subtle differences between their accounts mean that their priorities differ. For instance, egoism, one of the forms of consequentialism, states that a person ought to behave in such a way as to create the greatest good for her/himself (Burgess-Jackson 2013: 532) whereas classical utilitarians argue that we ought to act to create the greatest good for the greatest number – even if you have to sacrifice your own happiness.

Furthermore, there are moral relativists who claim that there is no definite set of moral rules that are applicable in all circumstances (Gowans 2004). According to these moral relativists, there is no objective right and wrong. Namely, what is right for one person is not necessarily right for another, or what is right in some cases is not necessarily right in another. What we all have is various cultures and societies which each have their own practices and their own norms. Our morality has been shaped over centuries through the combination of our genes and culture. So, the culture that we live in can influence our thought about what is right and wrong, but there is no universal right answer. For instance, being honest and respectful are the features that very often appear, but there might still be some differences across cultures. Each society and each community might have different moralities, that is to say, moral rules and values might be different for each society at a particular time.

Humans have created moralities according to their needs over time, and these moralities have changed. For instance, between the 15th and 18th centuries, there were witch-hunts in Western Europe. People who were thought to be witches were tortured and killed, and this was considered morally acceptable, even valuable. Another example

might be slavery. In the past, slavery was acceptable, but now people protest against it and see it as evil. So, morality changes over time. Similarly, morality differs according to location. For example, eating beef is not approved of in Hindu societies (Nesbitt 2004: 25-27).

If this line of thought is accepted, then it would not be possible to find one objectively true morality. And even if it is not, the fact remains that thousands of years of philosophical discussion has not brought about any agreement about what the objectively true moral system amounts to. But with millions of driverless cars in the world, decisions will inevitably be made in which someone has to be sacrificed. This kind of decision is incredibly difficult. The programmers will program the cars but who do they tell them to save? The decisions have to be made ahead of time. The problem is that there might be too many possible accident scenarios. There was an experiment about this carried out by MIT[30]. They asked people across the world to choose who driverless cars should kill in one possible dilemma. It seems that people most often agree about three things – save human life, save the greatest number, and prioritise children. However, they have found that morality changes according to country and there are different clusters of belief. For example, some societies decide according to age: Eastern countries do not have strong inclination to save young people, instead they choose elderly people to save. Some societies choose whom to save according to their gender. For instance, France and its subclusters demonstrate a strong preference for women over men. Social status is also a factor. For instance, when a country has a comparatively high economic inequality, there is a higher chance of choosing to save the life of someone of high status rather than the life of a homeless person.

[30] This information comes from the following website: https://www.moralmachine.net/

This study has tried to demonstrate how morality can change across societies. Everybody thinks differently, every culture has different thoughts about morality. As a result, people cannot find a common ground. It might therefore be claimed that there is no chance of the programmers being able to find common ground either. So, if programmers cannot decide which morality is correct, then driverless cars will all be programmed *differently*. This will cause two major problems. First, there will be chaos on the streets because of the cars with different 'moralities'. Secondly, some of the cars will seem more attractive to purchasers and thus, attract higher prices. As mentioned, hedonistic driverless cars might be the most popular because they will always prioritise the safety of the car owner. But, as a result, they will be much more expensive and only rich people will be able to afford them. This will cause more inequalities in society.

Now, let us make the scenario more complicated. Patrick Lin (2015) suggests a thought experiment. Imagine a future when you are a passenger in your driverless car in the middle of highway and there are other cars around your car. Suddenly, a big object falls off the lorry in front of you and your car will not be able to stop on time to escape the crash. So, it will have to make a decision about whether it should go straight on and hit the object or whether it should swerve right and collide with a strong car which has high passenger safety, or otherwise turn left and hit someone on a motorcycle. Should it prioritize your safety by crashing into the motorcycle, or reduce danger to others by not turning left or right even if this means crashing into the massive object and sacrificing your life? Remember, what a driverless car should do will already be programmed. If we, genuine moral agents, were driving that car, then whatever we decided, it would be considered as just a reaction; it would be an instinctive panicked action made without any bad feelings. According to Surden and Williams (2016: 121), 'theory of mind cognitive mechanisms allow us to extrapolate from our internal mental states in order to estimate

what others are thinking or like to do. These cognitive systems allow us to make instantaneous, unconscious judgments about the likely actions of people around us.' Furthermore, our practical ethical decision-making in an emergency like a road traffic accident, or a trolley problem, is not actually consistent and human decision-making can be influenced by different things such as what kind of mood you are in, or whether you had a cup of coffee in the morning. So, our ethical practical decision-making principles cannot be expressed as the kind of hard and fast rules that a computer requires (Dreyfus 1972: xxix). By contrast, driverless cars will proceed in a more predictable way. A programmer will have instructed the car on how to act in these conditions. For the Kantian, the instructions encoded in a utilitarian's program would look like premeditated murder. Driverless cars are forecast to decrease traffic accidents and casualties by eliminating human mistakes from the driving equation, but accidents will still occur, and when they happen, their results will have been determined months and years in advance by programmers or policy makers who will have had to make some difficult decisions.

Lin (2015) also offers a variation of his thought experiment with two motorcyclists. So, let us suppose that we have the same scenario as previously, but now instead of the strong safe car on one side, there is a motorcyclist who wears a helmet to your left and there is another one on your right without a helmet. Which one should your driverless car hit? If you say the motorcyclist with the helmet since s/he is more likely to survive, then are you not punishing the responsible motorist? If, rather, you save the motorcyclist without the helmet since s/he is acting irresponsibly, then, arguably, the driverless car is now apportioning street justice. Here, the ethical concerns are getting more complicated. In both scenarios, we will be systematically supporting or discriminating against hitting a certain kind of object, and the target vehicles' owners will

be suffering the negative results of this algorithm through no mistakes of their own (Lin 2015).

Someone might respond to the previous claims as follows: 'You have conceded that people in these situations simply act instinctively – how you react may depend on whether you had a cup of coffee that morning, as you said. So surely the fact that the programmer has at least programmed some moral considerations into the car is good – it is moral progress, because at least some moral consideration is being put into these situations and it is not being left to chance.'[31] The answer is that in that case, major moral deliberation will be required before we build these cars. Otherwise, contradictory moralities will be deciding our futures – rich people might be able to pay more to have cars that always prioritise the driver's safety, for instance. So maybe the cars could indeed introduce more morality into situations where a person would only act on instinct, but if the moralities conflict – and people can buy the moralities that suit them better – then this would just make matters worse. Therefore, I assume that the programmers and manufacturers must take moral responsibility, and as the previous examples about Google and Tesla suggest, they are not remotely ready to do that.

Sometimes, in real life, we might encounter scenarios similar to the trolley problem. When human drivers encounter them, they have to make split-second decisions. But driverless cars have to be pre-programmed as mentioned. So, before the scenarios occur, the programmers have to find answers to them and take the responsibility for doing so. We always seek someone to punish. We generally prefer a negative outcome to be the result of a human making a bad decision rather than for a negative outcome of the same value to be the result of a computer just making a statistically bad call. We prefer to have

[31] This objection was made by my Ph.D. supervisor.

somebody to blame even if that means we will be blaming humans more often. For instance, we often go to the counter instead of self-check in desks at the airport in case there is some mistake, because if there is a mistake then we will have someone to discuss it with and possibly blame. We can only blame and punish conscious beings. If the computer does something wrong which it would do almost instantaneously, then there is nothing that we can do. We cannot blame the computer itself; its programmer would be the one that we could blame – although their distance from the event makes it very unlikely they will receive justice.

In conclusion, we should remember that the decision of a driverless car to swerve or hit someone was already made long beforehand by the programmers. However, it will not be easy to blame the programmers or software companies who designed the cars and nor will it be easy to make them accept responsibility. But if we cannot, then some people and software companies will have the mysterious power to make life and death decisions for us. So, the regulatory system of the governments will need to be made ready before driverless cars become prevalent. This will also apply to the insurance companies. They will have to decide what they do in particular scenarios – even on the assumption that, in general, there will be less risk to insure. How the question of responsibility is faced will have a big impact on society and reflect people's different attitudes towards robots. The issues will not just revolve around driverless cars but around the whole transportation system. Eventually, the entire system will have to change.

2.ROBOTS IN THE WORKPLACE

Imagine a future in which robots take the place of babysitters, where robots will take care of elderly people, where diagnoses will be made by robots, where your food and drinks will be prepared and served by them, and where your hair will be cut by them.

These are just a few examples; there may be many more instances of unconscious robots working and interacting with humans in the future, and some people will treat them as if they are conscious.

Imagine a robot that is working as a babysitter in your house and suppose that your baby is spending most of the time with that robot. It is rocking the baby; soothing the baby when s/he cries; feeding him/her; later, teaching the baby how to walk and how to speak.[32] When the baby grows up, the robot can help her/him with homework; robots can 'know' many languages, thus they can teach the kids; they can paint or draw together; whenever the kids feels bored the robot can entertain them by playing games or telling a story. Thus, they become closer and closer to each other. Imagine that both parents are working; they would not have too much time to spend with their children; they might feel tired after work; so, the robots might step in, and eventually they become the closest one in the family to the child.[33] Robots can be smart, agile, flexible; they never feel tired; they might cook whatever the children want; they can mimic being sentient; as said before, they can act *as if* they are conscious beings. Moreover, they can 'manipulate' the emotion of the children; so, in time, the attitude of the children raised by the robots will likely be different than their parents' attitude towards robots. Parents might still see the robot as an unconscious tool while the children may think that it is no different from a human being and they might well get attached to robots.

This kind of close relationship might cause at least two main problems, personal and societal. Individually, the children might start to replace their parents with the robots

[32] There was a robot called 'Aristotle' which was programmed to serve children. It was able to read bed-time stories, play games with children and help with their homework. It was even able to order baby products via the internet. All these functions were liked by many parents and advertised as 'all-in-one nursery necessity'. However, after many complaints (such as privacy concerns), the company decided to remove the product (Hern 2017).

[33] This might remind the reader of the novel by Isaac Asimov, *I, Robot*. The first story 'Robbie' is about a little girl called Gloria and her nanny robot, Robbie (1950: 5-29).

and see the robots as real parents. You might think that that is not really a problem: if the child is raised by a robot s/he can learn different languages, not feel alone while the parents are away, always feel secure in their friendship with the robot, etc. But robots can damage the development of children and their abilities to relate to other people. As said in the previous chapter, robots cannot have empathy because the ability to empathise requires consciousness. When the children are raised by human nannies they experience genuine empathy, feelings, and relations. On the other hand, robots just mimic feeling empathy or emotions. So, we cannot give children to robots which pretend to empathise, and then expect that children will understand what real empathy is (Turkle 2015: 3-18). Empathy is important for humans because it is a feature that deeply affects our emotional and cognitive intelligence. We cannot give children to robots that only mimic having emotions and relationships and then expect our children to understand what a genuine relationship is. If robot childcare becomes a mass phenomenon, through widespread technological unemployment of babysitters, nursery school teachers, etc., then human emotional development may be very seriously affected.

The idea of 'technological unemployment' was introduced in 1930s by J. M. Keynes who said that 'we are being afflicted with a new disease of which some readers may not yet have heard the name, but of which they will hear a great deal in the years to come – namely, *technological unemployment*.' (Keynes 1930: 359-360). There have been three industrial revolutions which frightened people because of the prospect of mass unemployment: steam engines, electricity and internal combustion engines, and computers (Kapeliushnikov 2019: 90). The first revolution (between 1760 and 1830) was the transition to new manufacturing processes. Whereas average income and productivity increased significantly, the number of farmers started to decrease because of the mechanization; farmers had to move to the industrial towns to find new jobs in the

industry and those who found new jobs had to reskill and work under poor and dangerous conditions. In the second industrial revolution (the technology revolution) (from the late 19th century to the early 20th century), there was growth in pre-existing industries. For instance, thanks to the first industrial revolution, we had telegraphy – this product had a big role in the second industrial revolution when there was a huge expansion of rail and telegraph lines. These movements led to mass globalisation and further technological developments. This was the first era of mass-production, which meant specialisation and standardisation. Productivity continued to increase, and rich people continued to get richer. For instance, a better form of the Bessemer process produced a larger amount of steel at a cheaper cost. People lost their jobs again, but some retrained, to become electricians, for example.

The third industrial revolution (the digital revolution) (1960s - today) refers to developments in technology from mechanical devices to digital technology. There is mass production with small costs. Again, people lost their jobs because of the new technologies, but some retrained, and became, for example, software or hardware engineers. It is predicted that the fourth industrial revolution will be related to technological advances, namely, robotization, digitalisation and the creation of AI (Kapeliushnikov 2019: 89) and that this time, the amount of unemployment might have an even greater effect. The biggest impact of the fourth industrial revolution may be an army of unemployed. Gene editing, new forms of machine intelligence, robotics, 3D printing, nanotechnology and other technologies are already changing the way we live, work, and relate to one another.

Technological unemployment will happen when human workers are replaced by technological alternatives, such as machines, robots, and computer programs. If robots can perform more jobs at a cheaper cost than human employees and if they were to be

accepted by humans in more and more places as employees, then there will be technological unemployment. Robots do not need any breaks or holidays; they do not need to eat; they do not get sick; they completely 'focus' on the tasks given them, so they can finish their tasks faster than humans; they can make big calculations quickly; they carry heavy things easily. For economic and practical reasons, governments, advertisers, industries will promote them, and they might be supported by most people, companies, and workplaces. But after a while, when human employees realise that robots occupy most of the working areas, they will not accept their existence in the workplaces so easily – as we can see from history.

Let us contextualise the future transition we are imagining by focusing on the transition from feudalism to capitalism. In the feudal system, there was no problem with the creation of work since social mobility was low. Children followed in their fathers' footsteps. For example, the children of a farmer would know that they would be farmers; the eldest son of an aristocratic family would inherit the family estate, etc. The traditional conception of work and life was disrupted when feudalism gave way to capitalism, caused by new machines and increased social mobility. Political authorities tried to limit the use of machines in order to prevent unemployment, but they did not always intervene. Therefore, the employees tried to challenge the machines (Campa 2018: 21-22). Karl Marx says that '…in the seventeenth century nearly all Europe experienced workers' revolts against the ribbon-loom, a machine for weaving ribbons and lace trimmings…' (1867: 554-555 cited in Campa 2018: 22).

In 1770s, Ned Ludd also known as King Ludd, a weaver, demolished a loom because these machines started to lose employees their jobs. Ludd became the leader and founder of the movement called 'The Luddites' (Pistono 2012: 49-65). Later, in 1810s, one group of people, most of them textile artisans, started to organise attacks on the looms

and machines and destroy them. The protests arose because machines were 'stealing' their jobs. In that moment, automation was only water and steam-powered, so it would not entirely replace human work. But some thought that automation would increase to the point that the manufacturers would themselves be put at risk. In the 20th century, one manufacturer who was concerned about the social effects of automation was Henry Ford. He decided to pay double to his employees so that they could afford the cars that they produced. He realised that you need people to have money to buy the goods that you have produced, otherwise the connection between production and consumption will be disrupted (Pistono 2012: 53). The concern is that if robots replace human employees faster than they are able to find new jobs, society will have problems. Humans who lose their jobs will be upset; and, together with others whose jobs are at high risk, they may start to attack the robots.

Until the 1970s, mechanical innovations increased productivity, productivity increased mechanisation, and employees produced more; thus, they become more valuable and earned higher salaries. From then to now, productivity still continues to increase, but compensation remains stable for most sectors of society; this is an indication of the changing nature of our technology (Ford 2015). Whereas machines were previously only tools which made employees more valuable in many situations, now they are increasingly taking humans jobs and making their work less valuable. In the future, advances in technology may further disrupt the nature of the work, making people less necessary in the workplace as most work will be done by machines. For instance, robots will be doing our chores; our cars will be driven by them; our goods will be manufactured by them. While robots in the past were only tools used by humans, in this new era, they might replace employees (Ford 2015: 21).

Technological developments reduce the demand for labour and decelerate the process of matching employees with work, that is to say, they change not only the level of labour but also its structure. Therefore, developments in technology might cause mass unemployment. Some jobs will be out-dated; others will need more requirements such as education, skills, and new trades. But after a while, higher education may not be enough anymore.

Machines have been used as complementary to agricultural and manufacturing labour. With improvements in machine technology, including AI, they have started to be used to supplement cognitive and emotional labour too. One of the jobs that will be at risk is the job of care assistant. The populations of developed countries are aging quickly. By 2034, The United Kingdom is projected to have over sixteen million retirees, which is about 23.5 percent of the population. It was reported that by 2025 the UK will need one million more care workers (Ford 2015: 164). In Japan, by 2025, a third of the population will be over sixty-five years old and they already have around 700,000 fewer care workers for elder people than they need (Ford 2015: 161). So, this gives a big opportunity to develop robots in this area. Robots can assist elderly people in mobility; they can monitor and communicate with them; they can help them with the house tasks; and they can be accepted by humans as care assistants. For example, in South Korea, someone has already developed robots that can remind the elderly when it is time to take their medicine (Campa 2018: 126).

Although the common belief is that education and skills will guarantee a safe and prosperous future for employers and employees, as just said, it seems that in the future, having more education and abilities may not automatically offer a protection against work automation (Campa 2018: 55; Krugman 2013). For example, in medicine, AI is likely to help with diagnosis and treatment and sometimes they may prevent fatal mistakes. Some

cancer patients die every year because of wrong treatment or accidents. For example, in 2001, Wayne Jowett, a leukaemia patient in remission, was receiving his chemotherapy. There were two drugs that he should be injected with, and these two drugs should have come in two separate bags, one to be injected into the spine and the other to be injected only into a vein. But they were sent in the same bag and handed to two doctors who were not trained to administer the treatment. So, both drugs were injected into his spine and eventually, after one month, he died. He was one of the twelve thousand patients that die every year in the UK due to medical errors (Ford 2015: 155). Another job in medicine at risk might be that of radiologists. Radiologists study hard and they need special training to spot the abnormalities in x-rays in order to make accurate diagnoses. Once machines show that they can accurately 'diagnose' illnesses and 'offer' effective treatment, they perhaps will not be needed; at least they will not need to directly meet every patient (Ford 2015: 157).

This kind of narrative about technological unemployment has been criticised by neoclassical economists and labelled 'the Luddite fallacy' (Campa 2018: 78). They claim that new technology does not efface jobs; it just alters the structure of the jobs in the economy. They think that those who lose their jobs will eventually be hired by other companies or sectors, that is to say, technological developments cause only *temporary* unemployment because later new jobs will be created, just as happened during other industrial revolutions. Alex Tabarrok (2003), an economist, argues that, 'if Luddites were right, we would already have lost our jobs since productivity has been increasing for two centuries. There is no connection between productivity growth and job loss.' Automation increases productivity and eventually, more wealth will be generated. But the need for labour will not decrease because when the economy grows, our standard of living will grow, too. He points out that over the last two centuries, we use machines and we have

not yet been replaced by them. Conversely, new job opportunities have been created. Thanks to machines, we have become more productive and creative. Occasionally, there might be some unemployment, but it is just a matter of time until we return to normal.

One of the new jobs that AI might create is engineers who will develop the AI systems. When engineers create new robots, they will need more and more engineers so that they can program more robots. Everything will be related to programming; therefore, we will need data scientists. They will be making sense of all the data which the inter-connected world is producing. We will need more machine learning instructors who can produce enough job-ready individuals for an AI-based world, in particular at universities and institutes. We have argued in the fourth chapter that robots cannot act morally; however, there will be many people who believe that they do. So, in order to ensure that robots act in a manner which is in sync with human values and morals, they might create a need for ethics compliance managers.

The Luddite Fallacy claims that humans will be able to retrain and learn new skills at a rate which cannot be matched by developments in technology. But this cannot be true anymore because developments in technology are exponential (Kurzweil 2005: 14). Those who think that automation will not cause unemployment often claim that the same predictions recur but are never fulfilled; the level of unemployment might increase and decrease in time; however, it never happened that technological developments have caused a problem that we cannot reverse. In my opinion, and that of many others, this time looks different.

First of all, tech capacity is accelerating faster than ever before, as indicated by Moore's law[34]. Imagine that you are driving a car, and, in an hour, you are doubling your speed. Let us say that you start with 40, then double it to 80, to 160, so on. If you kept repeating this, you would be something of out of science fiction. Eventually, you would be as fast as a spaceship which travels at millions of miles per hour, and you would be able to reach another planet in a few minutes (Ford 2015: xii-xiii). Secondly, today's AI and information technology is increasingly taking on cognitive capability. This means that machinery is not only about muscle power anymore; there are machines that *seem* to think, make a decision, solve a problem, and to learn. Now we have the algorithms that can reveal a huge amount of data and thanks to this, machines can 'predict', and in a practical sense they can 'learn' (Zuboff 2019: 8-12). This is really extraordinary broad-based strong technology, in particular, in the area of deep learning. As such, it will be very influential on the job market and industries. There are many jobs that were previously done by people after an extensive period of study that are now done by machines. Thirdly, it is projected that all sectors will be affected by AI (Ford 2015: 184). Think of the mechanization of agriculture (Reimer 1984: 438): because of it, many people lost their jobs on farms although the productivity per worker increased enormously. Later, those who lost their jobs moved to factories; later, they had to move to the sector service. Here, the important point is that agricultural technology was very specific – it influenced only one sector of the economy. AI is much different; it will affect every single sector of the economy. Currently, machines are doing approximately 30 percent of all tasks; by

[34] Moore's law is about semiconductors, but it has implications for computing power; it refers to 'Moore's perception that the number of transistors on a microchip doubles every two years. Therefore, we can expect the speed and capability of the computers to rise every two years whilst the cost of the computers is halved'. (This quote comes from https://www.intel.co.uk/content/www/uk/en/history/museum-gordon-moore-law.html).

2025, it is estimated that there will be a balance of 50-50 between humans and machines (Kelly 2020).

Moreover, it is not easy to believe that unskilled workers and skilled blue- and white-collar employees will train in order to be physicists, computer scientists, molecular biologists, etc. (Rifkin 1995: 36). A 45-year-old taxi driver will obviously not be able to train to become a virtual-world designer, and with technology becoming more advanced, the skills needed by humans servicing it will become more advanced. Yet humans will not take less time to be retrained for new skills if the speed of technological change is increasing. Robots will create the job of taking care of robots, but that requires knowing computer programming, which is a highly specialised capability (Campa 2018: 79). Also, even though new jobs might be created, there might be other difficulties apart from the retraining, for example, moving to the cities or areas in which new opportunities are created may not always be possible. The least capable and the least educated will not be able to move to other sectors. Yuval Noah Harari (2017) thinks that by 2050, AI will lead to the creation of a new class, and he calls this the "useless class" – People who are not just unemployed, but unemployable (Harari 2017). Imagine a person whose task is washing dishes. That is the only skill that s/he has got. So, when the robots come to wash the dishes, that person will no longer be needed in this workplace because the robots will be doing the washing without feeling tired, wasting time by talking to other colleagues or checking their phones or needing any health insurance. That person may find it impossible to find another job.

It is possible that new jobs will be created, but we will need to create new jobs in which human beings can perform better than automated algorithms. Otherwise, robots will continuously take our jobs. At the beginning, humans might feel sympathetic to those robots that they think are conscious, and that seem to feel the same way as us; however,

when they see that they are taking humans' jobs, humans will start to change their attitude towards them. Here, government or corporate attitudes might be different than employees' because the automation will be good for some social classes. The owners of capital will become richer and richer, just as happened during earlier industrial revolutions, while others will be displaced because of technological unemployment. Thus, there will be increasing inequality in income and social status. Some people will earn insufficient money to support a family (Ayres 1998: 96). They may have to accept lower salaries in a world with more robots and this might make income inequality even worse (Ford 2015: xvi). There is a particular threat to middle-class jobs such as accountancy, law, etc. – and this provides the greatest threat to the social order because, arguably, the middle-classes have been the main driver of social change throughout history.

Even if the robots do not take our jobs entirely, it seems that they will apparently change day-to-day tasks in the workplace, and it looks like this will be bad news, especially for lower-skilled workers who might not be able to retrain for a new job. In order to solve this problem, one way is to redistribute wealth from the owners of capital to the unemployed. Technologists and futurists often suggest that a 'universal basic income' would solve this problem (Ford 2015: 256-260). The proposal entails that everybody receives a standard, unconditional payment at regular intervals. Thus, it is believed that a support will be provided to the people who remain unemployed while they are looking for a new job or trying to develop new skills.

As a result of technological unemployment, ethical and social problems are likely to occur. The first group consists of problems relating to income distribution, as just mentioned; capital owners will receive more benefit and income from robots than other classes. But the second, which we now need to discuss, is that people will need to find

something else to do with the time that they are not spending in work anymore (Danaher 2017). A job is usually considered as a meaningful feature of life; therefore, robots pose a threat to the meaning of human life. Without work, people will be required to seek other sources of meaning in life. Meaning in life may consist in feeling useful, having responsibilities, and socialising with other people. When jobs are taken over by robots, humans may feel that their lives are meaningless.

Work is often thought of as providing meaning to life. For instance, when we enquire, 'Hello! What is your name? What do you do?', people usually respond by saying that 'Hi, my name is …, I am a lawyer, I am a teacher, or I am a shop assistant, etc.' Most of the time, we do not ask people their jobs like 'what is your job?', we only ask 'what do you do?' and they take this to mean 'what do you do for a living?' This shows that to a great extent we identify ourselves with our jobs. What we do is who we are and what we do is work (Pistono 2012: 73). Work is an important part of human life; it does not matter what work we are doing; as Bryan Magee said, 'work gives meaning to your life however unimportant the work.' (in Cowley and Hardy 2021).

So, there are many reasons to worry about the fact that robots will take over most of jobs that are done by human employees today and cause technological unemployment. On the other hand, some people may think differently; they may not think that meaningful life is related to work. They may say that employees are working longer hours under worse conditions with more stress, less skills, less security, less power, less benefits, and less salary, all because of developments in information technology. During the technology era, humans have started to be deskilled, and some might think this is a good thing, since it leads us away from work that has been rendered less valuable and enjoyment by technological development.

Danaher thinks that there might be reasons to embrace robots in the workplace (2017: 42). When robots take over human labours, humans will be freer to follow their own conception of the good life; they will not have to spend time on boring tasks; they will not be degraded by needless stress. Danaher speculates about encouraging technological developments and integrating with them. He seems to support the anti-work critics and he justifies his position with two main arguments (2017: 47). He calls his first argument 'work is bad': technological unemployment should be embraced because it will take something which is bad away from us. He calls his second argument 'opportunity cost': we should embrace technological unemployment because non-work is better.

First let us discuss anti-work theories. Here are some examples of what philosophers have said about work. Bertrand Russell (1935: 3) says that when it is believed that work is virtuous, 'a great deal of harm' is done. He continues that 'work is of two kinds: first, altering the position of matter at or near the earth's surface relative to other such matter; second, telling other people to do so. The first kind is unpleasant and ill paid; the second is pleasant and highly paid.' Bob Black has one of the most extreme claims related to anti-work theory. He says that 'no one should ever work. Work is the source of nearly all the misery in the world. Almost any evil you would care to name comes from working or from living in a world designed for work. In order to stop suffering, we have to stop working.' (Black 1986: 17). Peter Fleming writes that 'once upon a time, in some faraway corner of that universe which is dispersed into countless solar systems, there was a planet upon which clever animals invented "work". Slowly, work lost its association with survival and self-preservation and became a painful and meaningless ritual acted out for its own sake. Taking on a hue of endlessness and inescapability, the curious invention consumed almost every part of the clever beast's

lives. Its constant presence kept the order; held certain divisions in place; lavished the few at the top with untold riches.' (Fleming 2015: 1).

According to Danaher, not many people are genuinely happy with their work. Some work causes distress; some is humiliating, with bullying and sexual harassment; some has physical risks, etc. (2017: 48). He continues that there are different forms of works; cleaning toilets is one of them and it is not the best quality job that people do. If we think that different forms of works give different status to people, then we end up accepting technological displacement. For, thanks to technological replacement, people might be displaced into more creative and meaningful forms of life. Today, in our society, in terms of economic and political issues, work is seen as a necessity if humans would like to access basic needs and luxuries. Therefore, work seems something intrinsically compulsory. When we describe 'work', it often seems that we are wage-slaves. But actually, there is no obligation to work other than economic and practical necessity (Danaher 2017: 48).

Julia Maskivker says that work is bad since it undermines freedom which is the core value of human well-being (2011: 31). When we are working, we have limited ability to select how to spend our time and govern our own lives. Sometimes working hours are really long, they are not always the regular 9-to-5, so people have very limited time for themselves. Danaher thinks that when new technology dominates every workplace, we will be better able to control our own lives (2017: 50). We may take on work for particular ends but, if so, we are not acting as genuinely autonomous agents while we do this.

Having given the 'work is bad' argument, we can move onto the second of Danaher's arguments – 'opportunity cost'. This argument claims that even though working may not be bad, non-work is better. If we are not doing paid work, we will have

much more time to spend on our hobbies or leisure activities (Danaher 2017: 51-52). According to Danaher, the absence of work facilitates meaningful and pleasurable human life (2017: 52). Work detracts us from the things that we really like and that provide subjective satisfaction. Some people might be working at what they like, however, not everybody is so lucky. Danaher maintains that work is not good for most people, most of the time (2019b: 55). Some jobs affect people badly because there might be humiliation, low salaries, stress, physical risks, etc. Therefore, we should do whatever we can to accelerate the obsolescence of humans in the workplace. These kinds of features usually vary according to the associated social class of work. If we consider the bad sides of some specific jobs, rather than the good sides of others, then he thinks we might think more favourably about embracing technological displacement in general.

After giving Danaher's ideas related to work, now let us discuss why they are mistaken. When we think of work, income is the first thing to come to the mind, since income allows us to pay for a lifestyle. There is a standard structure society – exchange of work for income, so we humans work for income which helps to decide our quality of life. We may also think that work has to do with social contribution, meaning, community, and achievement. When automation accelerates, we will lose both of these aspects; that is to say, humans will be deprived of income and other social goods concerning personal and societal well-being. When we work, we gain excellence in improving our skills; we contribute to society; we feel that we are part of a community, and we earn social status (Gheaus and Herzog 2016: 74-75). When we work, we try to be good at what we are doing; we try to change the world for the better; we want to be noticed by others (Gheaus and Herzog 2016: 78-79). It makes us feel engaged in human life, as we work in collaborations. Moreover, work gives us purpose in life. Since childhood, we have been taught that things cannot be taken for granted, we have to take

responsibility for what we want and to earn it. Namely, if we want something, we have to work to earn it (Pistono 2012: 73). This idea gives us ambition in the life, and with ambition we can better focus on our aims, then try to reach them.

In our society, work not only supplies financial support, allowing us to gain access to essential and non-essential products, but it is also our main means of feeling useful and belonging. It has a big role in individual identification, socialising, and networking. Individually, work helps us to find our social and personal identity, improve our physical and mental well-being, increase our confidence and self-respect, and feel worthy because we contribute to society. Furthermore, work is an important element of society: it promotes community harmony and social and economic improvement, by organising social life at a macro level in terms of communal goals and progress. Removing work from human life means taking away our sources of meaning and well-being. If there is no work, this causes depression, laziness, and weariness. If there is no work pressure, we might live a life of listless and unsatisfied boredom. Human well-being can be expected to decrease when there is an absence of work, that is to say, when the prevalence of automation increases. Therefore, when technological unemployment comes true, humans will have a lot to worry about.

Furthermore, we should not forget that human beings have always worked. The first work was hunting and gathering. Everybody contributed and basically tried to survive. Later, farming created different sorts of work. Also, work has been seen as a virtue in that it contributes to society. We work to fulfil our needs and wants, but also to be a part of, and contribute to, our society. We are aware that if we do not work, our future may not be bright and eventually there might be poverty, homelessness, starvation, so work motivates us. But apart from this material necessity, working may make us happier and healthier. For example, during Covid-19 pandemic lockdowns, many people

on furlough took up voluntary work; for when we receive respect from others for the work that we have done, we feel accomplishment and success.

Over time, how we work, where we work, and what we do when we are at work have changed continually, owing to cultural, economic, technological changes; however, one thing has remained the same – humans have always worked. Therefore, the decline of work would be perhaps the most radical change in our history. There is no way of knowing how such a change would affect us. This kind of wide and deep-rooted social change cannot be untangled to predict purely positive consequences, as opposed to its potential to bring chaos to society. Karl Popper made broadly similar arguments against Marxism. He prescribed, as an alternative, 'piecemeal social engineering' (1957: 64). The piecemeal social engineer focuses on solving problems from the individual's vantage point, with the modest, but arguably more effective, ambition of altering one institution at a time. The process is reliant upon trial and error. Thus, 'we make progress if, and only if, we are prepared to *learn from our mistakes*' (Popper 1957: 87). So, we should try to make reforms in one thing at a time. But the decline of work would be wholesale reform to all aspects of human life, all at once – a massive experiment that might plunge us into depression and/or chaos, since human history is so strongly affected by work.

I agree with Danaher that some people are not happy when they are working. But this is because they are not doing the jobs that they like – they did not willingly choose their jobs. One big reason why some people not able to choose the jobs that they like is, again, automation. When the degree of automation increases in the workplace, people will unfortunately continue to do jobs that they do not like. However, sometimes this is still better than doing nothing. People may not be happy, but they may still have a purpose and feel useful, especially when promoted or awarded.

Another point that we can make is that being able to do whatever you like with our day (because there is no more work) might suit imaginative intellectuals. Intellectuals, including artists, have different interests and foresight; thus, they can create new projects by researching deeply and thinking critically. They can first create ideas, then process them and try to turn them into a new reality. They have creative agency. But most people are not equipped to imagine new projects and a different way of life; they need guidance and set activities, which is what work provides. Work helps most people to spend their days productively, to find meaning and purpose in life; and it is in work that values are learned and created. So, if people who are not intellectual do not work, they would need to find something to guide them or something to make them active in the life. Otherwise, they would simply waste time and become prone to laziness and depression.

Imagine that one day you are the last employee remaining in workplace; everything is automated. Your robot pet dog wakes you up every morning and shows you some dog behaviour. Later, just before the job, imagine that you go to the doctor for medical checks and there is no real doctor, only robots. They check your health condition and tell you that you have an eye infection. A machine prepares your medication, and you collect it from the machine because there is no longer a pharmacist. But imagine that you were not able to collect it on time and the machine was already switched off; and there is nobody to talk to or complain to. Let us suppose that, before going to the job every morning, you used to go and get a cup of coffee and have a little chat with the barista. The coffee is now made and served by a robot. Now imagine that you are taking your self-driving car to your workplace, but on the way, you will perhaps see many homeless and miserable people that cannot find a job anymore because of automation. Then, finally you are at work. There is nobody there, I mean no human employees, only

robots – unconscious machines able to perform their tasks without *talking*. They are indeed talking, however, but it is not a real conversation; it is only *as if* they are having conversation; as if they wonder about you and show concern. When you have your lunch, it will be prepared by a 3D printing machine, and you will eat it alone because your robot colleagues never feel hungry and never need a break. When you work, after a while, you might cease to feel ambition, since robots can do the tasks faster and better than you. You might start to feel that you are becoming useless. When you are promoted, there is nobody in the job that you can share your happiness with. The story ends when you are fired. So, now imagine that nobody is working but only the machines. Danaher imagines that you would have all the more time to spend on your own interests. Let us say that you fancy going to the cinema but cannot afford it. People who own the technology will be richer, whereas everyone else will be trying to meet basic needs with the help of the state (perhaps with universal basic income). In this new world, people would not cultivate interests, because it would be hard enough to get through each day. If work is taken from all of us, life will become meaningless. This would be one of the biggest impacts that robots might have on society.

3.SEX ROBOTS

Sex robots once appeared only in science fiction, but now they have become a reality. They are usually designed to strongly resemble real humans; they might 'track' the user's eyes, 'answer' to the user's facial expressions, 'guess' if a subsequent action is desired or 'start an action' the user may enjoy; moreover, they can 'perform' emotions. So humans may learn to ignore the fact that they are only tools and treat them as if they are humans.

In fact, anthropomorphism is very common among humans. For instance, when we were children, we used to become attached to our toys; we gave them names; pretended they were eating or crying; punished them if they were 'doing' something wrong. Or we might even get angry at a table that we knocked against. Or when our computer is very slow, again, we sometimes get angry at it. Today, for most of us, our cell phones are very important. When we forget it, we usually return home to retrieve it; if we do not, we feel something missing. It has been found that in the USA, for example, people check their phone every 5.5 minutes (Wheelwright 2022). So, even though these objects do not look like human beings, people often imbue them with humans traits. Imagine then the situation with a robot that would totally resemble and act like a human being. They already resemble conscious human beings, in appearance, and in their actions. These robots are not only designed to take human forms, but also to mimic human behaviour. The way that they present their verbal messages, facial expressions, and gestures can momentarily deceive humans. Some of these robots can also 'learn' from humans and continuously develop themselves. Therefore, when they interact with humans, humans will intuitively treat them as if they are humans. The more robots *seem* to understand and analyse our actions, the more they will seem to meet our assumptions. This may well increase our attachment to them; we might value them emotionally. If we see that robots are abused, we might take action. We might like robots or even fall in love with them.

There was a robot, Pepper, that worked for a bank. One day an old man attacked it because he was mad at the clerk and instead of arguing with the clerk, he chose to assault Pepper (Weber 2015). After this incident, some people felt sorry for Pepper. This example shows us that humans are ready to feel sympathy for robots. One study wanted to prove this. In this study, some pictures were used; they were the hand-drawn pictures

of a human being and a humanoid robot that were both 'experiencing pain' – a knife was cutting their hands. When these pictures were shown to the participants (who were all adults), researchers measured their brain wave reaction. They found out that there were neural responses in participants which demonstrated that humans might attribute humanity to the robots and so empathise with their pain. Even though the participants said that they did not feel sorry for the robot because they knew that robots cannot feel pain, their brain scans told a different story (Suzuki et al. 2015: 6). In time, these empathetic brain responses might make humans more familiar with robots in our society and eventually might alter the way that humans feel about robots and normalise the idea of sex with them. As mentioned, the participants were all adults[35]. Even though they did not necessarily grow up in a digital age, they still felt empathy for robots, so just imagine children born into this era; most probably, they will be much more ready to accept the robots in their daily lives and treat them as though they are human.

It has been predicted that by 2050, marrying robots will be totally normalised (Levy 2007: 20-21). In Japan, it is common to have a relationship with virtual girlfriends (Wakabayashi 2010). Recall the Japanese man who got married to a virtual reality singer. He claimed that he preferred virtual partners to real humans because they are less complicated (Chandler 2018). This is not actually the first time that humans have lusted for AI. Let us recall the mythological robots in the first chapter. In Greek mythology, Pygmalion, who was a sculptor, fell in love with a statue named Galatea that he carved. He believed that he had created an unparalleled beauty for himself and treated Galatea as if she was his girlfriend; he was obsessed with his creation, and he even wished for

[35] Today (in 2023) we might think of adults as having been born as late as 2005, when technology was already relatively advanced. But this research was carried out in 2015; therefore, the adults would have been born in 1997 at the latest. Although robotic technology began a long time ago, as discussed in chapter I, in 1997 the technology was rudimentary as compared to today.

Galatea to be alive. Pygmalion prayed with all his heart and implored Aphrodite to turn his sculpture into a living woman. When Aphrodite saw the statue, she was amazed at Galatea's beauty, and she granted Pygmalion's wish. After a while, Pygmalion realised that Galatea was becoming soft and warm; later she was smiling at her creator. Then, they got married, had a child, and lived happily ever afterwards (Mayor 2018: 107-108).

Let us leave mythology and return to contemporary lives. Humans are already ready to treat their robot companions as if they were real human partners. But treating sex robots as if they are conscious beings will cause a big problem for society. Humans will become attached to them and try to build relationships with them and if that happens, the meaning of love will change. Perhaps, plenty of relationships will break down in the future because of robots. But the most important thing is that people will eventually lose their skills of empathy. In brief, robots do not have empathy, for the possession of empathy requires consciousness, but humans will ignore this and think that they are conscious; they will spend more time with their robot companions; and they will become socially isolated and not interact with human partners anymore. Not communicating with humans, but always spending time with robot partners will lead to a lack of empathy because humans learn and improve empathetic behaviours over time by communicating and interacting with conscious beings. So, after a while, humans will start to lose their empathy skills because they will not experience any real empathy from robots. Thus, humans who lose their empathy skills will feel more isolated since, with impaired communication skills, they will not be able to interact with others as easily; so, they will enter a vicious cycle. Meanwhile a society without empathetic people is likely to descend towards disorder and chaos. Sex is a powerful bond between people, an attraction that brings people together and who then sometimes decide to spend their lives together. The

existence of sex robots might make sex with a human seem second best, inferior – thereby creating a lot of lonely, or isolated, people[36].

Humans are sentient creatures with wide-ranging empathy, except in pathological cases. This enables participation in social life. For example, think about care assistants: they often empathise with their patients because they spend plenty of time with them. After a while, they often become attached to their patients as though they were family members. Also, as said, humans usually tend to attribute human traits to those who act like themselves. So, when these two characteristic features of humans come together, they create a strong tendency to treat unconscious robots as if they are conscious, which is likely to cause social problems. Humans may start to lose interest in other humans and spend more time trying to find the 'perfect' robot partner. Humans will easily accept robots as companions because of their capacities (Levy 2007: 22). Imagine a robot which always obeys your rules; never feels tired; never argues with you; does whatever you wish, whenever you wish; moreover, imagine that you can create every detail of its appearance. This might make humans feel even more attached to those robots and willing to accept them even more easily. But this attachment will only ever be one sided, because robots are unconscious; they cannot feel love or anything at all and their 'expressed' feelings are not genuine. When humans have a real partner, they often ask their opinion; they respect and tolerate each other; that is to say, love is a strong link between people that helps us to form relationships of respect and tolerance. Being in a relationship with someone means that we are able to connect on an emotional level and understand each other. In order to have and maintain a real relationship, we need the ability to empathise. By contrast, a robot which does not have consciousness can neither have genuine emotion

[36] This latter idea was suggested to me by my supervisor; see Tartaglia 2020: 161-180.

nor empathy skills; therefore, although the human might think they are having a reciprocal relationship, they are not.

Sex with robots will increase the number of people who do not feel empathy and people will continue to buy more of them. In general, when someone pays a prostitute, they recognise them as objects, as opposed to fully autonomous humans (Richardson 2015: 290). In one research, some men were asked how they feel about paying someone for sexual intercourse. One replied that 'he feels sorry for the girls, but this is what he wants'; the other said that 'it is just like renting a girlfriend'; 'it is like you just need to choose something from a catalogue' (Richardson 2015: 291). Thus, the prostitute is reduced to a thing whereas the buyer is the only subject. Thus, some people already demonstrate a lack of empathy and see others as objects, because they do not take into account the other person's genuine ideas and feelings (Richardson 2015: 291). But thanks to sex robots, this will be much more normalised in society and humans lose the ability to form the strong emotional relationships which are essential for an ethical society. Therefore, having relationships with unconscious robots is likely to lead humans to become socially isolated (Sullins 2012: 402), for feelings of empathy and intimacy can be developed only by conscious interaction – an interaction which requires mutual consent. Robots, because they do not have emotions, cannot genuinely meet humans' wishes in terms of a real relationship. Admittedly, robots can simulate being a companion for a human. They can simulate being in love; they can mimic feeling something for their human partners. Robots are programmed in a particular way to follow specific orders; therefore, they can simulate emotions. However, as we have shown, simulated emotion (also intimacy and empathy) is not genuine because consciousness is lacking – simulation is not duplication.

Empathising can be thought of as a process related to social cognition and social emotions. These social relations are the interaction of two conscious, intentional, and rational entities. Robots do not meet these criteria. However, behaviourists and functionalists would argue that robots can have empathy. For example, David Levy thinks that robots can develop empathy for humans by observing people's behaviour in various cases, and then making intelligent predictions related to what might be going on in these people's minds to predict future behaviour (2007: 107). Levy takes a behaviouristic approach; therefore, he accepts that whether someone or something can be a friend or a partner just depends on how they behave toward us. However, I would argue that empathy is an emotional or mental process which robots do not have. Also, in order to make predictions, we again have to have consciousness which robots lack. Moreover, there is no way of uploading empathy into a robot because it requires consciousness and conscious learning. Since we have already discounted functionalist and behaviourist theories of mind, the views of those like Levy can be discounted, since they are clearly premised on such theories.

As previously argued, as a result of being in a relationship with a robot, humans' empathy levels might decrease in time and eventually humans might start to not empathise with others at all; and lack of empathy will affect how we treat other humans. Whitby notes that 'an individual who consorts with robots, rather than humans, may become more socially isolated.' (in Sharkey et al. 2017: 21). Kathleen Richardson says that 'intimate relations with robots will lead to more isolation for the human race because robots are not able to meet the species-specific sociality of human beings, only other humans can do that.' (in Sharkey et al. 2017: 21).

As said, empathy is an emotional or mental processes which includes feeling what other people feel, or simulating it to some degree; caring about others; being emotionally

influenced by the emotion and experiences of other persons; imagining yourself into someone else's situation; deducing someone's mental states; etc. (Coplan 2011: 40-65). Without empathy, we cannot achieve a satisfactorily level of communication and social interaction with others. Therefore, it is necessary for intimacy, for trusting someone and for feeling a sense of belonging.

Empathy is important not only for personal relationships but also for society. It helps us to understand or guess what others are thinking or feeling and allows us to respond appropriately to their ideas and emotions. It also helps us maintain social order and cooperation. It helps us to improve mental and physical health. Practicing empathy is the key to understanding and interacting with other people in society. When we share our emotions such as anger, shame, and anxiety, we create an environment in which others can empathise with us. But robots, which cannot consciously have any experience, cannot share their emotions because they do not have any feelings. When humans interact with robots, after they understand that robots do not genuinely understand them, they might feel isolated and confused. When we do not experience empathy from others on the emotional level, we feel isolated. Furthermore, when we do not experience empathy for others, we can feel frustrated.

Imagine a world without empathy. If we were born without empathy, our connection to others would be shallow; it would be based on mutual interests, shared activities, and personal desires. There would be deception and manipulation. Eventually, lack of empathy might make people more selfish. For humans that lack empathy see others as mere objects. Remember the section discussing Kant and deontology in the previous chapter. It was argued that robots cannot treat humans as ends, but by following their programmer's intentions, the programmers would treat humans as mere means. It should now be added that if the programmers have been brought up around robots, they

might have no empathy that might otherwise prevent this. If people develop this lack of empathy, then on a Kantian account, people in our future societies may lose their capacity for moral action.

Without empathy, our relationships would become more like that of other animals. In 2017, some researchers from Georgia State University observed a group of chimpanzees that showed psychopathic behaviour (Young 2017). The researchers realised the chimps were faking emotions to get attention and to affect the others around them. Each relationship which they had was only used as a means to an end based on a constant exchange of favour. Moreover, they did not have any problems about killing each other to get what they wanted. It is fortunate that we possess considerably more empathy than chimps. Imagine that we are already living in a very competitive society in which everybody might do anything to reach their aims. If we all were thinking about only ourselves, society would be different today; for instance, we would be much more likely to cause pain in order to satisfy our own needs.

However, not everybody thinks that society without empathetic people would be problematic. Paul Bloom writes that 'from a moral standpoint, we are better off without empathy' (2016: 10). For him, empathy is a bad thing and makes the world worse. He thinks that empathy is not a solution for society, instead it creates the problems in society. He argues that empathy might be thought to make us good since it makes us more likely to care for others and more likely to help them. But it blinds us to the long-term consequences of our actions. Bloom seems to imply that empathy is selfish moralizing. He says that empathy sets our priorities in an absurd way; it misdirects our action and is used as a tool for violence and aggression. According to him, lack of empathy helps us in taking decisions in the interests of the majority of people. Empathy makes people behave for self-serving reasons alone. He follows a consequentialist account of ethical

theories; therefore, he thinks that we can make the best decisions by considering costs and benefits (Bloom 2016: 87) and he thinks that empathy does not allow us to make this calculation. For example, empathy might make the whole world care more about a baby stuck in a well than about global warming or terrorism. He maintains that most of our failures to make the world a better place and most of our bad actions are motivated by an emotional moralistic rush. Empathic engagement, being caught up in the suffering of victims, is usually the number one argument in a democratic country for going to war. It is how governments convince citizens to go to war, for example, by empathy for suffering victims. Bloom also thinks that sometimes we assert feelings which might not actually exist and could bring us to the wrong decisions (2016: 63). That is to say, sometimes we might misinterpret what someone else actually feels. For instance, just because humans themselves like to be hugged, we think that dogs also like to be hugged, but they actually suffer when they are hugged.

So, Bloom gives us a different perspective by arguing that empathy is a social problem. In my opinion, it is not empathy that is the problem but a lack of empathy and self-awareness. Only those who are without empathy are able to manipulate people to support a war. As humans, we are very open to be manipulated. That is why sex robots can easily 'manipulate' us, or rather their programmers can. Secondly, humans empathise with something that they think is in some way close to them. Therefore, we focus on babies or animals or even robots because we think that they are conscious, and we know what it is like for a conscious being to be in pain. It is wrong to characterise empathy, as Bloom does, as selfish moralizing. For instance, if a person fell in the middle of the street, we would hopefully not just look the other way. We would try to help her/him and ask if s/he needs something more. Sharing the pain of other people cannot be a selfish

moralising. In a world without empathy, we would be much more selfish, and each person would create their own morality.

Without empathy, we cannot understand other people's emotions. In a world where there is no empathy, people could grasp how we feel only if they had exactly the same experiences we had. (If they did not, they would not have any idea what we were talking about.) In such a world, we would feel solitary and isolated. Even if our basic needs were still met, such a world would still be lifeless. As Joan Halifax writes 'a world without empathy is a world which is dead to others – and if we are dead to others, we are dead to ourselves. The sharing of another's pain can take us past the narrow canyon of selfish disregard and even cruelty, and into the larger, more expansive landscape of wisdom and compassion' (Halifax 2018: para. 11).

Moreover, in such a world, there would be no way to feel loved. It would not be possible to trust someone else because trust is based on human compassion. There would not be any way to experience the credibility and autonomy of other people if they did not entirely satisfy our expectations. Eventually, we would be living in a world where no one could be trusted because everyone would fail to meet our expectations. In such a world, everybody would live only for their own pleasure; nobody would matter for anyone else – we would only be objects to each other. Conflicts between humans would increase because nobody would be willing to compromise, and this would bring us to chaos. In short, empathy is very important for humans; the loss of empathy would cause the loss of our own identities. We would continue to exist without it, but our existence would need to be devalued, and it would remain always empty. We would always feel that we were not taken seriously, or not understood. Life would come to feel meaningless.

Let us move onto the second problem. As was said, sex robots are likely to make humans feel lonely and unable to form relationships with other humans. So, those who cannot interact with humans and those who find real sexual relationships overwhelming will start to spend more time with sex robots because sex robots will be easier (you will just pay and get it; you will not need to earn their trust), for they are there for this single function. After a while, in order to meet demand, companies might establish brothels including sex robots. Those who buy sex robots and those who pay for sex in the brothels will devalue sex since when you pay for sex, you objectify the person; that is to say, you make the person a thing, an object or a tool. Here I should emphasize that I accept that sex robots are only tools, so, there would not be any problem if humans see them as tools, but the problem starts when people pay for them as companions and thereby normalise and validate this idea – because companionship itself is devalued. Likewise, sex robots will change the meaning of sex. As previously mentioned, currently sex is a significant bond that brings people together, who sometimes then decide to spend the rest of their lives together: sex robots trivialise this bond.

Psychologist Madeleine Fugere gives us some reasons why sex might make us feel closer to our partners (Dodgson 2018: para. 3). First of all, oxytocin, known as the love hormone, induces feelings of compassion and makes us feel connected to others; it makes us closer to our partner, especially in the initial stage of romantic attachment; it increases feelings of reliance and intimacy between people as well as loyalty (Schneiderman et al. 2011: 1279). The second reason that she gives is that sex allows couples to have a conversation. They might share personal information and sharing secrets might increase intimacy (Aron et al. 1997: 363-373). So, sex is not only about physical closeness, but also about emotions. Knowing each other better might reduce the

risk of disloyalty (Meltzer et al. 2017: 588-591). Thus, sex sustains romantic relationships between people – between conscious beings.

However, sex robots cannot be a part of this romantic relationship. First of all, they cannot have any mental states, so, they do not feel any emotions; they cannot be loyal; they cannot share their personal information because they lack consciousness. That is to say, being loyal and feeling connected to someone can only make sense for those who have consciousness. Humans can be loyal to anything – to a person, to a brand or to an institution – since they have consciousness. But sex robots, lacking consciousness, would not have the capacity for loyalty. That is not to say that every conscious person is loyal to every other, but in order to demonstrate loyalty, someone has to have consciousness as a necessary condition. For example, you might always choose the same hammer to repair everything even if it gets old. But there would be no sense in expecting the hammer to choose you. This applies also to sex robots. There would be no sense in expecting a reciprocal emotion from them because they are *only* tools, just like a hammer. Robots might be programmed to mimic some emotions, but these would not be their genuine feelings because they cannot have any feelings. Therefore, whatever sex robots 'share' with you, there will not be any genuine feeling in it. For instance, we have just mentioned that we might share personal information with our companions. What sex robots can have as personal information is the information that is coded by programmers; that is to say, it can only be the programmer's information. A sex robot that mimics having fun can be created but this is again only simulation. Sex robots cannot have any desire, again, because they lack consciousness. So, it does not matter how intricately it is designed, having sex with a sex robot cannot be the same as having a real sexual relationship.

Secondly, sex is important for humans because it is how we reproduce; that is to say, having sex is necessary to maintain the population and continue the human species. Procreation is the characteristic of all living creatures on Earth. Generally, and basically, when two people love each other they decide to spend the rest of their lives together and have children. Some may choose to not reproduce; however, the majority still want to. Humans with robot companions would of course not be able to reproduce.

So, sex is important for humans not only physically but also for emotional needs and procreation. As said before, sex robots cannot have these needs and interests. But robots might still attract humans by their physical appearance, never getting tired and old, and by their simulated emotions; and humans can treat them as if they are humans too. The danger is that when humans spend more time with robot companions, they will no longer be able to interact with humans and their lives will become lonely and isolated. Sex would become solipsistic, if the person is aware that the robot is unconscious, or delusional, if they are not. Sex will no longer involve a reciprocal attraction between two people. It would lose its meaning as a bond between two people and also its reproductive function.

So, if sex robots became cheap and readily accessible, there is a danger that sex would lose its importance. But first of all, sex should not be something that you can buy anyway. When you buy and sell sex, it means that you are turning it into a commodity. Here I should emphasise again that sex robots are indeed tools, but their use will normalise and validate the idea that sex is a commodity. At a superficial level, this might be attractive. Humans might like the idea that whenever they want, they can pay for sex; they do not need to spend time and effort to earn someone's respect and trust because sex

robots are programmed for this. However, life would become empty and shallow – and sex too.

Moreover, when humans reach things very easily, this makes them feel bored. When you reach something without spending any effort, it loses its magic. Imagine that you work extremely hard for something; maybe for a degree; for a relationship – the result is a hard-earned achievement. When you spend more effort, the reward seems sweeter (Lallement et al. 2014: 348). Spending time and energy on the thing that you really want is important, even though it can be devastating when we do not get it.

Sex robots, designed for that specific function, would not refuse you. But although they might still be the 'subject' of sexual obsession, something would be missing. Humans have sex to show their love to their partners and feel loved by their partners, to give and take pleasure, to make babies, etc. Sex robots would not allow for any of these, and so sex would lose much of its meaning. For the fact is that sex is more than a commodity; and when it is realised that sex robots do not feel genuine empathy, there will be more loneliness. As a result, robots might be made yet more life-like, but we would not get any closer to sex as an act of genuine value (with another conscious and autonomous human being).

Before ending this section, I would like to briefly discuss why some philosophers actually support the use of sex robots and are more optimistic for the future. Some may think that sex robots could replace sex workers and they might decrease rape. In a robotics conference, Ronald Arkin suggested that child sex robots might be used to cure people who have paedophilic inclinations (in Danaher 2019c: 553). That is to say, sex robots might have the potential to be therapeutic tools. Also, Levy does not see anything wrong in having sex with robots, especially for those who cannot achieve a relationship with

other humans; he thinks sex robots can fulfil a genuine need for such people (2007). He believes that the sex trafficking industry can be stopped by sex robots. Levy continues that it would be better for paedophiles to use robots as their sexual outlets than to use children (2007: 14).

For over a decade, a Japanese company, Trottla, has manufactured sex dolls that look like children. However, advances in robotics have not decreased prostitution; actually because of the Internet there has been a considerable growth in this business. In 1990, the percentage of men who paid for sex was 5.6 whereas the percentage rose to 8.8 in 2000 (Richardson 2015: 291). In addition, I think the use of sex robots might cause more paedophilia because it encourages the idea that sex is normal between adults and minors. Crimes might get even worse, for robots sexually objectify children. Rape is a crime (legally and morally); therefore, rapists should not be encouraged to find another outlet for their criminal desires. Furthermore, as Richardson says, sex with robots might simply be seen as a different option on the menu; therefore, it would not necessarily reduce sex trafficking (in Taylor 2017). So, it is doubtful that sex robots can prevent problems related to sexual harassment. But, even apart from these problems, the more significant point is that rapists and paedophiles are in greater need of therapy than of sex robots.

4.KILLER ROBOTS

The final robot group that we will discuss, in order to exemplify the social impact of people treating robots as if they are moral agents, is killer robots. Killer robots are designed to ‘choose’ and ‘engage’ targets without human interference, implying that these unconscious robots will be ‘taking life and death decisions’ (Altmann et al. 2013: 73).

Recently, some countries, such as China, Israel, South Korea, Russia, the United Kingdom, and the United States, have been investing in developing killer robots. A lot of countries have already developed precursors to them such as armed drones. For now, a major part of these weapons is remotely controlled by humans; so, humans decide when the trigger should be pulled (Heyns 2013: 8). But it seems that with technological advances, this will inevitably change and eventually these weapons will not need to be operated by any human operator. That is to say, the algorithm will 'decide' if a human being should live or die.

These robots have been described as the third revolution in warfare after gunpowder and nuclear arms. They can be faster and more durable than a human, and they never feel tired, frightened or depressed; they can kill anyone without hesitation; they do not develop mental disorders from the stress of combat; and they do not require a salary, so they make good economic sense (Scharre 2018). Sometimes, in times of danger and stress, human soldiers, because of their instincts for survival, can make highly dangerous and unpredictable decisions, whereas robots will be pre-programmed to deal with these situations. For all these reasons, killer robots might bring advantages to the countries that own them. Also, in a war, many soldiers die or get badly injured. Even though governments honour those who sacrifice themselves, their families face the results. But imagine an army consisting of drones. First, these drones will enter the city and clear it; after that, real soldiers will invade the city. These drones will be 'learning' as well, so they will improve their knowledge; thus, they will be 'making decisions' to kill or defuse the target without human interference, while continually improving their 'decision'-making abilities. (Piper 2019). Therefore, it is believed that killer robots may reduce casualties significantly by helping with dangerous and risky missions such as entering explosive or radioactive areas or entering combat zones. Moreover, it is assumed

that armies or governments can get better results where killer robots are used since these machines can be more precise and 'humane'. For example, in the World War II, when atomic bombs were used, they killed many civilians and flattened cities. Killer robots would not need to kill so many people, nor destroy cities, because they would be able to be more precise in dispatching military targets. Additionally, it is easy for human soldiers to accept killer robots as team members. Remember Boomer, the bomb disposal robot mentioned in the fourth chapter; recall how other soldiers felt upset when it 'died', and they held a funeral to show it respect. It was even awarded two medals (Garber 2013). We can expect that in the future people will often attribute human features to these tools and accept them into their groups, their armies, and their lives, treating them as valued team members.

There might be many advantages in producing and using killer robots. However, creating killer robots and treating them *as if* they are no different than any human soldiers in the army would be one of the biggest mistakes of humankind. First of all, when they hit or kill the wrong target, there would be a responsibility gap. Who should we hold responsible – programmers, commanders, governments, advertisers, manufacturers, or the killer robots themselves (Heyns 2013: 14-15)? Some people will try to blame killer robots and hold them responsible. But as was said previously, in the first section (driverless cars), in order to be able to take responsibility for an action, it has to be done by a moral agent, and robots have no moral agency. A moral agent is someone who has volition and an intention to carry out her/his aims. In order for something to be an agent, it has to be acting freely and be capable of reasoning (Schwarz 2022: 182). In order to have these features, the agent must be conscious. Killer robots, and other machines, are not conscious. Therefore, they cannot be held responsible for the damage they cause. This problem will recur whenever they damage anything. Consider a landmine. A landmine

does not distinguish anybody or anything and cannot be held responsible for the death and destruction it causes. The responsibility belongs to the person setting it. The same applies to the robots. Robots are not moral agents; therefore, they cannot be responsible for any action. Perhaps, the programmers might be the first to blame for any accidents and casualties that occur because of these robots. However, the governments, commanders, advertisers, manufacturers, each of them will be responsible for the killer robots' actions. But it will not be easy to make them accept any faults. Therefore, there will be conflict in society in attributing responsibility.

I will here keep discussion of the responsibility gap problem brief for, in an earlier section, it was already discussed in relation to driverless cars. But there is a further problem here. We have said that killer robots may bring advantages to governments and soldiers, and soldiers might even become attached to them and believe that they are part of the team, perhaps trusting them to find targets and enemies in dangerous places. But should we allow a machine to make the most important decision about a human life; how can we – humans – allow tools to 'decide' who to kill? The machine which does not have any genuine human mental states, for example, the ability to empathise, feel compassion and to consciously reason – all of which are necessary constituents of a moral being – will kill humans without any hesitation.

Killer robots have automated targeting systems, in which the criteria for whom or what gets targeted is decided by an algorithm that is *written by programmers*. It means that programmers can code anything as a target that they choose. For instance, these robots can be programmed to kill one specific kind of person or group in order to enact genocide. But even if we presuppose that the robots are engaged in a just war, how moral can it ever be to program a robot for killing a person deliberately? When we allow

machines to ‘make these decisions’, we arguably degrade human dignity (Docherty 2014).

Using killer robots would be against human dignity since killer robots do not understand what they are doing. Since they cannot respect the value of life, they cannot imagine the importance of its loss; they cannot calculate the consequences of killing someone (Docherty 2014; Goose and Wareham 2016; Heyns 2017; Ulgen 2016). As Kant says, humans have an intrinsic worth, a value and dignity as rational agents that can make free decisions; pursue aims; and control our actions by reason (1797: 6:387). Human dignity cannot be earned or relinquished; it is an innate status. It does not matter which social class you come from, every person is of value. Remember one of the interpretations of the Categorical Imperative was ‘treat humanity as end in itself’, according to which we should behave towards each other as if we are all subjects that have dignity rather than mere price (Kant 1785: 4:429). But killer robots will degrade humanity by ‘treating’ them as inanimate objects, that is to say, as only a means, never as ends. When humans are targeted by the robots, they are merely objects which must be destroyed (Heyns 2017: 63).

Furthermore, using killer robots is against human dignity since humans might in effect be turned into property by others willing to use the robots to coerce people. That is to say, killer robots might be used to coerce people to hand over money, to work for them, to lie, or anything else that they would not otherwise have done freely. For instance, one company, called Desert Wolf, developed a drone ‘skunk riot control copter’, and sold them to the mining companies. It is designed to deal with rioting crowds. When there is such a problem, it delivers pepper spray and fires plastic balls (Kelion 2014). By using the drone to act as a deterrent to unauthorised gatherings, companies keep control over their workers.

So, these robots might be used to threaten humans in order to make them do the things that the killer robots' owners want. Let us recall the second section of the previous chapter, in which we discussed Kantian ethics and robots. We said that the programmers might design robots with their own morality and indirectly treat humans as if they were objects. For instance, robots can 'manipulate' people or 'lie' to them. Thus, humans may have to do whatever they are asked to do. For example, the robot's owner wants me to leave the cinema. So, s/he might use her/his robot to force me to leave the place and if I did not leave, it might threaten me with physical force and kick me out. Or imagine a drone hovering over your house. It says 'leave this house' if someone tries to enter without permission. If that person does not leave, then it may fire a taser, pepper spray, or give her/him an electric shock, until the authorities arrive to arrest that person (Sharkey 2015). Another example might be as follows. The robot owner, in this case, let us say a loan shark, might collect money from people using the robot: 'you either give me the money or my robot will kill you'. Killer robots can hit targets without any hesitation. So, killer robots will spread fear in the civilian world and the robot owners will force the powerless to obey them. They will treat humans as a means, not as ends themselves, for anyone can be 'treated' as objects by killer robots. It will not matter about your education level, your economic situation, your race, your age, etc., anyone can be targeted. You could be studying and working hard to be a software writer, but one day you might be a target of the thing that you have programmed, or you might be killed by it.

You may not support building these killer machines; however, there will be many governments, nations and terrorists who will find them hard to resist. These weapons may be used by terrorists or despots to achieve their aims, or to kill innocent people, and they might very well pose an unacceptable threat to humanity. One day, these tools may escape from our control and turn into unstoppable killing machines. They could start a war and

cause massive destruction. Tools which do not have consciousness can bring humanity to an end when used by unscrupulous people. With the second revolution in warfare, countries which own nuclear weapons determine the destiny of the world. But once killer robots exist, they will change the game. Therefore, before killer robots occupy all streets and armies, they should be stopped. Otherwise, as many robotic experts have said, 'once Pandora's box is opened, it will be hard to close.'[37].

Some philosophers do not agree that killer robots present a threat to human dignity. Dieter Birnbacher claims that human dignity ought to be applied only to individuals, not to all the members of a species (Birnbacher 2016: 108). Therefore, he does not accept the claim that using killer robots that can 'decide' who to kill is against human dignity. He also asks why it can be ethical for someone to be killed by a human soldier or manned weapons, but not by a killer robot. 'Of course, machines cannot comprehend the value of human life. But why should this make a difference to their victims if alternatively, they are threatened to be wounded or killed by manned weapons like bombers. For the victims whose dignity is at stake it is a matter of indifference whether the threat that they are exposed to comes from manned or unmanned weapons.' (Birnbacher 2016: 120).

The understanding of dignity may change across cultures, times, countries, societies; it may be, in one sense, subjective. For example, slavery is considered to be against human dignity today, but in the past, it was accepted. Aristotle, in *Politics*, thought that slavery could be justified, for example (in Ambler 1987: 390-410). In Islamic societies when someone dies, they are buried because cremation is unacceptable, while in Christianity both cremation and burial are acceptable. The Yanomami tribe practices

[37] This quote comes from the following website: https://www.bbc.co.uk/news/technology-40995835

Endocannibalism, which involves eating the flesh of dead people because they believe that the soul can only rest if relatives eat the flesh and the body is burned (Ukiwe 2018). However, despite these differences, all of these rituals are carried out with respect for the dignity of the dead. It does not matter where you come from, or what social class you are from, every human being deserves dignity, value, and respect. In a general sense, dignity has a shared meaning among all humanity.

Now, let us turn to Birnbacher's second argument that it is no more unacceptable for a robot to decide to kill than a human. The problem is that robots cannot decide at all. Without *meaningful human control*[38] killer robots can only 'decide' who should live or die. That is to say, death will come from an algorithm. Killer robots do not understand anything; they do not have awareness; they cannot take the responsibility for their 'actions', but they will 'decide' who they should kill. Robots will be programmed for a specific target, but later with machine learning systems, they will be able to 'learn' and find the targets without humans' help. Perhaps, they will be able to 'identify' humans as figures after training; nevertheless, it is impossible for an algorithm to *understand* what kind of intentions humans might have. So, there might actually be insoluble problems with identification. For example, someone might be carrying a baby, but the robot might 'think' that it is someone carrying a bomb. Or one child may be playing with a toy gun. Or killer robots might not be able to separate the civilians who are escaping a war zone from the soldiers who are just making a tactical retreat. Here, I should emphasize that I am not saying humans do not also make mistakes, but I am saying that when robots do there will be a serious responsibility gap, and the robots cannot be held back by uncertainty, or afterwards regret what they have done. The speed and scope of machine

[38] Meaningful human control means that humans would still have control over choosing and engaging targets. Each individual attack should be under the control of humans. We can guarantee this only by banning the use of killer robots.

systems will likely create misidentifications because these killer weapons will be ruled by a targeting algorithm. They can spread their information very fast, but, in the chaos of war, they cannot understand the intentions of the people in front of them.

It is important to know the rules but obeying them and making sure others are obeying as well is more important. Wars have always occurred – plunder, ethnic cleansing, massacres – but rules have been established to limit them. These rules indicate how people who take part must act and they require enemies to be respected. Early rules were established by civilizations according to their customs. Later, religious and ethical sources emerged. Modern humanitarian law[39] was founded in 1864 (The First Geneva Convention). It had one basic rule, which was to spare anybody in the war zone who was not taking part in the hostilities. In time, the scope of the law has expanded to protect other people influenced by warfare and some restrictions emerged about the way war is waged. All the countries in the world have officially accepted these rules. These rules are to spare civilians, spare the wounded and sick, spare people who are detained. Civilians cannot be targeted and all people who do not take direct part in hostilities should be saved by the belligerents at all costs. Moreover, hospitals, first aiders and ambulance staff must also be protected. Civilians and soldiers who are in the hands of the enemy are entitled to respect for their lives and dignity. Humanitarian law also bans the use of weapons that are indiscriminate. Distinction, proportionality, and precaution in warfare are important according to the law and have to be respected. International humanitarian law, or the law of wars, says that every human is worthy of respect; everybody has a right to be cared for

[39] By Modern Humanitarian Law, I refer to International Humanitarian Law, also known as the Laws of Armed Conflict, which is the law regulating the conduct of war. It legislates as to how wars will be conducted. Governments respect these laws in order to protect civilians from the effects of war. These rules guide what targets can be legally hit and how, based on a balance between military essential and essential humanity. It has three principles: (1) civilians must be distinguished from combatants at all times; that is to say, civilians must never be attacked; (2) proportionality, which is the principle saying that attacks against military targets are prohibited if these attacks are expected to cause civilian casualties; and (3) precaution, – during military operations there must be constant care for civilians.

when s/he is wounded. Moreover, the existence of this international law allows people to be punished for violations[40]. But we cannot punish robots for their 'actions'. These unconscious robots do not understand and respect the value of human life. However, they will have the power to take humans lives. The laws prohibit war crimes, for example, denying soldiers the right to kill civilians. But being killed by a robot is always against human dignity because they do not understand the necessity of laws, cannot respect human life, cannot understand human intentions (which might make it impossible to programme them to accurately distinguish civilians from soldiers), and above all, can never be held responsible for their 'actions'.

There are some movements against killer robots[41], which argue that robots with the ability to 'choose' who lives or dies, without human interference, cross an ethical boundary, as we have just discussed. On the other hand, there are some people who are optimistic that these robots will be able to make ethical decisions. Arkin thinks that killer robots without human emotions might behave more ethically than humans because humans have anger, hatred, timidity, desires to take revenge and fear, while killer robots can fight without feeling these emotions (2009: 108-113). He adds that killer robots would not loot towns captured in war (2007: 35).

One immediate answer to this is that programming could never be as wise as human judgment. Killer robots cannot 'understand' who is in front of them. There might be a civilian carrying a gun just to defend her/himself and killer robots may fail to 'recognise' this is not a combatant. They cannot understand intentions so they may never be able to separate civilians from soldiers. The target might be programmed as 'a person who wears a soldier uniform with a gun and helmet', but we cannot know in advance

[40] This information comes from the 'ICRC' Website.
[41] Further information can be found on the 'Stop Killer Robots' Website.

whether or not a civilian might wear a helmet, wear a uniform, and hold a gun. Or the programmer might design the robot to 'kill anybody who does not stop when you say freeze'. But in a battle with killer robots humans will be particularly anxious and not know whether to trust the programming of the robots; therefore, they may not stop when they are told to freeze. And if they do not stop, the robots would kill them without hesitation.

It is very important to have a conscious being in a situation which can affect all of humanity, as is vividly illustrated by the following story (Aksenov 2013). Lieutenant Colonel Stanislav Petrov was working at Serpukhov-15 station around Moscow in 1983. He was specialised in the area of a satellite nuclear warning system. So, his main task was to observe the missile early warning system and give an order in the case of an alarm going off, which would have meant a nuclear attack from the USA. The orbiting satellite would counter the missile threat from the USA. When there was any launch coming from the USA, it would be detected quickly by observing the missile plumes over the horizon. During the Cold War, the official protocol of the Soviet Union was to respond automatically when they received any missile early warning. One night, around 00:40, the Soviet missile early warning system gave a signal that the USA had fired a ballistic missile at the Soviet Union. The system indicated the highest level of confidence in this judgement. So, at high alert, everybody took their places. But the launch was not able to be confirmed visually. Petrov *thought* that this warning was a mistake because of the lack of visual confirmation, and he *thought* that if the USA was attacking, it would not be with only one missile. For this reason, he decided to wait without informing his superiors. The team was not able to actually find any errors in the system; therefore, they were getting confused, and the situation was becoming even worse. Then, another alarm went off indicating that there was another missile coming. Later, it continued with a third, fourth

and fifth. But Petrov decided not to launch a counter strike: he *believed* it might be a false alarm. He was right: after ten minutes, the radar stations could not find anything. If he had replied with a counter strike by believing the machine, then doomsday would have been the likely result – it was prevented by the thoughts and judgements of a conscious human being.

Imagine the same scenario but with robots instead of soldiers. So, let us say that the computer system detected a missile and sirens started to ring. Killer robots would immediately respond with a counterattack because the official protocol is to answer automatically when you receive a signal coming from the alarm system. These robots would be programmed already with this rule; therefore, they would not wait and 'think' about whether there might be a mistake in the alarm system. *Thinking* is exactly what they cannot do. They cannot second-guess the automatic system, because they are part of that system. Colonel Petrov understood that there was probably a mistake because he had specialised *experience* in the job. If there was no human soldier in this case but only killer robots, those robots would make a counter strike within seconds, like a 'runaway gun'.[42] If Soviet Union killer robots had launched a missile, the USA killer robots would undoubtedly have delivered a retaliatory nuclear strike. Then, thousands of missiles would be airborne. There would be chaos and mass destruction. The sun's rays would not reach the surface of the Earth, everywhere would be ash and soil, our planet would become a desert, and the robots that caused this destruction would not even realise that something had gone wrong.

[42] Paul Scharre, in his book *Army of None*, uses the metaphor of the runaway gun in order to show the difference between mistakes by humans and by machines (2018: 190-191). The runaway gun is a defective automatic weapon which continues to fire until it runs out of ammunition, without understanding that it is making an error. But at least the runaway gun can be directed by a human operator who can point the weapon at a safe place. The algorithms of killer robots, without being affected by a virus or a programming glitch, could lead to even worse mistakes.

Colonel Petrov decided to wait not only because he assumed that it was a false alarm but also because he thought that even if he was wrong and the alarm system was correct, by not responding fewer people would die and there would be less damage to the Earth. A robot could never be programmed to respond so cautiously, because this would be a strategic error – the opposing side would gain an advantage by programming their robots less cautiously. If we are to remain safe now that we have weapons of mass destruction, we need *human judgments* and *moral understanding*. The development of killer robots is a very dangerous step for us.

If we replace soldiers with these machines – and history has taught us how easily new weapons spread – we might become more tolerant of the idea of global war. Thanks to drones, the USA conducted operations in Afghanistan, Pakistan, and Yemen without worrying about casualties for their own soldiers. But if drones make it easier for governments to commit acts of war, imagine how easy it might be when there are killer robots (Singer 2011: 319). They will behave without thinking of their survival and their destruction will be inconsequential compared to a real soldier, so going to war will be easier. In effect, the use of killer robots will make war seems to matter less, but the more war that occurs, the greater chance there is of escalation. Robots do not have any emotion and compassion that can restrain them. Emotions often stop humans from prolonging wars whereas the lack of emotions in robots will make killing easier for the humans order them from far away. For example, the Vietnam war (1955-1975) a long, expensive, and disruptive conflict in which more than three million people including civilians were killed, only stopped because the US public could not tolerate so many people dying. The governments who will own killer robot army will not encounter death or injury and they will not worry whether their military forces might rebel; robots will follow whatever orders are given, and, without qualms, they will be able to carry out inhumane actions.

All of these factors will make it easier for leaders to consider declaring war (Singer 2011: 395). Thus, armed combat will increase, but civilians are likely to remain in the crossfire – which is against the Humanitarian Law. In any war, humans will die either as civilians or soldiers. Before entering any war, political leaders make cost-benefit calculations. If one army were to consist of killer robots, the number of casualties may significantly decrease; however, this does not mean that nobody would die. The danger is that, when killer robots are employed, the life of civilians might be ignored if the war is thought to bring advantages.

CONCLUSION

In this chapter, we have discussed how the society will be affected by those who will treat robots as if they are moral agents, by discussing the issues related to driverless cars, robots in the workplace, sex robots and killer robots. It has been argued that robots cannot be moral agents, but some humans will be more than ready to be persuaded that they are. When they see robots that act like human beings, that have human appearance and behaviour, that can mimic some human emotions, then they will believe they have consciousness and treat them as if they are subjects. But this will cause serious problems and ethical mistakes in society.

In general, when a moral agent makes a mistake or causes a problem or any damage, we blame and try to punish this person. Therefore, robots, when they are treated as if they are moral agents, should be expected to take responsibility for their actions, too. But they cannot be blamed or punished because they are not moral agents. If we cannot find someone to be held responsible, a responsibility gap occurs in society, and this can bring about chaos, because each institution will blame each other and try to avoid

responsibility for any damage. This was most clearly brought into focus in the discussion of driverless cars.

We then discussed how robots might steal the jobs that are done by humans today, occupying most of the workplaces. People who will lose their jobs might feel depression and start to lose meaning in their lives, for work is one of the vital components of a meaningful human life. If it is taken from humans, they may find themselves experiencing boredom, depression, and an empty life.

Furthermore, we have discussed sex robots in this chapter. Humans' attitudes towards them will again be different. Some will really like them and want to have them in their lives as companions. But this relationship will only ever be one-sided and will cause problems in society. Firstly, humans who spend too much time with these machines will lose their empathy skills and also start to lose their skills at interacting with human beings, especially for forming romantic relationships; therefore, humans will start to feel lonely. Secondly, we have said that sex robots might make sex with a human second best, thereby devaluing human relationships. Thirdly, we have said that they might encourage and normalize objectifying sexual practices, including illegal ones.

Finally, we have discussed killer robots, able to make life-or-death decisions. The programmers will design them with certain targets; after training, they will 'learn' what their targets look like; then they will 'choose' and 'engage' targets without meaningful human control. But the problem with this is that humans will be objectified; they will become simply things to be hit and destroyed. As we said, this is against human dignity because humans should not be treated as a mere means. Also, these unconscious robots cannot make moral judgments, adapt to new situations or grasp intentions, so life-death

decisions may come to be replaced by the 'arbitrary' consequences of algorithms[43] (Heyns 2017: 58).

In the future, humans' attitudes towards unconscious robots will greatly influence society. Some people will assume that they are conscious and like them, some will not think they have consciousness but still like them, some will hate them, and many will try to benefit from them. But the truth is that when they are treated as conscious, this will cause a lot of problems. We have discussed many examples in this chapter. Once humans choose them as employees in their organizations, once programmers design them with their own ethical ideas, once governments take them into their armies, and once humans accept them for companionship, everything will have to change – rules, laws, relationships, meaning in life, etc… But perhaps the situation is even worse than this. Perhaps at some point, technological growth might become uncontrollable, and at that point, these unconscious learners might surpass humans in many more ways than they do already. In the next chapter, we will discuss this 'singularity'.

[43] Heyns uses the term 'arbitrary' to refer to a machine's mistaken targeting: for example, 'targeting' and 'defusing' the child who is playing with a plastic gun which the robot 'thinks' is real.

CHAPTER 6: THE TECHNOLOGICAL SINGULARITY

INTRODUCTION

As discussed in the previous chapter, people treating robots as though they are conscious beings will cause many ethical problems, but because of their potential benefits we may still expect their growth, both in number and in the fields in which they are used. But we should not forget that these machines can unconsciously 'learn' and 'improve' themselves; therefore, they may not be always under the control of humans. Perhaps, we may arrive at a point in the future where the growth in almost every field of science and technology will be uncontrollable.

As previously discussed (in the Chapters 4 and 5), we now have different types of extremely capable machines, from driverless cars to killer robots, and in all of our lifetimes, the speed of technology has increased exponentially (Ford 2015). Consider aircraft, for example. Leonardo da Vinci sketched an 'aerial screw', a helicopter-type contraption, in the late 1480s, but it was not until over 400 years later, in 1939, that the world's first helicopter flew. Within 100 years of the Wright brothers' first powered flight in a small wooden one-man aircraft in 1903, we have built passenger planes which can carry hundreds of people, flown humans to the moon, and sent a rover to Mars. Today, we look for other planets to colonise. Thanks to the growth in technology, we are able to build new technology even faster. Furthermore, these technological tools are able to 'learn' and 'improve' themselves; thanks to machine learning, they can 'evolve' rapidly; thus, eventually, technological growth may become uncontrollable, and this brings with it the idea that one day machines will become smarter, in every respect, than humans. This phenomenon, called 'the singularity', is the subject of this chapter.

In this chapter, we will discuss what the technological singularity is; how likely it is; and, if it happens, what kinds of advantages and disadvantages it might bring. That is to say, what should be our practical concerns relating to the singularity, and last but not least, what should be our philosophical concerns relating to the singularity. The springboard for this discussion will be David Chalmers' article, 'The Singularity: A Philosophical Analysis' (2010).

1.DEFINITIONS OF THE SINGULARITY

Imagine a future where machine intelligence overtakes human intelligence. What would the consequences be? One idea is that an intelligence explosion will occur after this event and algorithms will turn into super intelligent machines surpassing the cognitive capacities of humans. There will be a point reached where an algorithm will be capable of self-improvement. The new generation of algorithms will be even more powerful and capable of improving themselves even more. This will lead to a chain reaction – a superintelligence that upgrades itself and accelerates growth at an incredible rate. Each generation will be able to design the next generation better than itself. This process will continue exponentially. Thus, we will reach the singularity. Even though there are some disagreements about the meaning of the singularity, most people agree that an event will occur in which we will observe the rise of superintelligence. Another point most people agree upon is that algorithms will increasingly improve themselves.

There are actually different ways to explain the term 'singularity' (Shanahan 2015: xv). Generally speaking, the term refers to a unique or unusual event with enormous impacts. In mathematics, the term 'singularity' is used to define the point of a function at which we are no longer able to describe its exact properties; namely, it is the point at

which a function is undefined[44]. For example, when we are using a calculator to divide numbers into smaller and smaller numbers, it eventually says 'error, division by zero'. This is the point where our understanding of mathematics breaks down and this point is called a 'singularity'. The term is used also in physics (Curiel 2009). Imagine a black hole and event horizon around the black hole. The event horizon is often called the 'point of no return'. It is the border in which gravity becomes very strong and anything that passes it cannot come back out. According to general relativity, after the event horizon, physics begins to behave differently. Our current knowledge tells us that gravity and density are infinitely large at the centre of the black hole. But when we say a physical attribute such as gravity is infinite, the laws of physics as we know them cease to function. Therefore, our understanding breaks down. This point is known as a 'gravitational singularity' (Earman 1995). So, in general terms, when things begin to act in a strange or unusual way and we cannot understand and find any answers, then we call it a 'singularity'.

Now, let us return to 'the singularity' in technological development. To provide a little background: Jon von Neumann, a mathematician, was the first person to use the concept of the singularity in a technological context in the middle of the 20th century. He referred to 'the ever-accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue' (in Ulam 1958: 5). Many writers have been influenced by this idea. The first point of interest is that he noted that 'human affairs, as we know them, could not continue'; the second point is his reference to 'the ever-accelerating progress of technology'.

[44] This information comes from the 'Free Dictionary' Website.

When he said, ‘human affairs, as we know them, could not continue’, he did not mean that the human race will die or become extinct. Instead, he means that humanity, as we know it, could not continue. For example, one idea is that humanity will abandon their biological bodies and human intelligence will be transferred to machines (Kurzweil 2005: 35-44). In any case, something that has never been seen before will cause humanity to experience a unique event, and our current knowledge, in this context related to human affairs, will break down – just as happens when the laws of mathematics and physics break down when encountering a highly unusual event. This is the origin of the idea of technological singularity – a concept whose roots are based on the mathematical idea of a point where an object cannot be defined, combined with the observation of an incomprehensibly rapid development that will have a profound consequence which we cannot predict.

Turning now to the second point, when Neumann wrote of ‘the ever-accelerating progress of technology’, he was referring to ‘the law of accelerating returns’. Today, technological progress is faster than ever before. Between 1910 and 1950, the speed of computers used to double every three years whereas between 1950 and 1966 it was every two years. Now, it is doubling every year. This development in technology is called the Law of Accelerating Returns (Kurzweil 2005: 44-106). According to the Law of Accelerating Returns, we will eventually reach a point at which technology expands so rapidly that it completely escapes our control. At that point, it will become impossible to guess its consequences for the future of humanity.

It is important to recognise that the growth of computing power is exponential. The exponential change in computers and other technological tools will result in iterative development; the newest generation will be faster than the previous one and will be used to create even faster tools. In order to understand its importance, let us use an ancient

Indian chess legend as an example (Aron 2015: 35). According to the legend, King Shahram of India was a big chess enthusiast and he used to challenge visitors to a game. One day, a traveling sage came to the King's palace, and he wanted to play chess with the King. The King accepted and asked the sage what he would want as an award in case he should win. The sage asked for rice. One grain of rice was to be located on the first square of the chess board, two grains on the second square, four on the third, eight on the fourth, sixteen on the fifth, thirty-two on the sixth, and so on, doubling the number of grains on each square until all the squares were filled. The King accepted and the game started. The sage won the game and the King wanted to grant his wish; therefore, he ordered a bag of rice. He started to place the rice on each square, doubling each time. When the King came to fill the 12^{th} square, he was shocked because of the number of grains was 2048. Then he figured out that for the 30^{th} square, he would need about 536 million grains; by the 40^{th} square, he would need 550 billion and for the last square 64^{th}, he would need 9 million 220 thousand trillion. So, in total, the amount of the rice would be around 18 quintillion grains. In mathematics, this kind of growth is defined as exponential. As Kurzweil says 'exponential growth looks like nothing is happening, and then suddenly you get this explosion at the end.' (in Lamb 2005). Now, imagine that the rice is a measure of computing power. If computing power increases in the same manner as the rice, we will reach a point at which technology will begin to expand so rapidly that it will be uncontrollable. The exponential growth will make any predictions about the future of humanity impossible. So, this is the point where we will enter a technological event horizon which is just like the event horizon around the black hole. And we will be completely unable to predict the future. This point is called 'the singularity'.

In 1960s, Irving John Good set up the basic argument for the singularity in his article, 'Speculations Concerning the First Ultra-intelligent Machine': 'let an ultra-

intelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus, the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.' (1964: 33). In 1980s, Vernor Vinge popularised and expanded the topic: 'we will soon create intelligences greater than our own. When this happens, human history will have reached a kind of singularity, an intellectual transition as impenetrable as the knotted space-time at the centre of a black hole, and the world will pass far beyond our understanding. This singularity, I believe, already haunts a number of science-fiction writers. It makes realistic extrapolation to an interstellar future impossible.' (1983: 10). It is assumed in these arguments that a future in which a robot is more intelligent than human beings will be one in which robots are better than humans at designing robots. So a machine may be capable of designing a machine which is more intelligent than the most intelligent machine which human beings might design. It will be able to design a machine which is more intelligent than itself, even. By the same reasoning, the machine-made machine will itself be capable of designing a machine which is more intelligent than itself (Chalmers 2010: 1-2). So, once the machines reach and surpass human-level intelligence, they will be able to create more sophisticated and advanced tools because they will be improving themselves and performing better; thus, they will be able to invent better versions of themselves – and from that point onwards the process will simply continue. It is assumed that AI will take off in a runaway response of 'self-improvement' era with each new and more intelligent generation occurring very quickly, causing an explosion in intelligence and resulting in a powerful super intelligence which surpasses all human

intelligence. In this manner, many believe that the machines will transcend human intelligence. Eventually, we will arrive at a point where there will be an *intelligence explosion*.

2.PRACTICAL CONCERNS RELATED TO THE SINGULARITY

There are possibly some physically limiting factors that might prevent the singularity. Firstly, the amount of data and the required memory storage of the proposed machines would be huge. Secondly, if our current knowledge about the theory of relativity and quantum mechanics are true, then indefinite extension cannot be expected because energy is finite in the universe. However, as Chalmers has argued, even if there are these physical limitations for the singularity, there are still reasons to consider that speed and intelligence explosion may be pushed to the limits of what is physically possible (2010: 2). Our understanding of the universe is still far from complete. Speed and intelligence may still far exceed human capabilities.

The coming singularity has been discussed for many years. In 1964, Good predicted: 'a singularity is more probable than not that, within the twentieth century, an ultra-intelligent machine will be built and that it will be the last invention that man need make.' (1964: 78). In 1993, Vinge predicted that 'within thirty years, we will have the technology means to create superhuman intelligence. Shortly after, the human era will be ended.' (1993). Alan Turing estimated that human-level AI would be achieved by 2000 (Sharkey 2012). Kurzweil (2005) claims that the processing power of computers will allow them to become artificially intelligent (self-generating) by 2045. Chalmers predicts that there will be human-level AI before 2100 (2010: 6). So, there is no widespread agreement as to when the singularity will happen.

If the singularity happens, it might bring many benefits. Humans would no longer need to do boring or dangerous tasks. Asimov envisaged that robots could be willing slaves that might do tasks that surpassed our capabilities. Thus, we would have more time for other activities. The singularity could reverse the effects of climate change; it could end poverty. Our medical technologies would become more advanced. AI might figure out a way to cure all diseases and ailments; for example, it could insert billions of nanorobots inside a human body to repair cell damage. Moreover, it might offer immortality. It is thought by some that we may be able to even upload our consciousness into a computer and leave our mortal bodies behind to live forever. That is to say, humans in the future may be able to overcome death. Kurzweil says in an interview that we will be able to live forever – 'I believe we will reach a point around 2029 when medical technologies will add one additional year every year to your life expectancy. By that I do not mean life expectancy based on your birthdate, but rather your remaining life expectancy.' (in Nagesh 2016).

The proponents of the singularity assume that we would be wrong to fear the end of humanity because what replaces us will be much better, just as homo sapiens were superior to homo erectus. However, many thinkers continue to worry about the fate of humanity in the face of this development. Stephen Hawking said that 'humans who are limited by slow biological evolution, could not compete and would be superseded' (in Cellan-Jones 2014). Bostrom writes that 'when we create the first super intelligent entity, we might make a mistake and give it goals that lead it to annihilate humankind, assuming its enormous intellectual advantage gives it the power to do so. For instance, we could elevate a subgoal to the status of a super goal. We tell it to solve a mathematical problem, and it complies by turning all the matter in the solar system into a giant calculating device, in the process killing the person who asked the question.' (2002: 7). So, advanced AI

might pose an existential risk. It might decide to behave quickly before humans have a chance to react with any countervailing action. It might decide to eliminate all of humanity for reasons which might be incomprehensible to us. It might seek to colonise the universe – either in order to maximise its powers of computation or to obtain raw materials for manufacturing new super-computers. Humans may not be able to entirely understand an artificial super intelligence because the intelligence of the artificial super intelligence would be much greater than the smartest humans. It is likely to be difficult for humans to accurately perceive or understand its calculating processes. Super-intelligence would not necessarily behave benevolently towards humans either. How we are treated might be based on our past behaviour. For instance, we do not hate ants; but we do not change the course of our roads in order not to harm them. Sometimes we might walk around them but if their presence seriously conflicts with our aims, then we annihilate them without any thought.

3.PHILOSOPHICAL CONCERNS RELATED TO THE SINGULARITY

The singularity raises not only practical questions but also philosophical questions. These will now be discussed alongside further discussion of the nature of intelligence and mental capacities of AI.

Chalmers, in his article, gives the following argument for a singularity: '1) There will be AI (before long, absent defeaters[45]). 2) If there is AI, there will be AI+ (soon after[46], absent defeaters). 3) If there is AI+, there will be AI++ (soon after, absent defeaters). 4) Therefore, there will be AI++ (before too long, absent defeaters)' (2010: 6). Chalmers uses AI to refer to an AI with human-level cognitive capacities. This is often

[45] Chalmers refers to global catastrophes as 'defeaters', such as nuclear war, asteroid impact, etc.; he refers to 'within centuries' as 'before long'. (2010: 6-7).
[46] He means within years by 'soon after'. (2010: 7).

called ‘artificial general intelligence’. (This is a computer equivalent of human intelligence which can ‘think’, ‘strategize’, ‘plan’, ‘decide’, ‘reason’, etc.) In order to make the idea clearer, let us put the argument as follows: If we could make AI, then we could use it to make AI+. If we can build a computer which is as intelligent as humans, then it ought to be able to do the same. Then, AI+ would produce AI++ and so on; and thus, there would be an ‘intelligence explosion’.

In order to reach the later premises, the first premise (*Premise I: There will be AI.*) needs to be established. There are several different ways in which that premise could be realised (Chalmers 2010: 7-9). I will discuss four of them here. The first way is to discover one basic learning algorithm and let the computer build its own intelligence. But as discussed previously, the computer can only do what it is programmed to do. That is to say, it cannot learn consciously; therefore, it will never independently build a human-level intelligence.

The second way to reach premise 1 would be to write all the software from scratch, which would probably include trillions of lines of code. Thus, we program the intelligence directly. In order for this idea to be plausible, we would have to argue that intelligence is ultimately only symbol manipulation. This is the claim made by Allen Newell and Herbert A. Simon. They claim that intelligence does not have to have a connection with any particular variety of biological or physical wetware or hardware (1976: 114), which is a typical functionalist belief. According to them, the manipulation of symbols constitutes the basis of every intelligence action. But we have already rejected functionalism. Intelligence gives humans the abilities for acquiring knowledge, understanding, reasoning, deciding, solving problems, communicating, establishing goals, building relationships, etc. It enables humans to experience, think and value. These cognitive abilities are not only symbol manipulations, for there are meanings that are

attached to them. Computers can only manipulate meaningless symbols (Searle 1980: 417-457) (see chapter three in which we discussed CRA). Intelligence requires more than formal symbol manipulation. Intelligence is not a result of the interaction of neurons, but a manifestation of the mind; and machines do not have minds. Intelligence is not an objective aspect of a system which can be reduced to the behaviour; it is a matter of subjective experience involving consciousness; therefore, it cannot be computed.

Today, there is 'artificial narrow intelligence' that refers to the AI specialised in only one area, such as, self-driving cars, a chess player robot, killer robots, sex robots, etc. We do not have artificial general intelligence that is equivalent to a human-level intelligence, and it seems that we never will because intelligence is not in the machines, but in the people who develop them. So, we cannot claim that a machine or a robot is potentially intelligent. However, sometimes we attribute features that belong only to conscious beings to objects that are not conscious, on the basis of their behaviour (as discussed in the previous two chapters). For example, we might say that the killer robots can understand and engage the target, driverless cars can decide which road they should take, or sex robots can share a life with their partners.

Searle refers to this metaphorical intentionality as 'as-if' intentionality (1998: 93).[47] For example, a statement like 'the care-house robots are very empathetic' would be nothing more than an expression of *as-if* intentionality. Thus, colloquially we may ascribe an intentionality to robots or machines which they do not actually have. Just because they behave *as if* they are intelligent, people have a tendency to ascribe

[47] Searle says that there are two kinds of genuine intentionality that are intrinsic and derived, but he adds that there is also 'as-if' which is a metaphorical intentionality. 'Intrinsic intentionality' does not depend on anyone, that is to say, it is not observer-dependent; it is a characteristic of mental states; for instance, 'I am hungry right now' states an intrinsic intentionality. By contrast, 'derived intentionality' depends on observers; it is a feature of their language. For example, a statement like, in French, 'J'ai grand faim en ce moment' meaning 'I am very hungry right now' refers to derived intentionality (1998: 93-94).

intelligence to them. However, just because a computer can calculate faster and more accurately than we do does not mean that it is as intelligent as a human being. It does not matter how advanced the machines are, their essence will remain the same. They will still 'perform' a predetermined set of instructions and rely on their programmer's ideas. When they become increasingly complex, some may think that they are becoming intelligent. But a machine that acts *as if* it understands or is intelligent is different from a machine that *actually* does understand or is intelligent. (We will return to this argument later.)

The third way is to argue that evolution will be simulated. According to Chalmers, achieving simulated evolution is a hard problem, but it is an easier problem than creating intelligence itself (2010: 10). The argument basically says that evolution is the process by which we developed from unintelligent matter. That is to say, human-level intelligence has been produced by evolution. Nature did it after all, so why should it not do it again? But this argument seems a bit problematic because evolution does not always lead to intelligence. For instance, modern humans have been on Earth for around two hundred thousand years (Howell 2015) while the dinosaurs were on Earth for roughly 200 million years, but no species of dinosaur became particularly intelligent. So, longevity does not necessarily lead to the evolution of intelligence. What made humans intelligent was a virtuous circle involving both biological evolution and social development.

Some philosophers think that if the human brain is copied very well, then in principle there is no reason for machines not to have consciousness, and thus intelligence. If we accept that consciousness is a biological process like digestion, and if the brain is a machine, then the first step is to find out how to copy the brain. Once we figure out how it functions, we can build a brain which can have an equally effective mechanism for causing consciousness. This is the idea of 'brain emulation' which will be the last argument that I will discuss in this section.

The idea implies that in principle it is possible to copy or upload a human brain onto a computer. The basic idea is that after all, human organs are machines, and the brain too is a machine even if it is the most complex one. If we can emulate a human brain, then we could have AI. 'Emulation' is similar to 'simulation'. However, emulation focuses more on engineering. All we need is a human brain, a computer, and a machine. The machine will scan your brain and transfer the data in your mind to a computer. So, you take a brain and scan its structure, then build its software prototype, and when used on a suitable hardware, it could act in essentially the same way as the original brain (Sandberg and Bostrom 2008: 7).

Functionalists have suggested translating all the inputs and sensory data in the brain via calculations into outputs and our behaviour. This is where the arguments of copying the functioning of a human brain and uploading it onto a computer emerge. According to functionalism, mental states are functional states; therefore, if the functioning of the brain can be mapped, they believe that our consciousness can be copied to a computer. In addition, remember that for functionalists, a non-biological system can be conscious as long as it is created appropriately. The brain is made of particular biological material, but, according to functionalism, what defines a brain is not what it is made of, but how it functions. Therefore, it is in principle possible to create an artificial brain that can do the same job as a biological brain. In the same way, a heart is a heart as long as it pumps blood, no matter what it is made of. So, if we could copy the functioning of a brain, then we can claim that this copied brain can produce all the mental states of a 'real' human brain.

As we have already seen, however, machines cannot have consciousness even if the human brain can be exactly emulated, but functionalism is a false theory of mind. We can design and create a very complex machine which is capable of performing various

tasks or mimicking our behaviours and it might be able to pass the Turing test. But that would just mean it was able to fool someone; the machine will still be following the instructions that are created by its programmer. Kurzweil predicts that by 2029, a machine will possess the intellectual and emotional capabilities of a human; thus, it will be able to pass the Turing test, that is to say, it will be impossible to tell whether or not we were speaking with a human being (2005: 167). But there is a big difference between making a machine that acts *as if* it understands or *as if* it is conscious and making a machine that *actually* does understand or it is *genuinely* conscious. In order for the singularity to happen, the latter is required because the singularity will only happen once AI is able to *genuinely* think, dream, create, decide. A singularity requires that the machines have intelligence and understanding that surpasses ours, but if they are not conscious then they have no intelligence and understanding at all. There is nothing we cannot understand, and nothing they understand either – it would just be that we have lost control of our tools, which are now acting in erratic and incomprehensible ways that might be dangerous to us.

A machine that only acts as if it is conscious would be nothing more than a zombie (Chalmers 2022: 285) – a digital zombie with the copied human brain but without genuine consciousness, intelligence, or any mental states at all. However, if these zombie machines can act more efficiently than us, steal our jobs, take a big role in our societies, or even destroy us, and if society treats them as if they have minds, and as if they are moral agents instead of merely tools, then it does not much matter if they are actually unconscious. To a brick, it does not matter how precious a glass is, if it hits the glass then the glass will break. The same thing applies to these machines. In a practical sense, it does not matter whether they are conscious or not if one day they will steal our jobs and companions, if they will make us less empathetic, if they will cause responsibility gaps

in society, if they will hit us like a brick and perhaps even terminate us. Therefore, it does not matter if they are conscious or not from the point of view of the threat they pose to our societies.

Once again, I shall make the claim that robots with emulated human brains but without consciousness would not be enough for the singularity because we need AI with human-level consciousness in order for the singularity to happen; and in order for the dream of immortality to come true, for instance. We can build complex killer robots, sex robots or worker robots, and they might do their tasks efficiently, but they will never be intelligent. Computers might become 'smarter' than humans; they can 'learn' quicker than us; they can 'calculate' faster than we do; they can certainly be stronger than us; etc. But these features do not prove that machines are capable of genuine intelligence. So, there is no evidence to suggest that machines will ever surpass human intelligence.

However, once machines or robots *seem* to have consciousness and human cognition, people will treat them as though they are genuinely intelligent. Once they can pass the Turing test, when people will not be able to distinguish between a human being and a machine, some people will see them as intelligent as humans, and as equivalent to persons in society. But this will not be a singularity – if people ever genuinely believe this then it will only be a singular philosophical mistake.

CONCLUSION

The technological singularity has been a topic of interest not only for science fiction authors but also philosophers, scientists, and engineers for many years. But so far, AI capable of *apparently* thinking, making its own decisions, independent problem solving and reasoning, does not exist. Such tools may never exist. I have argued that AI

capable of *actually* thinking, making its own decisions, independent problem solving and reasoning, however, will never exist.

The singularity could occur only once AI had reached a human-level of intelligence. In order to achieve this, the whole brain emulation route seems to be the most promising route. But when a device is emulated by another device, insufficiently understood features of the original may be missed out, and in this case it would be human cognition, consciousness and intentionality. So, even if one day we can emulate the brain, it does not follow that the emulation itself will have a mind or be intelligent. The system will always lack such integral aspects as consciousness, intelligence, understanding, and it is mainly functionalism that makes people think otherwise. There have of course been many improvements in technology and there will be many more to come. It is predicted that machines and robots will continue to surpass us in many different areas. However, I remain sceptical that human-level AI will ever come to exist and surpass our own intelligence, even as a mere simulation.

CONCLUDING REMARKS

Humans create new technologies and products and improve them according to the knowledge of their times. From ancient times to today, we have always looked for better technologies. We have never stopped inventing technology to try to make life easier. One of these technological products are robots, the idea of which has amazed human beings since at least ancient Greece. We began building them as a tool to help us with daily tasks. Then, we developed them to get them do more and more tasks. Later, we made them calculate, walk and speak. And today, we have arrived at a point where discussions about whether robots can be conscious do not seem odd.

In this thesis, I have argued that robots cannot be conscious. In order for something/somebody to be conscious, I have often emphasized that there should be genuine understanding, knowing, learning consciously, intentionality, sense, perception, belief, desire, self-awareness. There should be something that is it like to be that thing. The main aim of this thesis, as I hope you have realised, is to emphasise the internal features of mentality such as genuine understanding, intentionality, reasoning, rationality, empathy skills, the ability to use language with true understanding, and the experience of pleasure and pain. None of these features are emphasised by those who support the idea that we can build conscious robots. When they discuss them, it is to answer people who object to their idea.

In the first chapter, I explained how in ancient times, humans were thought to be special whereas robots are seen only tools. Then when we arrive at the 17th century, Descartes challenges this idea. According to him, the human body is a machine that is controlled by the mind. However, this does not mean that he claimed that human beings are robots. The human mind is still special in Descartes' thought. Humans have

rationality, reasoning, free will, consciousness; therefore, they cannot be a machine. Nevertheless, the human body works like a machine. It gives certain responses to the certain stimuli. In addition, Descartes makes a clear distinction between animals and humans. He thinks that animals are machines since they lack consciousness. Thus, I believe that his ideas are very important in the studies of 'conscious robots' since he might be claimed to be the first philosopher who thinks living things (animals) are essentially machine. Later, La Mettrie refused to draw a distinction between humans and animals. He denied that there are two separate substances, mind and body. He reduced everything to the material. Thus, the idea of machine man emerges. Humans are machines too. Descartes paved the way to the idea of conscious robots, but left the obstacle of the immaterial mind, which later philosophers like La Mettrie sought to remove.

Descartes' dualist approach has been criticised for many centuries, especially by materialists – the obstacle he left to conscious robots needed to be removed. It was argued that Descartes had not adequately explained the mind nor the interaction between mind and body. In the second chapter, I analysed four materialist approaches (behaviourism, the identity theory, eliminative materialism and functionalism) to discuss how they seek to address these problems, and how encouraging they are to the idea of conscious robots.

Behaviourism basically claims that the concept mind derives entirely from bodily behaviour and inclinations of the body to behave in particular ways. Thus, if a robot shows predictable dispositions to behave in certain ways in certain situations it can be said to have consciousness, on one (non-eliminative) interpretation of behaviourism. If behaviourism were right, and the mind is nothing more than its exhibited behaviour, then there would be no doubt that a robot can be conscious. But I argue that behaviourism cannot be right because the mind is not just behavioural. There are three reasons for this. Firstly, behaviourism cannot explain the causal roles of mental properties and mental

states. So, this causes a circularity problem. When we try to analyse beliefs in terms of behaviour, we refer to desires; when we analyse desires in terms of behaviour, we refer to beliefs; etc. Secondly, behaviour does not always reflect someone's actual mental states. For example, an actor may just mimic being in pain even though s/he is not in pain at that moment. Finally, I argue that behaviourism does not explain our individual conscious experiences and their qualia. For example, my experience of eating chocolate may be different from yours, although we both might display the same behaviour. For these reasons, I feel justified in dismissing the behaviourists' claim that it is possible to build conscious robots.

I then move onto the identity theory, which claims, basically, that the mind is identical to the brain, or more exactly, mental states are identical to physical states in the brain. If that is right, then if we can physically replicate the brain's neural network, robots can become conscious. I argue that mental states cannot be identical with brain states, however. First of all, if we claim that two things are identical, both should have the same aspects, and yet neural states have plenty of properties which conscious properties lack, and vice versa. The second major problem is that there seem to be some animals that feel pain like humans but do not have the same neural structures. And anyway, although we might associate particular physical states with mental states, this does not demonstrate identity – It could instead be a causal link, as Descartes thought. For these reasons, the mind cannot be identical to the brain and there is no good reason to think it is. Therefore, even if we could build a robot that has the same neurophysiological aspects of a human being, the claim that it would be conscious is unjustified.

I then discuss eliminative materialism (or eliminativism). Eliminativism denies the existence of mental states. So, according to eliminativism, our beliefs, desires, pains, consciousness do not exist. And if there is no such a thing as consciousness, then we

cannot raise the issue of whether we can build conscious robots – although you might say that we can build them, in the sense that the eliminativist thinks the idea of a conscious robot is as confused as the idea of a conscious human, so there is no reason to think the robots we build will be greatly different to us due to lack of consciousness. The best question to raise, then, is whether humans are unconscious robots. If we are unconscious machines and if we can create machines that resemble and act the same as us, then there would not be any significant difference between them and us, if eliminativism is right. But, there are extremely good reasons to doubt eliminativism. The theory is self-refuting in that it claims that there are no beliefs and yet it itself proposes a belief. Hilary Putnam said that, many philosophers agreed, and I have not found any good reply.

I then examine functionalism, which seems to present the most promising argument for the idea that we can build conscious robots. It is also the argument that has been most influential on the idea of building conscious robots. It seems to solve some of the problems of behaviourism and the identity theory and is widely popular among philosophers. Functionalism claims that mental states are functional states. The brain is like computer hardware and the mind is like its software. In fact, for some forms of functionalism this is more than an analogy: the brain actually is a computer. This computational view of mentality suggests that mental states are multiply realized. The same computational process might be carried out by physically different computing machines – by digital computers or by computers with gears and wheels. They all perform the same process of computation. If minds are like computers, then mental processes are computational processes. In the same way that distinctive physical devices may perform the same computational programs, different biological or physical systems ought to be able to execute the same cognitive processes. This is the core idea of functionalist theories. If the mind is a computer program that runs in our brains as functionalists claim,

then when we get the right program, we can create consciousness. Functionalism seems to provide the strongest argument that it may be possible to build conscious robots, as well as the most practical engineering project.

However, in the third chapter, I argue that functionalism fails to explain mental states. I believe that mental states cannot be functional states. The human mind does not work like a computer. All computers do is manipulate formal symbols (syntax), but these formal symbols are not meaningful for the computers. Whereas, the mind works by more than syntax and computation. The mind has intentionality, meaning, understanding (semantics). By contrast, machines can only simulate these mental states, which is different from duplication. In the course of making this argument, I examine Searle's famous thought experiment of the Chinese Room. Searle's thought experiment suggests that formal symbol manipulation is not sufficient for genuine understanding and simulation is not duplication. I also agree with Searle's later claim that even syntax is mind-dependent – that computers only manipulate symbols systematically as interpreted by conscious people.

Moreover, I claim that qualia cannot be exhausted by their functional roles since they are subjective conscious experiences; they are concrete whereas functions are objective and abstract. So, I argue that even if two systems are functionally the same, this does not prove that they experience the same qualia. Thus, when we may look at the same thing, our behaviours might be functionally identical, but one of us may not be feeling anything at all or may be feeling something totally different. Functionalism is also vulnerable to the multiple realizability argument. Functionalists themselves use this argument to refute identity theory, but it also poses a problem for functionalism. According to functionalism, pain is identified by certain functional roles such as crying, groaning, or looking for a pain killer. But it might be that only humans look for a pain

killer when they have pain; millions of alien species who feel pain may do something completely different. So, mental states cannot be defined by their functional roles. Furthermore, if mental states are functional states, that is to say, if they are abstract, then there occurs a problem with explaining the interaction between the physical and mental states. Functionalists cannot explain how they can relate to each other. This is also a problem faced by dualists. So, my conclusion is that there is no good reason to think that the problems with functionalism can be overcome. The functionalist argument that conscious robots could be produced is flawed.

Nevertheless, people might still think that robots can be conscious and treat them as if they are conscious beings. In fact, humans often attribute human features to objects that are not conscious. This is because we have empathy skills. When we see something that looks like a human and that is in pain, we may immediately feel sorry. And, regardless of its many flaws, many people may find functionalism appealing, in that psychologically its analogy (the brain is a computer, and the mind is its software) makes sense for this current age. The human mind and the functioning of a human brain have amazed us for many centuries. Similarly, computers and their programs are the latest technology which fascinate us, as well. An overarching theory explaining both of these phenomena has an intuitive appeal. Nonetheless, for the reasons given, I do not believe that robots can ever have consciousness.

In chapter four I move onto some ethical issues. I argue that even if people believe that we can produce conscious robots, there is no school of ethical thought, from among the major ones of the history of Western philosophy, that suggests we can build a robot that acts ethically. I argue that robots cannot be said to act ethically or non-ethically towards us because they lack consciousness. They will always follow their programming/software. In making this argument, I focus on three main ethical theories –

consequentialism (esp. utilitarianism), deontology and virtue ethics, and I claim that none of these three ethical theories allow robots to be ethical.

I begin with utilitarianism which claims that good and bad action depends on the consequences. According to classical utilitarianism, if an action causes the highest pleasure for the greatest number, this action is ethically good. In order to choose the action that will cause the highest pleasure, consciousness is required. However, consequentialists focus on only outcomes, and so they ignore the decision-making process. Therefore, consequentialists might claim that the robots can ethically act, for they can make calculations and 'choose' the right action for the greatest number and the highest pleasure. But we should not forget that when the robots make calculations, these calculations will be programmed before any actions are performed. Therefore, the programmers will need to predict the possible actions and any possible outcomes. But these programmed actions will be based on the moral standpoints of their programmers. Most damagingly for this idea, however, I believe that pleasure and pain cannot be calculated – and that even if they could, the robots would have to perform impractically huge calculations in order to 'choose' the best action for the highest pleasure. In addition, there is no way of knowing beforehand which action would cause the highest pleasure for the greatest number anyway.

I then moved on to deontology, which claims that good or bad action depends on the intention or will behind the action. In order to make a moral decision, we have to use reason, we have to recognise the aim of our actions; and this requires conscious processes that robots lack. Only conscious beings can have intention and will. In deontology, the decision-making process is an important part of making moral decisions. Robots can be programmed to simulate decision-making processes; however, the decision will again be ultimately dependent on the standpoint of their programmers. So, there is no way in which

robots will be free of the programmers' moral understanding. Furthermore, according to Kant, humans should be treated as ends in themselves, and never as means to something else. In order to act towards someone as an end requires empathy skills which robots do not have, because these skills require consciousness.

Finally, I discuss virtue ethics which claims that ethically good action is related to cultivating of a good character. According to virtue ethics, if someone is acting as a virtuous person would act, then it means that s/he is acting in an ethically good way. A virtuous person is someone that develops ethically good habits. When we develop good habits, we consciously learn and practice them, however, so robots cannot be virtuous because they lack conscious experience. They can learn and improve themselves, but this is not conscious behaviour and they can never really have a virtuous character. Thus, I end this chapter by claiming that there is no reason to support the idea that robots can be ethical, and there is no reason that we should treat them as if they are moral agents.

The main idea related to ethical issues is that robots will always do whatever their programmers design them to do. When we talk about their morality, we will actually be talking about their programmers' moral understanding. Moral understandings of programmers might vary, however. As we discussed in the fourth chapter, humans do not have a single agreed moral understanding. Even if we accept that robots can be programmed by moral understandings, there would be robots with different moral understandings in the streets, houses, workplaces. This would cause chaos in society, I predict.

However, people may still believe that robots can be conscious and can act ethically and think that they can be subjects, even though all this is false. In the fifth chapter, I give examples of problems that might be caused by treating robots as if they

are moral agents/subjects. There might be many negative consequences of treating robots as if they are subjects, such as driverless cars, worker robots, sex robots and killer robots. The robots that do not have consciousness cannot be held responsible because of their 'actions'. So, when something bad happens in society, we will be confused about whom to blame and punish. A responsibility gap will arise. Moreover, people will start to lose their jobs when they are replaced by robots in factories, hospitals, care homes, houses. In fact, many people have already been replaced by robots in many areas; and this is likely to continue, increasingly. Work is essential to humans for a meaningful life. If robots take over work on a large scale, then the world will encounter limitless boredom and meaningless lives because humans would not know what to do if there is no work. I then discuss sex robots. If sex robots became common, sex would become less significant, and the meaning of love would radically change. Moving onto killer robots in the military, I argue that using killer robots will be against human dignity because humans will be seen as objects that should be targeted. (As seen with drones in Ukraine, they exacerbate the horror of war.)

Governments will need to prepare new laws before robots become more widespread in society. For, despite their problems, humans will still want to see them around. Human attitudes towards robots will be very important to shape the future. However, treating unconscious robots as if they are subjects would not only be a significant ethical mistake but would also, practically, cause chaos.

In the final chapter, I discussed the projected technological singularity. The technological singularity is a hypothetical point where robots (or other machines) will surpass humans in every respect; they will be better than humans at programming robots, even, so humans will lose control over them. When the singularity occurs, it is believed that AI will be able to genuinely think, dream and create, and although I have already

dismissed this, it is still possible that they could emulate this kind of behaviour as enacted by a superior conscious being. So in this chapter, I discuss practical issues related to the singularity and its good and bad consequences. For example, it might end poverty, it might make humans immortal, or the robots might make humanity extinct. More importantly, I discuss the philosophical concerns related to the singularity. In order for the singularity to come true the robots would have to have genuine intelligence. However, in order for the brain emulation argument to be plausible (which the projected scenario relies upon), functionalism would have to be correct – and, as seen in the third chapter, this is doubtful. I argue that human cognition cannot be emulated since it is not computable. Moreover, emulation is only simulation, and therefore an emulation itself does not have a mind or consciousness. I conclude that we should not expect a singularity.

That is what I argued in the thesis and now I would like to give my final thoughts on these matters. Consciousness in relation to robots with AI is likely to become increasingly discussed in philosophy. Improvements in AI and robotics may convince people to think that robots can have consciousness. But I remain highly sceptical, and I have given good reasons, I think, that the most promising theory, that is to say, functionalism, cannot overcome its inherent flaws.

I would like to conclude this research by quoting Searle. He writes that 'because we do not understand the brain very well we are constantly tempted to use the latest technology as a model for trying to understand it. In my childhood we were always assured that the brain was a telephone switch board. I was amused to see that Sherrington, the great British neuroscientist, thought that the brain worked like a telegraph system. Freud often compared the brain to hydraulic and electro-magnetic systems. Leibniz compared it to a mill, and I am told that some of the ancient Greeks thought the brain functions like a catapult. At present, obviously, the metaphor is the digital computer.'

(1984: 44). Maybe, in the future, some will compare it with our Galaxy, The Milky Way, because advances in astrophysics will have given us much more information about it, and it will be the latest significant knowledge in our lives, so we will feel confident to say that the human brain works like The Milky Way. Today, the analogy between catapult and brain seems to be absurd; in the future, perhaps, we will see that the comparison between the human brain and the computer is also absurd.

BIBLIOGRAPHY

Aggarwal, C. C. (2018). *Neural networks and deep learning: a textbook*. (1st ed.). Springer.

Aksenov, P. (2013, September 26). Stanislav Petrov: The man who may have saved the world. *BBC News*. Available from: https://www.bbc.co.uk/news/world-europe-24280831

Allen, C., Varner, G. and Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12 (3), pp. 251-261.

Altmann, J., Asaro, P., Sharkey, N. and Sparrow, R. (2013). Armed military robots: editorial. *Ethics and Information Technology*, 15, pp. 73-76.

Ambler, W. (1987). Aristotle on nature and politics: the case of slavery. *Political Theory*, 15 (3), pp. 390-410.

Anderson, D. L. (2006). Searle and the robot reply. *Consortium on Cognitive Science Instruction*. Available from: https://mind.ilstu.edu/curriculum/searle_chinese_room/searle_robot_reply.html

Anderson, M., Anderson, S. L. and Armen, C. (2005). Towards machine ethics: Implementing two action-based ethical theories. *Proceedings of the AAAI*. Available from: https://www.aaai.org/Papers/Symposia/Fall/2005/FS-05-06/FS05-06-001.pdf

Annas, J. (2008). The phenomenology of virtue. *Phenom Cogn Sci*, 7, pp. 21-34.

Apollonius of Rhodes (3rd century B.C.E. 1993). *Argonautica.* (R. Hunter, Trans.). Oxford University Press.

Aristotle (ca. 350 B.C.E./2009). *The Nicomachean ethics*. (D. Ross, Trans.). Oxford University Press.

- (ca. 350 B.C.E./1998). *Metaphysics*. (H. Lawson-Tancred, Trans.). (Revised ed.). Penguin Classics.

Arkin, R. C. (2007). Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture. *Technical Report* GIT-GVU-07-11.

- (2009). *Governing lethal behaviour in autonomous robots*. CRC Press.

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D. and Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *PSPB*, 23 (4), pp. 363-377.

Aron, J. (2015). Exponential growth. *New Scientist*, pp. 34-35.

Asimov, I. (1950). *I, robot*. New York: Bentam Books.

Atkinson, S. (2017, August 24). Robot priest: The future of funerals? *BBC News*. Available from: https://www.bbc.co.uk/news/av/world-asia-41033669

Ayres, R. U. (1998). *Turning point: the end of the growth paradigm*. London: Earthscan Publications.

Baker, L. (1987). *Saving belief*. Princeton: Princeton University Press.

Belpaeme, T. et al. (2018). Social robots for education: a review. *Science Robotics*, 3, pp. 1-9.

Bendel, O., Schwegler, K. and Richards, B. (2017). Towards Kant machines. In the AAAI [Symposium] on *Artificial Intelligence for the Social Good Technical Report*, pp. 7-11.

Bentham, J. (1789/2000). *An introduction to the principles of morals and legislation*. Batoche Books.

Berkeley, G. (1710/2020). *A treatise concerning the principles of human knowledge*. Independently Published.

Birnbacher, D. (2016). Are autonomous weapons systems a threat to human dignity? In N. C. Bhuta, S. Beck, R. Geib, H. Liu and C. Kreb (Eds.), *Autonomous Weapons Systems: Law, Ethics, Policy*, (pp.105-122). Cambridge University Press.

Black, B. (1986). *The abolition of work and other essays*. Loompanics Unlimited.

Block, N. J. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9, pp. 261-325.

- (1990). Inverted earth. *Philosophical Perspectives*, 4, pp. 53-79.
- (1995). On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, 18, pp. 227-287.

- (1996). What is functionalism? In D. M. Borchert (Ed.), *The Encyclopedia of Philosophy Supplement* (pp. 27-44). Mac Millan.

Block, N. J. and Fodor, J. (1972). What psychological states are not? *Philosophical Review*, 81, pp. 159-181.

Bloom, P. (2016). *Against empathy: the case for rational compassion.* Harper Collings Publishers.

Boghossian, P. (1990). The status of content. *Philosophical Review*, 99, pp. 157-184.

Bostrom, N. (2002). Existential risk: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, 9, pp. 1-36.

- (2016). *Superintelligence: Paths, dangers, strategies*. (Reprint Ed.). Oxford.

Bremmer, J. (1989). *Interpretations of Greek mythology*. Barnes & Noble Books.

Brooks, T. (2012). *Hegel's philosophy of right*. John Wiley & Sons.

Buckner, C. and Garson, J. (1997/2019). Connectionism. *Stanford Encyclopedia of Philosophy*.

Burges-Jackson, K. (2013). Taking egoism seriously. *Ethic Theory and Moral Practice*, 16, pp. 529-542.

Buxton, R. (2013). *Myths and tragedies in their ancient Greek contexts*. Oxford: Oxford University Press.

Campa, R. (2018). *Still think robots can't do your job? Essays on automation and technological unemployment*. D Editore.

Caughill, P. (2017, August 8). Germany drafts world's first ethical guidelines for self-driving cars. *Futurism*. Available from: https://futurism.com/germany-drafts-worlds-first-ethical-guidelines-for-self-driving-cars

Cave, S. and Dihal, K. (2018). The automaton chronicles. *Nature*, 559, pp. 473-475.

Cellan-Jones, R. (2014, December 2). Stephen Hawking warns artificial intelligence could end mankind. *BBC News*. Available from: https://www.bbc.co.uk/news/technology-30290540

Chalmers, D. (1996). *The conscious mind*. New York: Oxford. Oxford University Press.

- (2002). *Philosophy of mind: classical and contemporary readings*. OUP USA.
- (2010). The singularity: a philosophical analysis. *Journal of Consciousness Studies*, 17, pp. 7-65.
- (2022). *Reality+: Virtual worlds and the problems of philosophy*. Allen Lane.

Chaminade, T. et al. (2010). Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. *PloS One*, 5 (7), pp. 1-13.

Chandler, S. (2018, November 16). Unholy matrimony: Japanese man marries video-game hologram in bizarre $18000 wedding ceremony. *The Sun*. Available from: https://www.thesun.co.uk/tech/7756888/japanese-man-marries-video-game-hologram-in-bizarre-18000-wedding-ceremony/

Chisholm, R. M. (1957). *Perceiving: a philosophical study*. Ithaca, New York: Cornell University Press.

Churchland, P. M. (1981). Eliminative materialism and propositional attitudes. *Journal of Philosophy*, 78, pp. 67-90.

- (1984). *Matter and consciousness*. The MIT Press.

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, 30 (8), pp. 748-757.

Cole, D. (2004). The Chinese room argument. *Stanford Encyclopedia of Philosophy*.

Coplan, A. (2011). Will the real empathy please stand up? A case for a narrow conceptualization. *The Southern Journal of Philosophy*, 49 (1), pp. 40-65.

Cowley, J. and Hardy, H. (2021, July 28). Pensées by Bryan Magee. *New Statesman*. Available from: https://www.newstatesman.com/ideas/2021/07/pens-es-bryan-magee

Crane, T. (1995/2003). *The mechanical mind.* (2nd ed.). Routledge.

- (2001). *Elements of mind*, Oxford: Oxford University Press.

Crane, T. and Mellor, D. H. (1990). There is no question of physicalism. *Mind*, 99 (394), pp. 185-206.

Curiel, E. (2009/2019). Singularities and black holes. *Stanford Encyclopedia of Philosophy*.

Cuervo, M. J. P. (2017, December 1). Brazen heads: the curious legend behind fortune-telling automata. *Mental Floss*. Available from: https://www.mentalfloss.com/article/502537/brazen-heads-curious-legend-behind-fortune-telling-automata

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, 18, pp. 299-309.

- (2017). Will life be worth living in a world without work? Technological unemployment and the meaning of life. *Science and Engineering Ethics*, 23, pp. 41-64.
- (2019). The philosophical case for robot friendship. *Journal of Posthuman Studies*, 3 (1), pp. 5-24.
- (2019b). *Automation and utopia: Human flourishing in a world without work*. Harvard University Press.
- (2019c). Regulating child sex robots: Restriction or experimentation? *Medical Law Review*, 27 (4), pp. 553-575.
- (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics*, 26, pp. 2023-2049.

Darling, K. (2012). Extending legal rights to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In Robot Law, Calo, Froomkin, Kerr (Eds.), *We Robot Conference*, University of Miami.

Davies, M. (1998). The philosophy of mind. In A. C. Grayling (Ed.), *Philosophy 1: A Guide Through the Subject* (pp. 250-335). Oxford University Press.

Dawkins, R. (2006). *The blind watchmaker*. (1st Ed.). Penguin Books.

Dennett, D. C. (1978/1981). *Brainstorms*. (40th Anniversary Ed.). A Bradford Book: The MIT Press.

- (1984). Cognitive wheels: the frame problem of AI. In C. Hookway (Ed.), *Minds, Machines and Evolution* (pp. 129-150). Cambridge University Press.
- (1991). *Consciousness explained*. Penguin.

- (1994, September 1-3). Consciousness in human and robot minds. In ILAS *Cognition, Computation and Consciousness* [Symposium]. Kyoto. Available from: https://ase.tufts.edu/cogstud/dennett/papers/concrobt.htm
- (1996/1997). *Kinds of minds: toward an understanding of consciousness*. (Illustrated Ed.). Basic Books.
- (2017). *From bacteria to Bach and back: the evolution of minds*. Penguin.

Descartes, R. (1637/1998). *Discourse on method*. (D. A. Cress, Trans.). (4th ed.). pp. 1-44. Indianapolis/Cambridge: Hackett Publishing Company.

- (1641/1998). *Meditations on first philosophy*. (D. A. Cress, Trans.). (4th ed.). pp. 46-103. Indianapolis/Cambridge: Hackett Publishing Company.
- (1646/2017). Letter to Cavendish. In (J. Bennett Ed.) *Selected Correspondence of Descartes*, (pp.189-191). Available from: https://www.earlymoderntexts.com/assets/pdfs/descartes1619_4.pdf
- (1649/2017). Letter to More. In (J. Bennett Ed.) *Selected Correspondence of Descartes*, (pp.212-216). Available from: https://www.earlymoderntexts.com/assets/pdfs/descartes1619_4.pdf
- (1677/1985). Treatise on man. (J. Cottingham, R. Stoothoff & D. Murdoch, Trans.). In *The philosophical writings of Descartes.* Volume 1, (pp. 99-109). Cambridge: Cambridge University Press.

Descartes, R., Elisabeth and Shapiro, L. (1643/2007). *The Correspondence between Princess Elisabeth of Bohemia and Rene Descartes*. (L. Shapiro, Ed. And Trans.). Chicago: The University of Chicago Press.

Docherty, B. (2014, May 12). The human rights implications of killer robots. *Human Right Watch.* Available from: https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots#

Dodgson, L. (2018, July 20). 4 ways sex brings couples closer to each other, according to science. *Insider*. Available from: https://www.insider.com/how-sex-makes-couples-closer-2018-7

Dreyfus, H. (1972). *What computers can't do?* Harper & Row.

Dretske, F. (1981). *Knowledge and the flow of information*. Oxford: Clarendon Press.

Dugatkin, L. A. (2018). The silver fox domestication experiment. *Evolution: Education and Outreach*, 11 (16), pp. 1-5.

Earman, J. (1995). *What is singularity? Bangs crunches, whimpers, and shrieks: singularities and acausalities in relativistic spacetimes.* pp. 28-31. Oxford University Press.

Ebrahim, N. (2020, November 5). Egyptian inventor trials robot that can test for covid-19. *Reuters*. Available from: https://www.reuters.com/article/health-coronavirus-egypt-robot/egyptian-inventor-trials-robot-that-can-test-for-covid-19-idUSKBN2852F6

Essinger, J. (2004). *Jacquard's web: how a hand-loom led to the birth of the information age*. Oxford University Press.

Ewalt, D. M. (2012, November 27). 30 great moments in the history of robots. *Forbes*, Available at: https://www.forbes.com/sites/davidewalt/2012/11/27/30-great-moments-in-the-history-of-robots/ (Accessed: 13 October 2018).

Feigl, H. (1958). *The 'mental' and the 'physical'*. Minneapolis: University of Minnesota Press.

Fieser, J. (2017). Utilitarianism. Available from: https://www.utm.edu/staff/jfieser/class/300/utilitarian.htm

Filho, R. V. T. (2020). Phineas Gage's great legacy. *Dement Neuropsychology*, 14 (4), pp. 419-421.

Fleming, P. (2015). *The mythology of work: how capitalism persists despite itself*. Pluto Press.

Fodor, J. A. (1987). *Psychosemantics*: *the problem of meaning in the philosophy of mind*, Cambridge, MA: MIT Press.

- (1990). *A theory of content and other essays*, Cambridge (MA): MIT Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis, *Cognition*, 28 (1-2), pp. 3-71.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, 5, pp. 1-6.

- (1995). Does moral subjectivism rest on a mistake? *Oxford Journal of Legal Studies*, 15 (1), pp. 1-14.
- (2002). *Virtues and vices and other essays in moral philosophy*. Clarendon Press.

Ford, M. (2015). *Rise of the robots: technology and the threat of a jobless future*. New York: Basic Books.

Garber, M. (2013, September 20). Funerals for fallen robots. *The Atlantic*. Available from: https://www.theatlantic.com/technology/archive/2013/09/funerals-for-fallen-robots/279861/.

Gheaus, A. and Herzog, L. (2016). The goods of work (other than money!). *Journal of Social Philosophy*, 47 (1), pp. 70-89.

Gibbs, S. (2015, July 17). Crash involving self-driving Google car injures three employees. *The Guardian*. Available from: https://www.theguardian.com/technology/2015/jul/17/crash-self-driving-google-car-injures-three.

Gisborne, T. (1789/2015). *The principles of moral philosophy investigated and applied to the constitution of civil society.* Sagwan Press.

Godfrey-Smith, P. (2016). *Other minds: the octopus and the evolution of intelligent life*. William Collins.

Goff, P. (2019). *Galileo's error: foundations for a new science of consciousness.* Rider.

Goldman, A. (2006). *Simulating mind: the philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Good, I. J. (1964/1966). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6, pp. 31-88.

Goodman, L. E. (1992). *Avicenna.* Routledge.

Goose, S. D. and Wareham, M. (2016). The growing international movement against killer robots. *Harvard International Review*, 37 (4), pp. 28-33.

Gorvett, Z. (2018, May 31). How humans bond with robot colleagues. *BBC*. Available from: https://www.bbc.com/worklife/article/20180530-how-humans-bond-with-robot-colleagues.

Gould, S. J. (2007). *The richness of life: the essential Stephen Jay Gould*. Vintage Books, London.

Gould, S. J. and Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critic of the adaptationist programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 205 (1161), pp. 581-598.

Gowans, C. (2004/2021). Moral relativism. *Stanford Encyclopedia of Philosophy*.

Grayling, A. C. (1998). *Philosophy 1: a guide through the subject*. Oxford University Press.

Groves, M. and Mundt, K. (2015). Friend or foe? Google translate in language for academic purposes. *English for Specific Purposes*, 37, pp. 112-121.

Haikonen, P. O. (2003). *The cognitive approach to conscious machines*. Imprint Academic.

Halifax, R. J. (2018, February 25). Discovery at the edge of empathy. *Upaya*. Available from: https://www.upaya.org/2018/02/discovery-at-the-edge-of-empathy-by-roshi-joan-halifax/.

Harari, Y. N. (2017, February 24). The rise of the useless class. *Ideas Ted*. Available from: https://ideas.ted.com/the-rise-of-the-useless-class/.

Harnad, S. (1989). Mind, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, 1 (1), pp. 5-25.

Harris, W. (2021). Who invented the computer? *How Stuff Works?* Available from: https://science.howstuffworks.com/innovation/inventions/who-invented-the-computer.htm (Accessed on: 22 March 2022).

Hauser, L. (2006). Searle's Chinese room. *Internet Encyclopedia of Philosophy*. Available from: https://iep.utm.edu/chineser/.

Heal, J. (2003). *Mind, reason and imagination: selected essays in philosophy of mind and language*. Cambridge University Press.

Hern, A. (2017). Give robots 'personhood' status, EU committee argues. *The Guardian*. Available from: https://www.theguardian.com/technology/2017/jan/12/give-robots-personhood-status-eu-committee-argues.

Heyns, C. (2013, April 9). *Report of the special rapporteur on extrajudicial, summary or arbitrary executions*. UN Human Rights Council.

- (2017). Autonomous weapons in armed conflict and the right to a dignified life: an African perspective. *South African Journal on Human Rights*, 33 (1), pp. 46-71.

Hockstein, N. G., Gourin, C. G., Faust, R. A. and Terris, D. J. (2007). A history of robots: from science function to surgical robotics. *Journal of Robotic Surgery*. 1, 113-118. DOI 10.1007/s11701-007-0021-2.

Homer (c. 8th B.C./1990). *The Iliad*. (R. Fagles, Trans.). Penguin Classics.

- (c. 8th B.C./2017). *The Odyssey*. (E. Watson, Trans.). W. W. Norton & Co.

Howell, E. (2015, January 19). How long have humans been on earth? *Universe Today: Space and Astronomy News.*

Hursthouse, R. (1997). Virtue theory and abortion. In R. F. Chadwick & D. Schroeder (Eds.), *Applied Ethics: Critical concepts in philosophy*, (pp. 112-129).

- (1999/2002). *On virtue ethics*. (Revised ed.). Oxford University Press.

Huxley, T. (1874/2007). On the hypothesis that animals are automata, and its history. *The Fortnightly Review*, 16 (95), pp. 555-580.

Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32 (127), pp. 127-136.

Kang, M. (2011). *Sublime dreams of living machines, the automaton in the European imagination*. Cambridge: Harvard University Press.

Kant, I. (1785/1998). *Groundwork of the metaphysics of morals*. (M. Gregor, Trans & Ed.). Cambridge University Press.

- (1793/1998). *Religion within the boundaries of mere reason and other writings*. (A. Wood and G. di Giovanni, Eds. and Trans.). Cambridge University Press.
- (1797/1996). *The metaphysics of morals*. (M. Gregor, Trans & Ed.). Cambridge University Press.

Kapeliushnikov, R. (2019). The phantom of technological unemployment. *Russian Journal of Economics*, 5, pp. 88-116.

Katz, E. R. (2017, May 12). Sophia the robot has been given the perfect drag queen makeover thanks to legendary queen. *MIC*. Available from: https://www.mic.com/articles/186553/sophia-the-robot-has-been-given-the-perfect-drag-queen-makeover-thanks-to-legendary-queen-aquaria.

Kelion, L. (2014, June 18). African firm is selling pepper-spray bullet firing drones. *BBC News*. Available from: https://www.bbc.co.uk/news/technology-27902634.

Kelly, J. (2020, October 27). U.S. Lost over 60 million jobs-Now robots, tech and artificial intelligence will take millions more. *Forbes*. Available from: https://www.forbes.com/sites/jackkelly/2020/10/27/us-lost-over-60-million-jobs-now-robots-tech-and-artificial-intelligence-will-take-millions-more/.

Keynes, J. M. (1930/1932). Economic possibilities for our grandchildren. In *Persuasion* (pp. 358-373). New York: Harcourt Brace.

Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT Press.

- (2011). *Philosophy of mind*. Westview Press.

Kirk, R. (1974). Zombies v. materialists. *Proceedings of the Aristotelian Society*, 48, pp. 135-152.

- (1999). Why there couldn't be zombies. *Proceedings of the Aristotelian Society*, 73, pp. 1-16.
- (2003). *Mind and body*. Chesham: Acumen.

Korsgaard, C. M. (1998). Introduction. In I. Kant, *Groundwork of the metaphysics of morals*. (M. Gregor, Trans & Ed.), (pp. vii-xxx). Cambridge University Press.

Kozima, H., Nakagawa, C. and Yano, H. (2004). Can a robot empathize with people? *Artif Life Robotics,* 8, pp. 83-88.

Krugman, P. (2013, June 13). Sympathy for the Luddites. *The New Yok Times*. Available from: https://www.nytimes.com/2013/06/14/opinion/krugman-sympathy-for-the-luddites.html.

Kurzweil, R. (2005). *The singularity is near.* Duckworth.

- (2014). *How to create a mind*. Duckworth.

La Mettrie, J. O. (1747/1996). Machine man. (A. Thomson, Ed. & Trans.) In *Machine man and other writings* (pp. 1-39). Cambridge University Press.

Lallement, J. H., Kuss, K., Trautner, P., Weber, B., Falk, A. and Fliessbach, K. (2014). Effort increases sensitivity to reward and loss magnitude in the human brain. *Scan*, 9, pp. 342-349.

Lamb, G. M. (2005, October 13). Progress at light speed. *The Cristian Science Monitor*. Available from: https://www.csmonitor.com/2005/1013/p14s01-stgn.html.

Levin, J. (2004/2018). Functionalism. *Stanford Encyclopedia of Philosophy*.

Levin, S. and Wong, J. C. (2018, March 19). Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. *The Guardian*. Available from: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

Levy, D. (2007). *The evolution of human-robot relationships: love+Sex with robots*. Harper Collins.

Lewis, D. (1966). An argument for the identity theory. *The Journal of Philosophy, 63* (1), pp. 17-25.

- (1970). How to define theoretical terms. *The Journal of Philosophy,* 67 (13), pp. 427-446.
- (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50 (3), pp. 249-258.
- (1980). Mad pain and Martian pain. (N. Block, Ed.). In *Readings in Philosophy of Psychology Volume 1,* (pp. 216-232). Cambridge, MA: Harvard University Press.

Lin, P. (2015). The ethical dilemma of self-driving cars. *TED Talks*.

Littlewood, A., Maguire, H. and Wolschke-Bulmahn, J. (Eds.). (2002). *Byzantine Garden culture*. Washington DC: Dumbarton Oaks Research Library and Collection.

Loemker, L. E. (1989). *Gottfried Wilhelm Leibniz: Philosophical Papers and Letters*. 2nd ed. Kluwer Academic Publisher.

Loon, R. J. and Martens, M. H. (2015). Automated driving and its effect on the safety ecosystem: how do compatibility issues affect the transition period? *Procedia Manufacturing*, 3, pp. 3280-3285.

Lycan, W. G. (1981). Form, function, and feel. *The Journal of Philosophy*, 78 (1), pp. 24-50.

Maskivker, J. (2011). Employment as a limitation on self-ownership. *Human Right Review*, 12, pp. 27-45.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6, pp. 175-183.

Mayor, A. (2018). *Gods and robots: myths, machines and ancient dreams of technology*. Oxford: Princeton University Press Princeton.

McCarthy, J. (1979). *Ascribing mental qualities to machines*. Available from: http://www-formal.stanford.edu/jmc/

McClelland, J. L. and Cleeremans, A. (2009). Connectionist model, In T. Byrne, A. Cleeremans and P. Wilken (Eds.), *Oxford Companion to Consciousness*, New York: Oxford University Press.

McGinn, C. (1987). Wittgenstein on meaning: An interpretation and evaluation. *Behaviourism*, 15 (1), pp. 66-72.

Meltzer, A. L., Makhanova, A., Hicks, L. L., French, J. E., McNulty, J. K. and Bradbury, T. N. (2017). Quantifying the sextual afterglow: the lingering benefits of sex and their implications for pair-bonded relationships. *Psychological Science*, 28 (5), pp. 587-598.

Mill, J. S. (1863/2002). *Utilitarianism*. (G. Sher, Ed.). (2nd Ed.). Hackett Publishing Co.

Millikan, R. G. A. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, Mass MIT Press.

Mitchell, W. C. (1918). Bentham's felicific calculus. *Political Science Quarterly*, 33 (2), pp. 161-183.

Mitchell, D. and Jackson, F. (1996). *Philosophy of mind and cognition*. Blackwell Publishing.

Muller, O. (2020). An eye turned into a weapon: A philosophical investigation of remote controlled, automated, and autonomous drone warfare. *Philosophy & Technology*. DOI: 10.1007/s13347-020-00440-5.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83 (4), pp. 435-450.

Nagesh, A. (2016, May 9). Apparently we're all going to live forever by 2029. *Kurzweil: Tracking the Acceleration of Intelligence*. Available from: https://www.kurzweilai.net/metro-apparently-were-all-going-to-live-forever-by-2029.

Nesbitt, E. (2004). *Intercultural education: ethnographic and religious approaches*. Sussex Academic Press.

Newborn, M. (1997). *Kasparov versus deep blue: computer chess comes of age*. (1st Ed.). Springer.

Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19 (3), pp. 113-126.

Newell, A., Shaw, J. C. and Simon, H. A. (1959). A general problem-solving program for a computer. *The International Conference on Information Processing*. Paris, France.

Nof, S. Y. (1999). *Handbook of Industrial Robotics*. (2nd Ed.). Wiley.

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.

Nyholm, S. and Frank, L. (2018). From sex robots to love robots: is mutual love with a robot possible? In J. Danaher and N. McArthur (Eds.), *Robot Sex: Social and Ethical Implications*, (pp. 219-245). The MIT Press.

O'Neill, O. (2000). *Bounds of justice*. Cambridge University Press.

Parke, P. (2015, February 13). Is it cruel to kick a robot dog? *CNN Business*. Available from: https://edition.cnn.com/2015/02/13/tech/spot-robot-dog-google/index.html.

Parthemore, J. and Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness*, 5 (2), pp. 105-129.

Penrose, R. (1994). *Shadows of the mind: a search for the missing science of consciousness*. Oxford University Press.

Pettit, P. (2015). *The robust demands of the good: ethics with attachment, virtue and respect*. Oxford University Press.

Piper, K. (2019, June 21). Death by algorithm: the age of killer robots is closer than you think. *Vox*. Available from: https://www.vox.com/2019/6/21/18691459/killer-robots-lethal-autonomous-weapons-ai-war.

Pistono, F. (2012). *Robots will steal your job but that's ok: how to survive the economic collapse and be happy*. CreateSpace Independent Publishing Platform.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Philosophy*, 47 (1), pp. 44-50.

Popper, K. P. (1957). *The poverty of historicism*. Boston the Beacon Press.

Powers, T. M. (2009). Machines and moral reasoning. *Philosophy Now*, 72, pp. 15-16.

Purves, D., Jenkins, R. and Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethic Theory Moral Prac*, 18, pp. 851-872.

Putnam, H. (1960). Minds and machines. (S. Hook, Ed.). *Dimensions of Mind: A Symposium*, pp. 20-33. New York University Press.

- (1967/2008). The nature of mental states. In *Mind, Language and Reality* (pp. 429-440). (Subsequent ed.). Cambridge University Press.
- (1968/2008). Brains and behaviour. (D. M. Rosenthal, Ed.). In *Mind, Language and Reality.* (pp. 325-342). (Subsequent ed.). Cambridge University Press.
- (1975/2008). Philosophy and our mental life. In *Mind, Language and Reality*. (pp. 291-304). (Subsequent ed.). Cambridge University Press.
- (1991). *Representation and reality*. Cambridge: Mass MIT Press.

Ravenscroft, I. (1998). What it is like to be someone else? Simulation and empathy. *Ratio (new series),* 11, pp. 170-185. Blackwell Publishers.

Reimer, B. (1984). Farm mechanization: the impact on labour at the level of the farm household. *The Canadian Journal of Sociology*, 9 (4), pp. 429-443.

Reppert, V. (1992). Eliminative materialism, cognitive suicide, begging the question. *Metaphilosophy*, 23, pp. 378-392.

Richardson, K. (2015). The asymmetrical 'relationship': parallel between prostitution and the development of sex robots. *Sigcas Computers & Society*, 45 (3), pp. 290-293.

Rifkin, J. (1995). *The end of work: The decline of the global labor force and the dawn of the post-market era.* New York: Putnam Publishing Group.

Robinson, W. (1999/2019). Epiphenomenalism, *Stanford Encyclopedia of Philosophy*.

Rolls, E. T. and Treves, A. (1997). *Neural networks and brain function*. Oxford University Press.

Russell, B. (1935/2004). *In praise of idleness*. Routledge.

Ryle, G. (1949/2000). *The concept of mind*. Penguin Classics.

Sandberg, A. and Bostrom, N. (2008). *Whole brain emulation: a roadmap*. Technical Report, Future of Humanity Institute. Oxford University.

Sartre, J. P. (1943/2018). *Being and nothingness: an essay in phenomenological ontology*. (S. Richmond, Trans.). Washington Square Press/Atria.

Scharre, P. (2018). *Army of none: autonomous weapons and the future of war.* W. W. Norton & Company.

Schlosser, M. (2015/2019). Agency. *Stanford Encyclopedia of Philosophy*.

Schneiderman, I., Zagoory-Sharon, O., Leckman, J. F. and Feldman, R. (2011). Oxytocin during the initial stage of romantic attachment: Relations to couple's interactive reciprocity. *Psychoneuroendoctinology,* 37, pp. 1277-1285.

Schmaltz, T. M. (1992). Descartes and Malebranche on mind and mind-body union. *The Philosophical Review,* 101 (2), pp. 281-325.

Schwarz, E. (2022). Delegating moral responsibility in war: lethal autonomous weapons systems and the responsibility gap. In H. Hannes-Magnusson and A. Vetterlein (Eds.),

The Routledge Handbook on Responsibility in International Relations (pp. 177-191). Routledge.

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3 (3), pp. 417-457.

- (1983/2008). *Intentionality*. Cambridge: Cambridge University Press.
- (1984). *Minds, brains and science*. Penguin Group.
- (1989). Artificial intelligence and the Chinese room: an exchange. *New York Review of Books*, 36 (2).
- (1990). Is the brain's mind a computer program*? Scientific American Journal*, 262, pp. 26-31.
- (1992). *The rediscovery of the mind*. Cambridge: The MIT Press.
- (1993). The problem of consciousness, Social Research, 60 (1), pp. 3-16.
- (1998). *Mind, language and society*. Basic Books.
- (2002). Why I am not a property dualist. Journal of Consciousness Studies, 9 (12), pp. 57-64.
- (2018). Status function. In M. Jankovic & K. Ludwig (Eds.), *The Routledge Handbook of Collective Intentionality* (pp. 300-309). Routledge.

Shaffer, J. (1961). Could mental states be brain processes? *The Journal of Philosophy*, 58 (26), pp. 813-822.

Shanahan, M. (2015). Ascribing consciousness to artificial intelligence. *ArXiv*, abs/1504.05696.

Sharkey, A. J. C. and Sharkey, N. (2019). Connectionism, In *The Routledge companion to philosophy of psychology*, (2nd ed.). Routledge.

Sharkey, N. (2012, June 21). Alan Turing: the experiment that shaped artificial intelligence. *BBC News*. Available from: https://www.bbc.co.uk/news/technology-18475646.

- (2015, June 21). Alan Turing: The experiment that shaped artificial intelligence. *BBC News*. Available from: https://www.bbc.co.uk/news/technology-18475646.

Sharkey, N., Wynsberghe, A., Robbins, S. and Hancock, E. (2017). Our sexual future with robots. *Foundation for Responsible Robotics.* The Hague, Netherlands.

Shumaker, R. W., Walkup, K. R. and Beck, B. B. (2011). *Animal tool behavior: the use and manufacture of tools by animals*. Johns Hopkins University Press.

Singer, P. W. (1983). *Hegel: A very short introduction*. Oxford University Press.

- (2011). *Wired for war: the robotics revolution and conflict in the twenty-first century*. The Penguin Press.

Sinnott-Armstrong, W. (2003/2019). Consequentialism. *Stanford Encyclopedia of Philosophy*.

Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, 68 (2), pp. 141-156.

- (1965). Philosophy and scientific realism. *British Journal for the Philosophy of Science*, 15 (60), pp. 358-360.

Smids, J. (2020). Danaher's ethical behaviourism: an adequate guide to assessing the moral status of a robot? *Science and Engineering Ethics*, 26, pp. 2849-2866.

Sparrow, R. (2016). Kicking a robot dog. In *Proceeding of the 11th ACM/IEEE International Conference in Human-Robot Interaction*, IEEE, Christchurch, pp. 229-229.

Squires, E. J. (1990). *Conscious mind in the physical world*. (1st Ed.). CRC Press.

Strawson, G. (1994). *Mental reality*. The MIT Press.

- (2018). *Things that bother me: death, freedom, the self, etc.* The New York Review of Books.

Sullins, J. P. (2012). Robots, love, and sex: the ethics of building a love machine. *IEEE Transaction on Affective Computing*, 3 (4), pp. 398-409.

Surden, H. and Williams, M. (2016). Technological opacity, predictability, and self-driving cars. *Cardozo L. Rev*., pp. 121-181.

Suzuki, Y., Galli, A., Itakura, S. and Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, 5, pp. 1-9.

Tabarrok, A. (2003, December 31). Productivity and unemployment. *Marginal Revolution*. Available from: https://marginalrevolution.com/marginalrevolution/2003/12/productivity_an.html.

Tallis, R. (1997). *Enemies of hope: A critique of contemporary pessimism*. (1st Ed.). Palgrave.

- (2016). *Aping mankind*. Routledge.

Tartaglia, J. (2020). *Philosophy in a technological world: Gods and titans*. Bloomsbury Academic.

Taylor, L. (2017, July 20). Sex robots: Perverted or practical in fight against sex trafficking? *Thomson Reuters Foundation*. Available from: https://news.trust.org/item/20170720040410-iyta2/.

Todes, D. P. (2014). *Ivan Pavlov: a Russian life in science*. Oup USA.

Tortoreto, A. (2022). Intentionality and dualism: does the idea that intentionality is the mom necessarily entail dualism? *Phenomenology and Mind*, 22, pp. 83-91.

Truitt, E. R. (2015). *Medieval robots: mechanism, magic, nature, and art*. Philadelphia: University of Pennsylvania Press.

Turing, A. (1936). On computable number, with an application to the entscheidungsproblem. *Proceeding of the London Mathematical Society*, 2 (42), pp. 230-265.

- (1950). Computing machinery and intelligence. *Mind*, 59 (236), pp. 433-460.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies*, 8 (3), pp. 501-517.

- (2015). *Reclaiming conversation: The power of talk in a digital age*. New York: Penguin Press.

Twomey, S. (2010). Phineas Gage: neuroscience's most famous patient. *Smithsonian Magazine*. Available from: https://www.smithsonianmag.com/history/phineas-gage-neurosciences-most-famous-patient-11390067/.

Ukiwe, U. (2018, August 8). Yanomami tribe bury their dead by eating their flesh. *Life*. Available from: https://guardian.ng/life/yanomami-tribe-bury-their-dead-by-eating-their-flesh/.

Ulam, S. (1958). John von Neumann 1903-1957. *Bulletin of the American Mathematical Society*, 64, pp. 1-49.

Ulgen, O. (2016). Human dignity in an age of autonomous weapons: Are we in danger of losing an 'elementary consideration of humanity'? *ESIL Annual Conference*, Riga, 8-10 September 2016.

- (2017). Kantian ethics in the age of artificial intelligence and robotics. *QIL, Zoom-in*, 43, pp. 59-83.

Urmson, C. (2015). How a driverless car sees the road. *TED*. Available from: https://www.ted.com/talks/chris_urmson_how_a_driverless_car_sees_the_road.

Usher, A. P. (1929). *A history of mechanical inventions*. (1st ed.). McGraw Hill Book Company. New York.

Verschuuren, G. M. N. (2012). *What makes you tick? A new paradigm for neuroscience*. Solas Press.

Vinge, V. (1983). First word. (E. Datlow and D. Teresi Eds.). *Omni*.

- (1993). The coming technological singularity: how to survive in the post-human era. *Whole Earth Review*.

Višňovský, E. (2015). Homo biotechnologicus. *Human Affairs*, 25, pp. 230-237.

Wakabayashi, D. (2010, August 31). Only in Japan, real men go to a: hotel with virtual girlfriends. *The Wall Street Journal*. Available from: https://www.wsj.com/articles/SB10001424052748703632304575451414209658940.

Wallach, W. and Allen, C. (2009). *Moral machines: teaching robots right from wrong*. Oxford University Press.

Watson, J. B. (1924/2017). *Behaviourism*. Routledge.

Weber, G. (2015, November 5). Study: Humans feel empathy for robots experiencing 'pain'. *Slate*. Available from: https://slate.com/technology/2015/11/study-shows-humans-feel-empathy-for-robots-in-pain.html.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M. and Bargh, J. A. (1998). Control and automaticity in social life. In, (Eds.) D. Gilbert, S. Fiske, G. Lindzey, (4th ed.), *Handbook of Social Psychology,* pp. 446-496. New York: McGraw-Hill.

Wheelwright, T. (2022, January 24). 2022 Cell phone usage statistics: how obsessed are we? *Reviews Org*.

Wilde, R. (2019). The first computer. *ThoughtCO*. Available from: https://www.thoughtco.com/first-computer-charles-babbages-1221836.

Young, E. (2017, August 9). Researchers are studying psychopathic chimps to better understand the human variety. *Research Digest*. Available from: https://digest.bps.org.uk/2017/08/09/researchers-are-studying-psychopathic-chimps-to-better-understand-the-human-variety/.

Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power.* Generic.