A Model of Autonomy for Artificial Agents

Master Thesis

Jurjen Idskes
s1625071

University of Twente, Faculty of Behavioural, Management, and Social Sciences
Enschede, the Netherlands

Supervisor: David Douglas
Second Reader: Philip Brey

Programme: MSc Philosophy of Science, Technology and Society (PSTS)

Table of Contents:

Introduction

An increasing number of tasks and responsibilities are being delegated to artificial agents. In areas such as healthcare, traffic, the household and the military, artificial agents are being adopted and assigned to execute tasks that were previously performed by human beings. Artificial agents are deployed in an open, uncontrolled, real-world environment and actions they perform directly affect human beings. Researchers in computer science are attempting to construct artificial intelligences that can execute increasingly complicated tasks. Current discussions underline the importance and relevance of the topic, such as: discussions regarding military robots (Hellström, 2013)(Noorman & Johnson, 2014), artificial agents in medical care (Van Wynsberghe, 2013) and autonomous cars (Hevelke & Nida-Rümelin, 2015). Recently, the first fatal car crash caused by an autonomously driving car led to discussions regarding safety and questions of responsibility when it comes to autonomous cars (The Guradian, 2016).

Because artificial agents are operating with more responsibilities and an increased level of autonomy, they are required to be able to make important decisions in situations that involve certain level of unfamiliarity and uncertainty (Wardziński, 2006). These artificial agents are deployed in an open and uncontrolled environment. In this thesis, open and uncontrolled environments refer to real world situations in which artificial agents are expected to perform their tasks. Therefore, they have to be able to operate when circumstances require it to make decisions in a situation outside of its intended purpose. Uncertainty occurs when an artificial agent is in a situation where it has to make a decision with incomplete knowledge about the environment and the consequences of their actions. I will give two examples to illustrate the importance of correct decision making in these situations.

The first example is an autonomous car in the dangerous situation of a possible accident. It has to be able to make an appropriate decision to avoid an accident or at least limit the negative consequences of a collision. In order to make justifiable decisions, the autonomous car needs to be able to anticipate the actions of the other traffic users involved. Additionally, the autonomous car needs to anticipate the consequences of its own actions. It is required to make rational and morally justifiable decisions: rational because it should anticipate potentially dangerous situations and attempt to avoid them, and moral as it must decide between prioritizing the safety of passengers in the car and human beings outside of it.

The second example is a military artificial agent that is deployed to guard a certain area. The area requires protection to the extent that harmful trespassers need to be incapacitated. In the event of an approaching human being, the agent needs to anticipate the threat this individual might pose. If the individual has malicious intentions, he or she might endanger the lives of other human beings. The artificial agent is required to make a decision on how to act. The first option is that the artificial

agent identifies the trespasser as dangerous and prevents the trespasser from going further with potentially lethal outcomes. The second option is to interpret the situation in such a way that it believes the trespasser to be harmless and not take rigorous measures. In this example, the agent has to make an estimation about the possible dangers and decide on how to act based on this estimation. The selected response to the situation is based on anticipation. Because the consequences of making a wrong judgement are severe, the agent has to be able to make reliable judgements.

Both of these examples underline the problematic situations that artificial agents with a high level of autonomy and responsibility can end up in. A high level of anticipation is required in order to adequately deal with uncertainty; for instance, to make correct judgements in situations that are unfamiliar to the agent or when decisions have to be made based on incomplete information. The literature in the fields of philosophy and computer sciences have linked anticipation with autonomy (Collier, 2006). Autonomy is required in order to make justifiable decisions in an open and uncontrolled environment. In this thesis, artificial agents refer to embodied agents that operate in open and uncontrolled environments.

Because artificial agents operate autonomously in society, they should be able to operate safely and make justifiable decisions. As I will argue in this thesis, artificial agents do not need to determine their own goals. Instead, their goals are determined by designers and users. Therefore, their autonomy should allow them to fulfil these goals by autonomous action. Designers of artificial agents have the moral responsibility to implement artificial agents in society that have a positive contribution to society. Therefore, artificial agents need to be able to make justifiable decisions to reach their pre-determined goals. In an open and uncontrolled environment, there are many unpredictable factors that should be taken into account in order to make justifiable decisions. These factors include other agents and their actions, the outcomes of certain actions (including those of the artificial agent itself) and parts of the environment that are impossible to observe at the time interval in which the decision needs to be made. To make justifiable decisions autonomously under these conditions, it is important to have a solid representation of the environment. From these representations, multiple future states can be anticipated. These anticipations are based on actions by the artificial agents itself, actions by other agents, the possible outcomes of actions by the artificial agents and other agents and the occurrence of unpredictable events. This allows the selection of the most beneficial future state and selecting actions that will result in this most beneficial future state. I will expand further on the relationship between autonomy, decision making, representation and anticipation in chapter three. I will discuss two different aspects of morality in this thesis. First, I will discuss the moral responsibility that designers of artificial agents need to fulfil in order to implement artificial agents in society. Moral codes of conduct for designers

include that their new developments should have a positive contribution to society. This will result in certain requirements for artificial agents in their autonomous functioning. I will discuss this in the first chapter. The second aspect of morality is the moral autonomy and decision making by the artificial agent itself. Since artificial agents operate in an open and uncontrolled environment which include human beings, their actions have moral consequences. This requires an appropriate model of autonomy that allows the agent to make justifiable moral decisions.

Problem Statement

The increase in the responsibilities of artificial agents should come conjointly with an increase in the reliability of judgements by artificial agents. Therefore, they should possess reliable decision making capabilities. The decision making processes should be in accordance of what is expected of an artificial agent operating in a certain function. Several theories strongly associate judgement making processes and anticipation with autonomy (Dubois, 2003) (Davidsson, Astor, Ekdahl, 1994, p. 1427). In order to gain an insight on what degree of judgement making in artificial agents is necessary to justify their employment, it is therefore required to analyse the concept of autonomy. In this thesis, I will argue that reliable judgements of artificial agents in open and uncontrolled environments depends on anticipation. Straightforward input-output mechanisms are insufficient for decision making in situations that require anticipation. Instead, internal representations of the current state of the environment, such as entities and processes, and about the self are formed. Based on the internal representation of input, future states can be anticipated. Based on this anticipation, the most likely output for reaching a predetermined goal is selected. This type of input-output processing is strongly linked with a certain degree of autonomy. Furthermore, in order to have solid and reliable input-output processing, representation of the environment is crucial. Representations allow artificial agents to use information about the environment. I will address these topics in my sub-questions.

Research Question:

The main research question of my thesis is:

What is an adequate model of autonomy for artificial agents that fulfils the designer's moral responsibility?

In order to answer this question, I have divided the main problem in five different sub-questions.

- Q 1.1: What is the moral responsibility of designers of artificial agents?

- Q 1.2: What aspects of autonomy are required for artificial agents?
- Q 1.3: How can the anticipation of future states contribute to autonomous decision making?
- Q 1.4: How does knowledge representation influence the autonomous decision making of artificial agents?
- Q 1.5: What is an adequate model for moral autonomy in artificial agents?

Chapter Overview

In the first chapter of my thesis, I will address the first sub-question of this thesis. I will discuss why there are minimal requirements for artificial agents in order to fulfil the moral responsibility of the designers. In this chapter, I will describe why it is important that artificial agents should have capabilities that live up to certain standards. Designers have a moral responsibility when they aim to implement new innovations in society. They should strive to avoid causing harm to society and instead attempt to make society benefit from their innovations. Society will benefit from artificial agents if they are able to make justifiable judgements. The aim of this chapter is to provide insight on the necessity of minimal requirements for autonomous decision making in artificial agents that operate in an open and uncontrolled environment.

In the second chapter I will address what the role of autonomy is in artificial agents. I will argue that there is not one type of autonomy and that aspects of autonomy can exist in different gradations. I will make a distinction between three different types of autonomy: personal, constitutional and behavioural. The role of an agent is what is most important in determining what the requirements for autonomy are. Autonomy has a different role in artificial agents compared to biological organisms. Artificial agents are not responsible for their own survival or for setting their own goals. They have to execute tasks and reach goals that are determined by human beings. Their model of autonomy should originate from the need to have justifiable decision making capabilities in order to execute these tasks and to reach these goals. These capabilities are part of behavioural autonomy. This chapter will provide insight about what aspects of autonomy are required in artificial agents and what aspects of autonomy are not necessary or even undesirable in artificial agents.

The third chapter will describe some of the necessary capabilities an artificial agent should posses in order to meet the requirements of autonomy described in chapter two. This chapter addresses the third and fourth sub-question of this thesis. I will describe why anticipation and representation are important for behavioural autonomy and justifiable decision making. The

anticipation of multiple future states is crucial for making justifiable decisions in open and uncontrolled environments. The most desirable of these future states should be selected and the artificial agents should plan its actions based on reaching that future state. In order to anticipate future states, an artificial agents needs to have a solid representation of the current state. I will discuss the requirements for reliable representation in an open and uncontrolled environment. The aim of this chapter is to provide an understanding of the relationship between behavioural autonomy, anticipation and representation. I will provide requirements for anticipation and representation of artificial agents in order to have justifiable, autonomous decision making in an open and uncontrolled environment.

In the fourth chapter of my thesis I will address moral decision making of artificial agents. Because the artificial agents discussed in this thesis are operating autonomously in an open and uncontrolled environment, their actions have moral implications. Therefore, they are required to make morally justifiable decisions. I will discuss different approaches with different levels of moral autonomy. There are two main approaches that can be identified to accomplish moral decision making in artificial agents. The first is to have a top-down approach, which means that an ethical framework for decision making is directly implemented in an artificial agent. The second option, a bottom-up approach, will allow artificial agents to develop their own ethical framework. I will discuss the advantages and disadvantages of both approaches and conclude that a hybrid approach is most suitable for moral decision making of artificial agents.

The last chapter is a concluding chapter in which I will answer the main research question. I will describe a model of autonomy based on the conclusions of the answers of the sub-questions.

Chapter 1: The Moral Responsibility of Designers of Artificial Agents

Introduction

In this chapter, I will address the standards of autonomous decision making that artificial agents should possess in order to fulfil the moral responsibilities of designers when they implement artificial agents in society. The responsibility addressed in this thesis refers to the responsibility of scientists, designers and engineers for the societal impacts of their developed products (Hersh, 2014); in this case artificial agents (ACM, 1992). In order to construct a model of autonomy for artificial agents, it is important to have insight on the requirements based on the moral responsibility designers have when they implement their products in society.

Being morally responsible means being worthy of a moral judgement after performing certain actions or making certain decisions (Eshleman, 2014). These reactions can be both positive and negative. For example, someone can be regarded as praiseworthy after saving another person from harm or can be regarded as blameworthy after causing events that lead to harming others. That a person can be held moral responsible means that he or she has certain moral obligations. These obligations arise from the ability human begins have to choose their own actions, and can therefore be considered responsible for those actions. The ability to plan actions, reflect on their situation, and have intentionality (to deliberately perform actions in pursuit of a certain goal) means that human beings are moral agents. Being capable of intentionality does not mean that a moral agent needs to have intended a certain outcome from an action for him or her to be held morally responsible for that outcome. Moral agents are morally responsible for any outcome when there is a causal relationship between their actions and that specific outcome (Forge, 2000).

Responsibility and Moral Agency

Certain aspects of autonomous artificial agents make questions of responsibility difficult. Responsibility for one's own actions is relatively straightforward. It becomes more complex to attribute responsibility to designers when their products are used in unexpected and unintended ways. Things are further complicated for the responsibility of the actions of artificial agents, because the product itself can be the cause of unexpected and unintended outcomes. While they have a designer, who has a certain responsibility, the very goal of artificial agents is to let them operate as autonomously as possible (Grodzinsky, Miller & Wolf, 2008). This raises question regarding the allocation of responsibility of actions by artificial agents to either the agent itself or its designers (Franklin & Graesser, 1996). Making a distinction between different levels of agency can contribute to determining responsibility (Floridi & Sanders, 2004). An important aspect in these

discussions is whether moral agency can be ascribed to an artificial agent (Sullins, 2006)(Johnson, 2006). These discussions are related to the topic of this thesis, however, the goal of this chapter is not to determine to what extent designers of artificial agents are responsible for the actions of their creations. I do not aim to clarify how responsibility is allocated between artificial agents and their designers. I predominantly aim to discuss requirements for autonomy in artificial agents that will allow them to make justifiable decisions autonomously. Meeting these requirements should result in a safe implementation of artificial agents into society. For this discussion, the outcome of the functioning of artificial agents is the most important aspect for determining whether the moral responsibility of designers is met. Therefore, the aim of this chapter is to provide constraints on the requirements of autonomous decision making of artificial agents. These constraints should provide insight on how designers of artificial agents can make a positive contribution to society. This will not answer the questions of moral agency. However, setting standards for artificial agents can contribute in determining whether the implementation of artificial agents into society is justifiable.

Moral Responsibility of Designers

Ethical codes of conduct in research emphasizes the importance of non-maleficence and beneficence (Shamoo & Resnik, 2009). These principles originate from medical ethics, but have been implemented in the ethics of scientific research as well. The principle of non-maleficence states that researchers should not cause unnecessary harm with their research. The principle of beneficence states that researchers should strive to do good with their research (Gillon, 1994). These two principles can be extended to the impacts new innovations have on society (Benčin, Strle & Gurzawska, 2015). Individuals that take part in the development and implementation of new innovations should strive to make positive contributions and avoid causing harm to society. In the case of developing and implementing new products in society, the principle of non-maleficence entails that new products should not cause harm to society, or at least the intention of its designers should not be to cause harm. The principle of beneficence entails that designers should intend to make a positive contribution to society with his or her products. Developments of ethical theories have emphasized the importance of corporate social responsibility as well. These theories have focussed on ethical requirements for corporations to make a positive contribution to society (Carriga & Melé, 2004).

Allocating moral responsibility to designers becomes difficult when the products they developed are not used for their intended purpose. This concept is commonly referred to as *dual use* (Forge, 2010). Dual use means that an artefact has a primary intended purpose and a secondary unintended purpose. The intended purpose of a product is generally beneficial for society, while the

unintended purpose is an undesirable use of a product because it is harmful to human beings or society. The artificial agents as discussed in this thesis are capable of performing complex tasks autonomously This makes them particularly dangerous when used for malicious purposes. When the goals of these artificial agents are altered, they may become a threat to society, safety and even human lives. This is especially true for artificial agents that are designed for combat, such as robots that are deployed in warfare. It is very complex for designers to make dual use impossible for any sort of artefact. This is even more complex when these artefacts are artificial agents that are designed to be able to execute a wide variety of tasks and have advanced capabilities. It is therefore crucial that designers strive to make it as difficult as possible to alter the goals of artificial agents. However, it is naïve to think that not a single hacker will attempt and succeed in doing this, either with malicious intentions or just out of general curiosity. It is undesirable that anyone would change the programming of an artificial agent because it becomes almost impossible to hold artificial agents to a certain standard if anyone can tinker with their goals and programming. For example, consider someone who alters the programming of an autonomous vehicle because he or she thinks it will function better. This is not necessarily done with malicious intent, but it seems unlikely that not a single human being will get hurt if these alterations to the programming of artificial agents happens on a regular basis. Next to this, it is possible for human beings to alter the goals of artificial agents with the intention of involving it in any sort of criminal activity. Responsibility needs to be attributed accurately in cases where actions by artificial agents have negative consequences. In these situations, a third party needs to determine whether the programming of an artificial agent has been altered. Based on this, responsibility can be attributed to either the designer or the user.

It is difficult to come up with a method to make dual use impossible. There are a few measures that can be taken to reduce the utilization and alteration of artificial agents for bad intentions. Firstly, altering the goals and programming of artificial agents with certain capabilities should be made illegal by law. And secondly, designers should be legally obliged to make it as difficult as possible to alter the goals and programming of artificial agents. This should be included in the standards for artificial agents. There are possible benefits to personal alteration of the programming of artificial agents. Experts can make modifications at home that might contribute to developing artificial agents that are better at executing certain tasks. However, it is difficult to determine who would be allowed to do this based on expertise. Moreover, it is dangerous to experiment with artificial agents in an open and uncontrolled environment. Altering the programming of artificial agents could be allowed in closed, experimental settings, if this could be beneficial to the capacities of artificial agents in general. However, the responsibility for the actions of the artificial agent would shift from the designer to the user that chooses to modify the artificial

agent. This includes legal liability, which would shift from the designer to the user that modified the artificial agents as well. Because the programming of artificial agents is a crucial part of the product, altering that programming would mean to radically change the product itself, which most likely includes changing the intended purpose of the product. In cases with undesirable outcomes, it will be important to decide what party is responsible for the actions of an artificial agent. One way of doing this is to identify whether the code of an artificial agent has been altered by the user.

Standards for Artificial Agents

Technological innovations that are developed with the intention of replacing older products and methods are not excluded from the principles of non-maleficence and beneficence. This means that newly developed methods or products should make a larger positive contribution to society in comparison with older options or they should be able reduce harm to society. At the very least, new developments should not result in an increase in harm to society. Luciano Floridi argues that determining whether an agent is morally good or bad depends on how well it performs its intended function (Floridi, 2013). Since an artificial agent performs tasks that have moral consequences, an artificial agent can considered as being morally good when a certain standard, or threshold function, is met. An artificial agent is considered to be morally bad when it does not meet a satisfactory level of performance. In this thesis, the satisfactory level of performance is determined by the contribution of artificial agents to society in executing its tasks. As I have discussed in this chapter, it is important that new technological methods are not more harmful to society than the method it is replacing.

The artificial agents discussed in this thesis are designed to perform tasks that used to be, or currently still are, performed by human beings. The principles of non-maleficence and beneficence state that products of designers should not cause harm to society and that they should make a positive contribution to society. Considering that artificial agents are intended to replace human beings for specific tasks, artificial agents should be at least equally competent in performing these tasks as human beings. Additionally, artificial agents can be deployed to perform tasks in situations that are considered dangerous or harmful to human beings. Moreover, the tasks they are supposed to execute, and for which they are designed, should not be performed with the intention of causing harm. Instead, artificial agents should execute tasks that make a positive contribution to society. These standards can be compared with the threshold function suggested by Floridi. The threshold for artificial agents would be to be at least as good as performing its tasks as human beings.

The subject of this thesis are artificial agents that operate with a high level of autonomy and which have to make decisions in situations in which they might have to deal with unfamiliar

situations and uncertainty. In order to fulfil the designer's moral responsibility, these artificial agents need to be at least as sufficient in performing these tasks as human beings. Because these artificial agents operate autonomously and are required to make decisions in an open and uncontrolled environment, they require a high level of autonomous functioning. In this thesis, I will investigate what level of autonomy is required in artificial agents in order to fulfil the designer's moral responsibility. The type of artificial agents addressed in this thesis need to deal with uncertainty and unfamiliarity, which require it to have a robust representation of its environment and it has to be able to anticipate the actions of other agents and the consequences of its own actions. Next to this functional or operational autonomy, the decisions these artificial agents are supposed to make have ethical consequences, because they operate in the physical world and their actions affect human beings. In the upcoming chapters, I will analyse what model of autonomy would be most suitable for artificial agents that are operating under these conditions. I will argue that the required level of autonomy for artificial agent will be limited compared to human autonomy, because their goals are determined by human beings and it is undesirable that artificial agents develop their own ethical beliefs.

Chapter 2: Autonomy

Introduction

In this chapter, I will address the second sub-question of this thesis. I will discuss the requirements of autonomy of artificial agents based on the role they play in society. There are many different conceptions of autonomy. I will argue in this chapter that it is important to take the role of an agent into account when a model of autonomy is developed. Artificial agents differ significantly from biological organisms when it comes to the role and goal of their autonomy. For the type of artificial agent that is discussed in this thesis, it is mainly important that they can complete tasks in an open and uncontrolled environment. Human beings want artificial agents to complete tasks that are determined by human beings. Therefore, artificial agents do not need to authorize their own actions. Consequently, they do not need to posses those aspects of autonomy. Moreover, as I will discuss in more detail in this chapter, the autonomy of biological organisms is a product of evolution. Autonomy allows biological organisms to become more independent from the environment and therefore enhance the chances of survival. Artificial agent are not responsible for their own survival. Therefore, there are significant differences in the requirements of autonomy between artificial agents and biological organisms. In order to clarify these differences, I will identify three different ideas of autonomy: personal autonomy, which means to govern oneself, autonomy as the ability to preserve oneself, and behavioural autonomy, which means to be able to autonomously perform tasks. The role of autonomy is important when considering these three different ideas of autonomy. I will discuss how these different ideas of autonomy can be interrelated but can exist independently as well. I do not attempt to answer questions regarding which beings are autonomous or not. I want to examine what aspects of autonomy are required for artificial agents. Autonomy is not an all-or-nothing property, and can therefore exist in different gradations. In the final section, I will address the problems with ascribing autonomy to artificial agents.

Personal Autonomy

There are a variety of perspectives on what it means to be autonomous. One conception of autonomy is that someone is autonomous when he or she governs himself. Any action this person undertakes are authorized by the self (Buss, 2014). Many definitions and models of autonomy also include that a being is autonomous when it has a conception of morality and when it is able to form its own moral law. A well-known and influential understanding of autonomy that include moral autonomy is Kant's definition of autonomy (Reath, 2013). According to Kant, to be called autonomous, a being has to formulate its own moral laws and act upon these self-imposed moral

laws. This conception has been hugely influential to the way autonomy is understood. Many conceptions of autonomy are influenced by Kant's view, such as the *coherentist* view on autonomy. According to this view, an agent is autonomous when the actions it undertakes are in coherence with mental states (Ekstrom, 1993). These mental states, which function as motivation for deliberate action, can be beliefs, desires or evaluative judgements. This means that an agent is autonomous even when he or she acts upon desires she cannot resist, but of which the agent knows that they are wrong, for example in the case of addiction. Moreover, nothing is stated about the origin or content of her beliefs and evaluative judgements. This means that an agent can be misinformed or is irrational about the formalization of his or her beliefs and can still considered autonomous. Therefore, the coherentist view does not take into account how beliefs and plans are shaped nor how and if they relate to reality. The *reasons-responsive* understanding of autonomy counters these objections against the coherentist view. According to the reasons-responsive conception of autonomy, the motives and beliefs that moves an agent should be well-informed and based on a variety of reasons for and against behaving in a certain way (McKenna, 2000). An agent needs to be fully aware of the reasons why he or she acts in a certain way in order to be considered as an autonomous agent. *Responsiveness-to-reasoning* is an account on autonomous agency that stresses the importance of the reasoning process itself. An autonomous being must be able to determine what actions it should perform in order to best represent their beliefs and desires. These conceptions of autonomy are all aimed at human autonomy. There are important differences between the autonomy of human beings and artificial agents. Therefore, the conception of autonomy for human beings is not necessarily appropriate for artificial agents.

Features such as moral autonomy and acting according to one's own beliefs and desires are arguably required when artificial agents are supposed to function as if they were equivalent to human beings. However, as I will argue in this thesis, it is not required that the conception of autonomy for artificial agents should include these features. Firstly, it is impossible to develop autonomy of such a level in artificial agents with current technologies and with technologies in the foreseeable future. Later on in this thesis, I will describe in detail why it is complicated and possibly not feasible to implement autonomy in artificial agents that is equivalent to human autonomy. Secondly, in contemporary uses of artificial agents, such degrees of autonomy are not required. Moreover, artificial autonomy that is equivalent to human autonomy might not be desirable in artificial agents in general, I will address this point in more detail in the chapter about moral autonomy and artificial agents. Despite this, researchers in computer science are keen to refer to a conception of autonomy that includes moral autonomy as well (Allen & Wallach, 2012).

Instead, a model of autonomy should be constructed that best represents the intended

function of artificial agents and the requirements of that function. The goal human beings have for artificial agents that are currently employed in society is to execute tasks that are determined by human beings. These tasks are to be executed in an open and uncontrolled environment. Artificial agents should be able to make justifiable decisions in order to reach that goal. The function of artificial agents do not require that their actions are initiated by their own free will. In the next section, I will discuss what the differences of the role of autonomy is between artificial agents and biological organisms, such as human beings.

The Role of Evolution in the Development of Autonomy

There are crucial differences to be found between the nature and purpose of autonomy when autonomy is compared between biological and artificial agents. Autonomy in biological agents originates from methods of survival and *autopoiesis* (Maturana & Varela, 1980). Autopoiesis is the property of a system to maintain itself through self-regulation. The evolution of biological organisms has been crucial in the development of autonomy. In contrast, artificial agents, as they are currently deployed in society and as they are defined in this thesis, are not required to be self-sufficient. Instead, successfully executing tasks and reaching certain goals are the origin for autonomy in artificial agents. I will first explain how evolution and autonomy are related to each other in biological organisms.

Various theories agree that autonomy is a product of evolution. In order to survive, more complex organisms are required to become independent from their environment. A widespread belief in evolutionary biology is that there has been an increase in autonomy through the course of animal evolution (Rosslenbroich, 2009).

Maturana and Varela have provided one of the most influential understandings of the relation between autonomy and the development of living, biological entities (Maturana & Varela, 1980). All action is subordinated to the perseverance of the organization of the living system. Such a living system is capable of maintaining its identity by actively compensating for deformations of the environment (Maturana & Varela, 1980, p. 135). This understanding includes a certain difference between autonomous *action* and autonomous *being*. The actions an organism undertakes are dedicated to persevering the organisation of itself. This means that there has to be a certain organization inside the system itself that is the reason for executing certain actions and processes that maintain it. The maintenance of this organization is active and actions comes from within the biological entity itself. In this understanding of autonomy for biological entities, actions are not chosen based on external factors alone, but processes within the system itself determine the course of action. Furthermore, it underlines the importance of autonomy for the existence and perseverance

of life and living organisms in changing environments. This is performed by monitoring environmental agency on the biological system according to internally defined needs of the system. The organization of the system determines the actions of that system on the environment (Moreno, Etxeberria & Umerez, 2008). This viewpoint results in the understanding of two separate, yet interdependent, aspects of autonomy. The first is *constitutive autonomy*, which determines the preservation of the internal structure and identity of a system. The second is *interactive autonomy*, which is a product of constitutive autonomy and is responsible for controlling interaction with the environment (Moreno, Etxeberria & Umerez, 2008, p. 312). More advanced biological organisms use increasingly elaborate forms of interactive autonomy to maintain their constitutive autonomy (Vernon, Lowe, Thill & Ziemke, 2015).

An increase in autonomy results in an increase in the independence of an organism from the environment. Instead, intrinsic functions become more important for an organism. Because of this evolution of autonomy, influences from the environment are decreased, which leads to an increase in adaptability. Consequently, the organism is better equipped to adapt to changes in the environment and to surviving in different environments altogether. In evolutionary theory, an increase in fitness means an increase in survivability.

There is a difference between autopoiesis and autonomy. Autopoeisis means that an agent can maintain itself in an environment. Autonomy means that an agent is able to compensate for external dependencies. Autonomy contributes to autopoiesis in biological organisms. It makes it possible for an organism to actively adapt to the environment by selecting actions based on information about the environment (Christensen & Hooker, 2000). Autopoeisis is not a requirement for artificial agents. Therefore, the role of autonomy is different in artificial agents. As I will discuss later, the role of autonomy in artificial agents is to allow them to have anticipation in order to complete tasks in open and uncontrolled environments.

With the argument of the role of evolution in the development of autonomy I attempt to underline a crucial difference between the requirements of autonomy in an artificial agent compared to human autonomy. One of the main reasons for the difference are dissimilarities in the goals of the system and the reason for autonomy in that system. Biological autonomy originates from evolutionary developments and the need for the survival and reproduction of organisms. An artificial agent is in general not responsible for its own survival or maintenance. While it is plausible that artificial agents are responsible for some sort of self-maintenance, the goal of artificial agents, at least the type of artificial agents discussed in this thesis, is not aimed at survival. It is only required to fulfil a certain function that is determined by an external source: human beings. It requires a interactive autonomy, however, for artificial agents, interactive autonomy is not

dependent on the artificial agent's constitutive autonomy. Instead, constitutive autonomy, as the goal for interactive autonomy, is replaced by goals that are determined by human beings. Therefore, a limited account of autonomy (limited in the sense that it does not include all the aspects of autonomy that are generally considered to be part of human autonomy) is sufficient for artificial agents.

Behavioural Autonomy

As argued in the previous two sections, what is required from artificial agents is that they can autonomously perform tasks. It is not required that they determine their own goals and they are not responsible for their own survival. Being able to perform task autonomously is often referred to as behavioural autonomy. This type of autonomy is similar to interactive autonomy (Froese, Virgo & Izquierdo, 2007), and the terms and definitions can be used interchangeably. It is the type of autonomy that allows agents to interact with its environment. Artificial agents are required to operate and complete tasks without the intervention of human beings. Justifiable autonomous functioning should be the start off point for autonomy in artificial agents.

Margaret Boden identifies three different aspects of autonomy (Boden, 1996). She argues that autonomy is not an all-or-nothing property. Instead, it has several aspects that come in different gradations. The first aspect makes a distinction between the manner in which behaviour and actions are determined. In less autonomous agents, action is mainly determined by its environment. Actions are selected in a direct response to circumstances within the environment. Agents become more autonomous when the experience of that agents plays a larger part in the actions it decides to perform. This result in a indirect response to the environment. An agent needs internal mechanisms to determine what action to perform. Such an agent has decision making processes that result in a deliberate response to its environment. Deliberating between different responses and having internal mechanisms to select an action based on environmental conditions brings certain requirements regarding representation and anticipation. I will address these capabilities and their importance for autonomous decision making in the next chapter.

The second aspect states that there is a difference in autonomy when behaviour is self-generated or externally imposed. An agent is more autonomous when its behaviour is the product of self-organizing processes. This aspect of autonomy is complex in artificial agents, since they are developed by human beings in principle. However, there are different graduations of the manner in which behaviour is imposed. Artificial agents with machine learning properties have partly self-generated behaviour. This aspect of behavioural autonomy has overlapping properties with personal autonomy. The advantage of autonomy as a continuous property as suggested by Boden is that it

allows for a graduation of different aspects of autonomy as well. Moreover, the first aspect of autonomy identified by Boden states that deliberate decision making based on internal mechanisms is important for autonomy. When an artificial agent possesses this type of indirect response mechanisms to environmental conditions, it becomes more complex to impose certain behaviours on artificial agents. In order to be able to deliberate about what the appropriate action is in a certain situation, an agent has to take the outcomes of possible actions into account. This requires a solid representation of the environment and anticipation of future states. Since they operate in an open and uncontrolled environment, decisions of artificial agents have moral consequences. I will discuss anticipation and moral decision making and their role in autonomy in chapter three and four, respectively.

The third aspect of autonomy deals with whether an agent can reflect upon its own directing mechanisms and selectively modify these mechanisms. This is different from the first aspect because it means to alter the mechanisms of decision making itself. To be able to reflect upon and alter one's own mechanisms implies that the agent has a certain consciousness. Conscious artificial agents have not yet been developed and this aspect of autonomy is therefore currently unavailable for artificial agents. However, as I will discuss in more detail in the chapter about moral autonomy, such high levels of behavioural autonomy might not be desirable since it results in a lower level of control over artificial agents.

Behavioural autonomy is the type of autonomy we should be looking for in artificial agents. Its different aspects, which have different gradations, allow for a flexible approach to autonomy.

Ascriptionality of Autonomy

In this section, I will address whether autonomy can be justifiably ascribed to an artificial agent. One of the main problems with any form of science is that it is observer dependent (Rohde & Stewart, 2008). An observer cannot adopt the point of view of the object of study. Therefore, it is difficult to assess whether particular features, such as autonomy, are genuine or only a belief by the observer. The problem of whether a feature is genuine or not has been discussed in the field of artificial intelligence.

In Alan Turing's famous paper *'*Computing Machinery and Intelligence' (Turing, 1950), he proposes a scenario that he calls the imitation game, which is now known as the Turing test. In short, the goal of the Turing test is to assess if a computer can trick a human being into believing that it is a person via written interaction. The goal of the test was to determine whether a robot could converse like a human and consequently to determine if the computer is intelligent. Turing predicted that around the start of the 21th century, an average interrogator would have no more than

a 70% chance of making the right identification. One of the major limitations of this test is that there is no way to verify if the computer is actually intelligent or merely imitating to be intelligent.

This problem was addressed by John Searle in his paper ‘Mind, Brains and Programs’ (Searle, 1980). In this paper, he introduces a thought experiment, known as the Chinese room. In this thought experiment, there are two scenarios with two different persons: person A who knows Chinese, and person B who does not. Suppose person A was to be locked up in a closed room and pieces of paper with questions written in Chinese were sent into the room. The person in the room is able to send out pieces of paper with Chinese text as well. If person A sends out reasonable answers in Chinese, then it is to be expected by people outside the room that person A knows Chinese, as is the case. In the second scenario of the example, person B is locked up in the same room. Person B receives pieces of paper with text in the Chinese language as well, but he does not know Chinese. However, there is a table in the room that allows him to look up the Chinese texts and allows person B to respond to the initial text in Chinese. For a person outside the room, there is no difference in outcome between the performance of person A and person B. However, person A understands the Chinese language and person B does not. The argument shows that the similar performance does not necessarily requires the same procedures. Person A really masters the Chinese language and Person B does not. This example extends to features such as intelligence and autonomy. The same input-output behaviour does not necessarily imply that the underlying mechanisms are identical as well. The Chinese room experiment underlines one of the problems with the Turing test: it can only measure the input-output behaviour and not the underlying mechanisms and procedures.

Searle underlined the differences of ascribing properties to an agent and genuinely knowing that an agent possesses certain properties. However, there are theories and arguments that defend the ascriptional approach. One of the prominent supporters of this approach is Dennett (Dennett, 1971). He identifies three different stances that allow the prediction of actions of other entities: the *design stance*, the *physical stance* and the *intentional stance*. The design stance is generally adopted when actions of mechanical objects are being predicted. If it is known how a computer and its programs are designed, it is possible to predict its behaviour in certain scenarios, provided that the system works properly and does not break down. The essential feature of the design stance is that predictions about a system's behaviour are made solely based on knowledge or assumptions about the functional design of a system (Dennett, 1971, p. 88). With the design stance, it is irrelevant to know how a system is structured physically. In contrast, the physical state of an object is relevant in the physical stance. In this stance, knowledge about the laws of nature in combination with knowledge about the physical constitution of an object are necessary in order to make predictions. It is virtually impossible to predict the behaviour of a computer when using the physical stance. The

physical constitution of a computer is too complex to allow for accurate predictions. However, it is used when a system malfunctions and the cause of the problem is of a physical nature and easily detectable (Dennett, 1971, p. 89).

Modern computers have become too complex to analyse from both the design and the physical stance. In order to be able to predict behaviour of complex computer program, one has to attempt to explain its behaviour based on the goals of the computer program and the assumption that it will execute rational actions to get to that goal. This is the same method that is used to predict the behaviour of human beings and animals. To rationally explain or predict an agent's behaviour based on rational action towards a goal is called the intentional stance. In the case of artificial agents, rational action means optimal design relative to a goal, or optimal hierarchy of goals and a set of constraints (Dennett, 1971, p. 89). Furthermore, the prediction of a system's behaviour is relative to the nature and extent of the information a system has. Therefore, to adopt the intentional stance to predict a system's behaviour relies on three aspects: the knowledge the system has about the constraints, the knowledge or representation about the current state of affairs and the goals of a system. If a person can reliably predict a system's behaviour using the intentional stance, it can be stated that the system has a certain intentionality.

The intentional stance makes it possible to ascribe the necessary amount of autonomy to an artificial agent. Because the goals of artificial agents are known, as they are determined by human beings, we can determine if their actions are rational towards a goal. Additionally, it is arguably not required to find out whether the existing amount of autonomy is genuine. All that is needed from artificial agents is that they execute their tasks in a justifiable manner. If they are able to do this, it does not matter whether their autonomous capacities are genuine or simulated. However, it is questionable if the mere simulation of autonomy is enough to be able to execute tasks in a justifiable manner.

Conclusion

In this chapter, I have described that there are multiple facets to autonomy. Autonomy can include being able to maintain oneself, being able to formulate one's own beliefs and values and autonomous action, amongst other definitions. Because of this, questions regarding whether something or someone is autonomous cannot be answered with a simple yes or no. Therefore, when artificial agents are designed, the role of the artificial agent should be considered in detail to determine what aspects of autonomy should be included within the artificial agent in order to state whether the autonomous capacities of an artificial agents are sufficient for its intended task. For artificial agents discussed in this thesis, it is predominantly important that they are able to complete

tasks and reach goals that are determined by human beings. They are not required to determine their own goals and are not responsible for their own survival or maintenance.

Artificial agents are required to complete tasks in an open and uncontrolled environment. They do not have to make decisions about what their goals are, since those are determined by human beings. They do need to make decisions about how to reach those goals. This is complex in an open and uncontrolled environment for various reasons. Firstly, because there are other agents in the environment that might act in an unpredictable manner. Secondly, it is unlikely that an artificial agent, like any other agent, will have complete and perfect information about the environment and other agents in that environment. This results in uncertainty and situations and events that are difficult to predict. The environment they operate in also includes human beings. Therefore, the decisions artificial agents make affects human beings and have moral consequences. Operating up to standards that fulfil the designer's moral responsibility demands a set of capabilities for artificial agents. They need to be able to have a robust representation of the environment in order to make sense of it. Moreover, they are required to make moral decisions and consequently require to have an understanding about morality and moral consequences of their actions. In order to make correct decisions in an open environment, artificial agents need to be able to anticipate events in the environment. In the coming chapters, I will discuss why these aspects of autonomy are necessary and what is asked of the capabilities of artificial agents regarding these aspects in order to be able to make justifiable decisions and fulfil the moral obligations of the designers.

Chapter 3: Autonomy, Representation and Anticipation in Artificial Agents

Introduction

In the previous chapter, I have discussed that the function of autonomy of artificial agents differs from the function of autonomy of biological organisms. For artificial agents, it is predominantly important that they are able to make justifiable decisions in order to reach a pre-determined goal in an open and uncontrolled environment. In this chapter, I will discuss why anticipation and representation are necessary for autonomy of artificial agents in order to make justifiable decisions in these conditions. I will address the third and fourth sub-questions of this thesis. As I argued in the previous chapter, artificial agents require a restricted understanding of autonomy. In this understanding, goals are set externally and autonomy is a teleological function as it guides an agent in its functionality. The artificial agents discussed in this thesis do not have to determine their own goals. Therefore, the most important aspect of autonomy is autonomous decision making in order to reach goals that are set externally. There are two main aspects that are crucial for autonomous decision making in an open and uncontrolled environment. First, the artificial agent should be able to represent its environment and agents in that environment. Secondly, the agent should be able to anticipate future states based on the representation of the present. Artificial agents operate in an open and uncontrolled environment which include other agents and parts of the environment that are not observable but are important aspects for selecting justifiable actions. Therefore, both representation and anticipation become rather complex.

There are many factors that should be taken into account when a decision needs to be made in an open and uncontrolled environment. The artificial agent needs to anticipate the actions of other agents, possible unforeseeable events and the outcome of its own actions, amongst others. From a representation of the current state, an agent with the appropriate capabilities can anticipate future states. There are requirements for representational capabilities in order to have a level of anticipation that meet the demands of being able to operate justifiably in an open and uncontrolled environment. The agent needs to be able to represent moving objects in order to anticipate the position of that object at a certain time interval in the future, for example. To anticipate future events and plan its own actions based on these anticipations, the artificial agents requires the capability to predict the probability and desirability of events and outcomes. I will expand on these concepts in this chapter.

That the goals of an agent are set externally does not necessarily mean that the agent is dependent on direct input-output mechanisms. In agents with a direct input-output mechanism, the agent itself is unable to make any decision regarding the output that follows from a certain input.

There is a direct link between input and output and the same input will always result in the same output. Systems that operate in this way are *deterministic* systems. In order to have an agent that functions in more complex situations and environments, an agent must be capable of predicting, and potentially adapting to, future states. The agent must therefore be able to anticipate future states based on the input it receives and select the most desirable future state. Based on this information, the agent has to decide which output is the most suitable to reach this most desirable future state.

When an agent is capable of selecting its own actions based on possible future states, the same input (or collections of input) can result in different output. The agent has an internal selection procedure which can select the most optimal future state and is able to determine which outputs are most suitable to reach the most optimal future state. In this chapter, I will expand on how internal selection procedures and anticipation are crucial for obtaining then essential level of behavioural autonomy. This type of autonomy is required for agents to function properly in situations in which unfamiliarity and unexpected events are a factor. A crucial aspect for anticipation is the representation of current states. I will analyse the role of representations of the present and how they can contribute in making anticipations about future states. From this analysis, I will derive what capabilities an artificial agent must have in order to make anticipations about future states that allow it to make justifiable decisions.

Machine learning is a sub-field within artificial intelligence and is part of the discussion in the upcoming sections. In short, machine allows artificial agents to act without being explicitly programmed (Michalski, Carbonell, & Mitchell, 2013). Machine learning aims to develop computer programs that can teach themselves by modifying their model based on exposed data. Machine learning can be utilized to make computer programs detect patterns and adjust its actions accordingly.

Artificial Agents and Anticipation

Anticipation in an agent means it is able to adjust its current behaviour based on its knowledge about future states. A distinction can be made between causal systems and anticipatory systems. Anticipatory systems take future states into account when they plan their actions, while causal systems respond more directly to current events and do not have knowledge about future states (Davidsson, Astor, Ekdahl, 1994). Traditional artificial agents used sensor data to produce a model of the world. This model of the world is subsequently used to plan actions. A famous, early example is the robot named Shakey (Nilsson, 1984). Shakey operated in a room or multiple rooms and was told to perform tasks such as navigating around the room or push coloured blocks to pre-determined location. This type of artificial agent is capable of performing cognitive tasks, such as

planning and problem solving. However, they are not suited to making fast decisions that are required for more basic tasks (Brooks, 1991).

Reactive agents were created in order to solve this problem and to develop artificial agents that were able to rapidly perform tasks (Georgeff & Lansky, 1987). They operated by having an internal collection of behaviours of which one was selected based on input. This removed the need for artificial systems to make a model of the world and perform behavioural planning based on this model. This type of artificial agent is successful in performing simple tasks but is not particularly versatile. They heavily rely on their sensors and direct input from the world and are unable to perform tasks that require knowledge about the world, a certain level of reasoning or knowledge from memory. This creates problems when artificial agents are required to perform more complex tasks. There are various identifiable tasks that this type of agent is unable to perform. An example is an automated vacuum cleaner. When switched on, it will move in a direction until one of its sensors comes into contact with an object that is in the way of the vacuum cleaner. The agent would then turn for a certain number of degrees in order to be able to keep moving. Automated vacuum cleaners of this type are unaware of their tasks, the environment the operate in, and where they are within that environment. In general, artificial agents of this type are unable to respond effectively to events that are beyond the agent's current sensory inputs. They lack capabilities to make predictions of the behaviour of other agents, and are therefore not suited to be deployed in situations that involve other agents (Kirsh, 1991).

Artificial agents have to be capable to perform these tasks if they are to be deployed in more complex environments and are required to perform more advanced tasks. In the coming section, I will describe the relation between anticipation and autonomy.

Anticipation and Autonomy

Anticipation is to act on predicted future states. One of the most commonly used definitions of anticipation is by Rosen: “A system containing a predictive model of itself and/or its environment, which allows it to change state at an instant in accord with the model's predictions pertaining to a later instant.” (Rosen, 1985, p 339) In order to operate autonomously, anticipation is necessary to interact dynamically within an unfamiliar, changing environment. Based on its predictions of future states, an agent can perform certain actions in the present in order to reach future states that conform with its goals. In mobile agents, perception and representation are important aspects of anticipation. An agent that is moving forward and possesses sufficient perception capabilities can gather information about the environmental conditions it will have to deal with in the near future. Based on that information, it can start planning its actions. This implies

that perception and representation are crucial for anticipation and autonomy. In this thesis, I will not go into detail about perception of artificial agents. For the argument, I will assume that they have sufficient perception to construct solid representations of the environment.

Anticipation is crucial to autonomy because it contributes to dynamic interaction with the environment. An anticipatory agent is able to choose and modify its actions based on information about what it will encounter in the future. This is particularly important when an agent is trying to achieve certain goals in a dynamic environment, such as one that includes other agents. Being able to regulate and adapt one's behaviour to suit the context in order to achieve one's goals is referred to as *self-directedness* (Christensen & Hooker, 2000). The self in self-directedness refers to the normative goals selected by the agent. For the type of artificial agents discussed in this thesis, these goals are not determined by the agent itself. Instead, the goals are set by the designers and users of the artificial agent. However, these artificial agents still need to modify their actions in a dynamic, unpredictable environment in order to reach these goals. Therefore they still require sufficient anticipatory capabilities. Anticipation and directed action underline the difference between autopoiesis (being able to maintain and reproduce oneself) and autonomy. Autopoiesis can be achieved by dynamic processes and interactions with the environment. However, these processes and interactions are not goal directed actions. This is in contrast to autonomy, in which actions are performed to reach a certain goal. The relation between autonomy, anticipation and self-directedness emphasizes the distinction between autonomy and autopoiesis. Autopoiesis is not directed and does therefore not require anticipation. The artificial agents discussed in this thesis do not determine their own goals. Therefore, the directedness of artificial agents discussed in this thesis does not come from the agents themselves. However, the artificial agents still requires anticipation, and therefore autonomy, to reach their goals. In contrast, autopoiesis is required in all living organisms in order to maintain themselves. Artificial agents do not need to maintain themselves in that sense. Instead, they require to accomplish externally set goals and tasks in a dynamic environment. To reach these goals, they need to make acceptable decisions based on autonomous anticipation.

The Strength of Anticipations

In the previous section, I explained the importance of anticipation for planning autonomous action in a dynamic, unpredictable environment. In this section, I will attempt to provide an account of the necessary strength of the anticipation capacities in order to rely on artificial agents in dynamic environments which include other agents. The anticipatory power of an agent depends on the width of its anticipatory time window (Christensen & Hooker, 2000, p. 139). The width of the

anticipatory time window refers to how far into the future an agent is able to anticipate events. A longer time window will result in more anticipatory power because the agent is able to plan one's actions over a longer amount of time and further into the future.

An agent has to be able to do more than just anticipate its environment if it wants to select actions that are the most likely to help it succeed in completing its goals. In order to select the most suitable action to reach its goals, an agent needs to be able to anticipate its own actions. In this manner, it can select the action that will result in reaching the goal or select the actions that comes closest to reaching its goal (Lavigne & Lavigne, 2000). The agent needs to be able to anticipate its own actions and the environment. This is because the artificial agents discussed here are not only required to predict what will happen in the environment, but are also required to act upon the predictions they have made about the environment. In order to be able to anticipate the environment and its own actions, the artificial agent needs to have a model of the environment and a model of itself as part of the environment (Astor, Ekdahl, Davidsson, Gustavsson, 1991). Furthermore, the model should be dynamic because the agent should be able to change or update the model when itself or the environment changes, or at least when its perception or anticipation about the environment or itself changes. This means that the agent should be aware that its current model of the world might be incorrect. It should be able to revise its model when changes in the information it has of the environment makes it aware of errors in its current model.

To be able to anticipate what is going to happen, the agent needs to construct a model of what will happen in the future based on the model it has about the present. It needs to construct a predictive model in order to be able to anticipate the environment and its own actions. Thus, for anticipation, at least two points in time are crucial: the moment in which the anticipation takes place, and the anticipated moment. Two time intervals is the absolute minimal for anticipation (Chrisley, 2002). Agents that have anticipatory capabilities, for example human beings, are able to make a prediction for multiple future moments. For example, it is possible to anticipate what will happen in five minutes and what will happen in two months. Moreover, human beings can take past anticipations into account when predicting the future. For example, it is possible to include an anticipation that was constructed two minutes ago about what is going to happen in five minutes in one's current anticipation. One of the problems with this is that there is a virtually unlimited amount of predictions to be taken into account, especially since not all past predictions are relevant for the current situation. An agent should only take those past anticipations that are still relevant into account. Furthermore, past anticipations should not be taken as a separated prediction. Instead, they should contribute to the current anticipation. This is of critical importance when an agent is anticipating moving objects and other agents, since there are two observations in different time

intervals required to predict velocity and direction.

I will illustrate the necessity for an agent to construct models of the environment in different time intervals with an example of an autonomous car. In this scenario, an autonomous car is driving on a road and intends to overtake another car that is driving in front of it. To simplify this example, all factors except for the two cars are constant and there are no other agents on the road. There are numerous factors that the car needs to take into account in order to successfully and safely overtake the other car. It needs to be able to recognize its own velocity and its place on the road. It also needs to analyse the velocity and place on the road of the other car and anticipate the direction it is travelling in as accurately as possible. From the collected data, it should be able to create a model about the current state of the relevant factors of the world. From this model it should predict a future model of a certain time interval with information regarding where the car itself is and where the other car is. It is important to note that there is an endless number of future time intervals. An artificial agent does not have unlimited computational power. Therefore, the autonomous car needs to be able to identify certain critical future events and base its anticipation on the time interval in which a critical event is likely to happen. This does not mean that the artificial agent can only anticipate one future time interval. It does mean that it has to make a selection of possible future events and anticipate the most likely or dangerous events at a certain time interval. This can be compared with chess computers, which are able to see many turns or moves ahead (Thompson, 2014). However, they have certain selection methods to reduce the necessary computation power by analysing what the most likely and successful moves are for a given situation. A similar method is likely to be required for autonomous cars. However, this would be more complex than the selection procedures in a chess computer. In chess, there are a limited number of possible moves at any given interval and the possible actions are restricted to pre-determined time intervals. In real life scenarios, there are an unlimited amount of events possible in a continuous time flow.

An autonomous car needs to reduce its anticipations of events and time intervals in order to reduce the computational requirements to a feasible level. It also has to recognize events in the present that might lead to dangerous situations. For example, objects that come towards the car itself, or where the car itself is moving towards, require more attention when it comes to anticipation, especially when they rapidly move closer to each other. To return to the example, if the car in front moves with a constant, slightly slower speed than the car behind, the autonomous vehicle that wants to overtake the vehicle in front of it only needs to anticipate the moment in time where the cars are close enough to each other to start the overtaking procedure. Of course, it needs to closely monitor the car in front of it for changes in velocity or direction. However, it is neither possible or necessary to anticipate every possible change in velocity or direction.

To summarize, an artificial agent needs to construct a model of the present state of events, including itself. This model includes the location, direction and velocity of objects in the world. From this model, it can construct a model of the future. Based on this anticipated future model, the agent can plan its own actions. Objects in the environment need to be constantly monitored so that the model of the present is constantly updated. When changes occur in the environment, the model of the present (and models of the future based on it) should be altered accordingly. This allows the agent to change its actions and planned actions if necessary. When an agent has constructed a suitable model of the present and from that model is able to anticipate a model of the future, it can start planning its actions. The planning of an action includes the time interval in which the action is going to take place, therefore limiting the amount of specific predictions of future time intervals. Secondly, it can anticipate what changes in the environment can result in dangerous situations or an alteration in the chosen course of action. Examples of these changes are a change of direction or speed of an already monitored object, or the emergence of a new object in the model of the present.

Weak and Strong Anticipation

A system in the real world can never have true knowledge about future states of its environment. It is required to make anticipations that allow it to make predictions about the future. Dubois has made a distinction between two different types of anticipation based on an agent's capability to internally select input-output procedures (Dubois, 2003). He defines an anticipatory agent as a system whose present behaviour is not only based on past and present events but on anticipated future states based on these past and present events as well. A systems with anticipatory capabilities is an incursive system. An incursive system is able to predict possible future states which contribute in its decision making. Dubois makes a distinction between two different types of anticipation, weak and strong anticipation. In *strong anticipation*, an agent uses internally produced data to model its future internal state. In this type of anticipation, referred to as *endo-anticipation* (Dubois, 2003), an artificial agent can be relatively sure about the future states, since it depends predominantly on its own behaviour. In *weak anticipation*, an agent uses external data to internally model future states of the environment. This is *exo-anticipation* (Dubois, 2003), in which a system makes anticipations about its environment and the agents within this environment. This type of anticipation is more dependant on predictions than strong anticipation, since a system cannot be entirely sure about the future states of external systems. A *hyperincursive system* (Dubois, 2003) builds more than one future state. Because multiple states are anticipated, and an action is decided based on the most likely or desirable of those states, there is no direct input-output mechanism.

It is possible for a system to predict its own future states, as long as it is able to construct a

model of its own internal states (Collier, 2008). It is impossible for a system to be sure about its predicted future states of external systems. However, it is possible to anticipate multiple future states of data external to the system. Based on these states, a system can start planning its actions in order to reach a state that is most desirable. I will expand on this concept in the upcoming sections.

Modelling Future States

A model of the present needs to meet certain conditions in order to be able to predict future states. The model of the present is a logical model that should be generated in such a way that it represents the causal relations of systems (Rosen, 1991). This is based on the conception that every object in the world has a causal structure. In order to have a representation of itself or other systems, an agents needs to construct a logical model. The logical relations of this model should mirror the causal relations of the real world (Collier, 2006) If the logical model represents the world sufficiently, the agent is able to have a representation of the present. In the artificial agents discussed in this thesis, a sufficient model means that the representation allows artificial agents to make justifiable decisions. In order to make justifiable decisions, the artificial agent is required to anticipate future states. This is possible when the representation of the present that exists in the logical model can be projected to model future states. An agent with these capabilities is able to anticipate future states. In order to this, representation of the present is crucial.

Representation

As discussed in the previous sections, representation is crucial for anticipation and thus for decision making processes. Representations are models that are related to what an agent perceives of reality. What is represented in the model is based on what an agent considers to be important. They are constructed by living systems such as human beings in order to make sense of the world. A representation consists of information about the environment. It is the only kind of information that is available to the agent about the world (Bickhard, 2000). The only way to check if a representation is correct is by comparing it to other representations.

It is difficult to analyse to what extent representations correspond with the real world, especially since there is no other source or method to examine what it is that they are representing. However, it is unnecessary to explore whether representations of artificial agents are correct. For artificial agents, it is only required that their representations allow for sufficient anticipation which in turn should lead to justifiable decision making.

Measuring Representations

Since representations are our only source of what the world is like, it is complicated or even impossible to analyse their accuracy or correctness. Therefore, instead of directly examining representations, the outcome of actions based on these representations can be evaluated in order to get some sort of qualitative measure for representations. It is possible to evaluate whether certain goals are reached. These concepts originate from a pragmatist approach to representation. Pragmatism rejects the idea that mental states have content that exists as an intrinsic property of that mental state. It also rejects the claim that the content of a mental state has a meaning that is identifiable purely by accessing that mental state, without reference to the actual world (Hookway, 2016). Instead, pragmatism states that the content of a mental state has meaning in reference to the role it can play in actions which in turn should lead to accomplishing a certain goal. Therefore, the content of representations needs to be explained in terms of reference to what can be done with it. In pragmatism, mental states do not exist without reference to the real world. Additionally, the correspondence between the representation and the real world is a functional one, in contrast to having an informational correspondence with the real world, in which there has to be correspondence of informational content between the real world and representation.

An active agent that interacts with the environment has to choose certain actions. Actions and behaviours are selected based on the desires and goals of that agent. For example, in biological organisms, the sensation of hunger triggers the behaviour to search for food. This is a very general and simplified version of behaviour selection processes in a biological organism.

Every active agent that interacts with the world needs to select certain actions. In simplistic cases of action selection, this can be a trigger in the environment or in the agent itself that results in selecting a specific action or behaviour, as the example above shows. This type of direct action selection only works when there is enough certainty that the selected action is the correct one or when there are no or limited consequences when the selected action is the wrong one. This is possible in biological or artificial agents that execute tasks that are simple enough to be controlled by one-dimensional input-output mechanisms. For more complex agents, there are usually more inputs that guide action and a variety of possible actions that can be selected to reach a certain goal. This means that the agent much be able to anticipate that a certain action in a certain situation brings the agent to, or closer to, a desirable goal. A major advantage in artificial agents as they are discussed in this thesis is that none of their goals need to be determined by the system itself. Its goals are set by the designers or users of the artificial agent. Therefore, it is less complicated to detect from an outside perspective whether the representations of an artificial agent are correct. There is not the problem of not knowing its goals. From a pragmatic point of view, a representation is sufficient when it allows an agent to reach its goal. Since the goals of artificial agents are known,

it is possible to derive whether their representations are sufficient from their success in reaching a certain goal.

Knowledge Representation

Knowledge representation in artificial intelligence aims to find methods for artificial agents to represent information about the world in order to make sense of the world and utilize this knowledge to complete tasks. In one of the most influential works within artificial intelligence regarding knowledge representation, Davis, Shrobe and Szolovits define knowledge representation by dividing it into five different roles in can play (Davis, Shrobe, & Szolovits, 1993). The different roles illustrate the different requirements and properties that a representation can have. The set of roles combined should allow for a framework about what representations are and at the same time underline differences and similarities different roles require in representations.

First, knowledge representation functions as a surrogate. It allows a system to reason about the world instead of having to directly act in it. One of the major complications of reasoning is that reasoning itself takes place internally, while the objects that are the subject of the reasoning processes exist externally from the system. Actions that take place in reasoning processes are substitutes for real action in the world. Representations of objects in the world are never perfect. This is in principle impossible because nothing is identical to something except for the thing itself. While it is possible to have perfect representation of formal notions such as mathematical concepts, in most situations reasoning tasks usually involve objects that exist in the physical world. Therefore, imperfect representations are inevitable. In order to represent real-world objects accurately, it is necessary to use simplified versions of the objects as representations, because the almost limitless complexity of real-world objects needs to be reduced. An important conclusion that can be drawn from this is that there will always be a possibility of error when an agent is reasoning and interacting with objects in the real world. Because no model of the world can possibly be perfect, there will always be the possibility of mistakes. The risk of making mistakes can be reduced by finding representations that are sufficient for the goal to be achieved. It is important to note that more complex tasks are likely to require more complex representations. Therefore, more complex tasks increase the chance of mistakes in representations of the world.

Second, representation answers ontological question about how a system should think about the world. This is a result of the first role representations can play. Each representation is an approximation of reality. It cannot capture the entirety of reality. Therefore, a representation places a higher value in some aspects while it ignores other aspects of the world. Since none of the representations are perfect, a representation needs to be selected. Deciding upon a certain

representation is committing to certain ontological beliefs about the world. Selecting a certain representation means selecting aspects of the world that are believed to be relevant to the system. In this way, representations can tell an agent how to perceive the world. It contributes to focussing on the relevant and useful aspects of the world. Therefore, representations are crucial in interpreting how to act in the world.

The third role of representation deals with how to reason intelligently in the world. Intelligent reasoning typically determines the initial conception of a representation. This role is fragmentary because the belief or insight that initiated the representation is usually only incorporated partly in the representation and the belief or insight itself is only a part of the intelligent reasoning by an agent. This role deals with what it means to reason intelligently, the content than can be inferred from what is known, and what should be inferred from what is known. The first part, what it means to reason intelligently, is the most complex. There are different accounts on what intelligent reasoning is. Logic is the method that is used for intelligent reasoning in artificial intelligence (Charlesworth, 2014)(Walton, 2016). Reasoning intelligently is a process that can be captured in formal descriptions, such as in first-order logic.

The fourth role identified by Davis, Shrobe and Szolovits states that representations are a medium to perform efficient computations. Representations make it possible to organize information in order to make inferences from it. These include how certain information can be used and what can be expected from certain information, among others. They are therefore crucial to accomplishing tasks and reaching goals. Effective representations allows more effective computations. Therefore, representations are a crucial aspect for executing tasks and achieving goals.

The fifth role describes representation as a way of communication and expression. Representations are used to tell others, including artificial agents, about the world. We create and use representations as a medium of communication. This role deals with how easy or difficult it is to use a representation as a means for expression and communications.

To summarize, representations make it possible for a system to reason about the world before having to actually perform actions in it. A representation is never perfect or complete but allows the world to be interpreted in a way that makes is understandable for the system. In artificial intelligence, logical methods are used to represent the world. From these representations, an artificial agent is able to organize information and make inferences from them. In the next sections, I will evaluate how representation works in terms of logic and what is required to make anticipations based on these representations.

Logic for Representations in Artificial Agents

As discussed in the previous section, logic is crucial for representations in artificial agents. In this section I will analyse how representation takes place within an artificial intelligence system.

Russell and Norvig describe the role of logic in the representations of artificial agents (Russell & Norvig, 2014). A crucial aspect of representation in logical formulations is that there has to be a certain categorization of objects in the real world. Categorizing objects makes it possible to make predictions about their properties and behaviour. Logic can represent attributes of real world objects that are significant for decision making. For example, an autonomous car should be able to recognize other cars on the road. The colour or shape of the vehicle is initially not important, because it should avoid crashes with any car on the road. From the main category of road users, subcategories can be derived: based on colour, weight or size, for example. Members of a subcategory are ascribed particular properties that are more specific compared to properties of the main category. One of the main problems with representation in logic is that real world objects, or natural kind categories, have no clear definition (Russell & Norvig, 2014, p. 450). Objects of the same category, such as cars, can have many different shapes and sizes. It can therefore be difficult for an artificial agent to state with absolute certainty if objects really are what it has perceived them as being. This is further complicated by partially observable environments and objects, Such as being able to only detect the first half of an approaching car at an intersection. In addition to representing objects, an artificial agent in a dynamical environment should be able to make representations of actions and events, especially when they have to make anticipations of future events.

The Frame Problem

One of the major problems with designing artificial agents has been the frame problem (Dennett, 2006). This problem is both a logical and an epistemological one. From the viewpoint of many researchers in the field of artificial intelligence, the frame problem refers to the difficulties of developing an artificial agent that is is able to represent which aspects of the world do not change when an action takes place. From a philosophical perspective, the frame problem is often understood as being concerned with how an artificial agent can discriminate between relevant and irrelevant aspects of the world.

The frame problem is generally considered to be limited to artificial systems. The frame problem does not seem to affect human beings for two reasons: human beings have emotions that limit the amount of actions that are considered (Evans, 2002); and secondly, human beings can direct their attention to limit the amount of actions that have to be considered.

Overcoming the frame problem would result in many advantages for artificial agents, especially those that are operating in the real world. To be able to distinguish between relevant and irrelevant factors would severely reduce the required computational power. Consequently, an artificial agent can utilize its computational power to anticipate relevant factors, such as events that might be dangerous. This would result in an increase in anticipatory power because more available computational power will result in an increase in the width of the anticipatory time window. Being able to take into account more possible future scenarios and to predict events further into the future should result in better judgement by artificial agents.

There has not yet been a definite solution to the frame problem. However, researchers in artificial intelligence have made attempts to find a solution of the frame problem in logical descriptions of the world.

Situation calculus is the classical manner in artificial intelligence to describe situations in logical terms (Russell & Norvig, 2014, p. 453). While it is useful for describing situations, it is designed to describe worlds where actions are discrete and instantaneous, rather than describing dynamical, real worlds. Because situation calculus is based on situations and not on actions, it is not very useful for anticipation. Furthermore, it cannot be used to describe actions that take place at the same time. Every time an action is performed by an object, statements have to be added to make the system aware that other objects have not changed. Situation calculus requires that every situation or object that did not change from a particular action has to be described, next to the ones that did change. In real world scenarios, this means that a system needs an extreme amount of computational power to describe all the facts in the world that did not change. This factor, in combination with a system's relative inability to represent moving objects, results in situation calculus being part of the frame problem. Therefore, researchers in artificial intelligence have developed logical representations that are better able to detect fluent processes and do not require statements about objects that do not change after a certain action has taken place. One of the possible logical formulations that can deal with this is event calculus (Shanahan, 1999). Event calculus is better suited for the representation of continuous events, causal effects and simultaneous actions (Kowalski & Sergot, 1989). As mentioned above, while a definite solution for the frame problem has yet to be found, there have been developments that make it possible for artificial agents to represent the above mentioned situations and actions by using event calculus (Artikis, Sergot, & Paliouras, 2015) (Brandano, 2013) (Skarlatidis, Paliouras, Artikis & Vouros, 2015).

The Role of Probability and Utility in Applying Representations and Anticipations for Justifiable Decision Making

Anticipating and taking into account all possible future situations can become a problem in artificial agents. For example, an autonomous car plans to overtake a car that is driven by a human driver. There is always a theoretical chance that a human car driver swerves out of their lane and onto the next one. The autonomous car cannot know for sure if this will happen or not. However, it does not make a lot of sense to act upon this possible future scenario, because if the agent does so it will never be able to overtake a car that is driven by a human being. There are numerous imaginable situations like this because there is always a level of uncertainty when it comes to future scenarios. Therefore, it becomes difficult to state what the right thing to do is. The artificial agent cannot know for sure what actions will be successful, because it can not be inferred from the information it currently has. This problem is known as the qualification problem (Russell & Norvig, 2014, p. 489). In order to solve this problem, and make decisions that make it most likely that an intended goal is reached, an agent needs to take the relevant importance of a goal into account, the likelihood of certain events and the degree to which a certain goal is or can be reached. This is referred to as rational decision making. To be able to do this, an artificial agent need to place some degrees of belief into certain possible scenarios. With standard logic reasoning, an agent can either believe that something is true or false. What is required in situations like this is probabilistic reasoning, which puts a certain degree of confidence on a belief (Russell & Norvig, 2014, p. 490). It represents the the assumed likelihood on the occurrence of an event. By placing a degree on beliefs about current events and possible future events, an artificial agent is able to make better considered decisions. A crucial aspect for rational decision making is the value or preference an agent puts in a certain outcome.

Next to probability, an agent needs to take the utility of a certain outcome into account. Referring back to autonomous cars, there is usually a low probability of a crash, but a high utility value for avoiding a crash. However, wanting to avoid all crashes with absolute certainty will most likely result in an autonomous car that will never decide to actually start driving. However, there is a value of utility in driving the car with the purpose of reaching destinations. With suitably attuned utility preferences, the autonomous car will decide to start driving. It is important to note that utility comes down to preferences, therefore they will be pre-programmed subjective values that require fine-tuning. The autonomous vehicle of the example, that has as a goal to overtake a car driven by a human being, can put a very low degree of probability on the possibility that the car in front of it will swerve to the next lane. Therefore, it can allow itself to overtake the car, while still taking into account the possibility that the car might swerve in front of it. Another example would be a car that is about to cross an intersection. It might not be able to perceive whether there is a car or other traffic user that is coming from the right. However, it can put a certain value on the probability of

that happening. It can then decide to cross the intersection, based on the utility value, but at a lower speed, in order to anticipate that someone might be coming from the right.

Utility and probability in combination with machine learning can help an artificial agent in determining how to best reach a certain goal. Because artificial agents operate in an open and uncontrolled environment that include human beings, their actions have moral consequences. In the next chapter, I will discuss in more detail how moral considerations are important for deciding what the most desirable future state is.

Conclusion

In this chapter, I have analysed what is required from representation and anticipation to make justifiable decisions. Artificial agents need to be able to predict future states in order to make justifiable decisions autonomously. To be able to select the most desirable future state, a systems should be capable of predicting multiple future states. Such a system is a hyperincursive system. Anticipations are based on representations of current states. From these states, future states can be anticipated. Logic plays a crucial role in representations in artificial systems. In order to be able to predict future states from present states adequately, an agent should have the capability to represent moving objects and be able to represent multiple objects at the same time. Event calculus is a type of logic that is used in artificial intelligence to make that possible. One of the other advantages of event calculus is that it might contribute in finding a partial solution for the frame problem. An artificial agent should select the most optimal state and decide how to act based on that state in order to make the most desirable decisions. Calculations regarding probability and utility of certain states and goals are important in deciding actions of an artificial agent. Together, this should result in anticipation and representation that is sufficient for autonomous decision making in open and uncontrolled environments.

Chapter 4: Moral Autonomy and Artificial Agents

Introduction

The number of artificial agents that are deployed in society is increasing. Some examples are autonomous cars, robots that are deployed in healthcare settings and military robots. Moreover, the roles they play in society are becoming increasingly complex, which has resulted in more responsibility being placed upon them. Artificial agents do not necessarily operate in controlled environments, such as factories, anymore. Moreover, these artificial agents are operating in situations in which an automatic emergency shutdown itself is harmful, such as an emergency shutdown of a self-driving car on the road, for example. They are deployed in open, real world environments in which unpredictable events play a role and their actions affect human beings. Human beings have decreased control over these artificial agents regarding how tasks are executed. Instead of a human being having control over an artificial agent, the artificial agent relies on internal mechanisms to determine what to do. This includes moral judgements, which require additional mechanisms apart from those that allow an artificial agent to operate autonomously in an operational manner. Artificial agents are operating with higher levels of autonomy, therefore, their actions have moral consequences. This means that they should posses the ability to act morally. However, contemporary artificial agents are ethically blind. They are not designed to detect and consider ethical relevant features of the world and their decision making does not include moral considerations (Allen & Wallach, 2012).

Computer science research often refers to definitions of autonomy that include moral autonomy (Powers, 2013). As I have discussed in chapter two, this conception of autonomy is influenced by Kant. According to this understanding of autonomy, an agent is autonomous when it has a conception of morality and is able to form its own moral laws (Schmidt & Kraemer, 2006). As I argued in chapter two, it is not required nor desirable for artificial agents to form their own moral laws, because their goals are determined by human beings. Another problem with this conception of autonomy is that autonomous agents must be free from the moral values and will of the creator in order to form their own moral laws. In order to be free from the moral laws of its creator, the artificial agent has to obtain a certain level of independence from its creator (Di Paolo & Iizuka, 2008). Despite these complications, artificial agents will need to be able to make moral decisions. I will make a distinction between personal autonomy and moral autonomy. Personal autonomy means to formulate one's own moral law and act upon one's own will. With moral autonomy, I mean the ability to take moral considerations into account during autonomous decision making. The aim of this chapter is to provide a model of moral autonomy in artificial agents that will meet the

requirements to fulfil the responsibility of designers.

Types of Moral Agency

James H. Moor makes a distinction between four different types of moral agents: *ethical impact agents*, *implicit ethical agents*, *explicit ethical agents* and *full moral agents* (Moor, 2006).

Ethical impact agents are agents whose function have an ethical impact on the society and environment they are deployed in. This impact does not come from self-directed actions by the artificial agent, instead they strictly derive from the function of the agent and the manner in which it is deployed. An example would be an artificial agent that replaces labourers who work under dangerous conditions. Its ethical agency can be considered as beneficial for labourers since they no longer have to work under dangerous conditions. At the same time, the ethical agency of such robots can be considered as harmful because they put human labourers out of work. The ethical implications are of an indirect nature. The functions of these artificial agents have moral consequences but these consequences do not come from self-directed action. Therefore they do not have moral responsibility. Instead, human beings that decide to deploy these agents have moral responsibility for its consequences.

The second type of moral agents are implicit ethical agents. They are programmed in order to limit them to perform actions that are morally acceptable. Artificial agents of this type are prohibited from performing actions that are unethical. Furthermore, these agents have limited autonomy since they only execute tasks for which they are programmed. Consequently, these artificial agents cannot be held responsible for their actions, only their users or designers. An example of an implicit ethical agent are the automatic pilots in aircraft. They are programmed to operate within pre-defined parameters that are considered to be safe, and cannot choose to do otherwise. The actions of implicit ethical agents can only have unethical consequences when they are malfunctioning. There are various imaginable causes that can lead to malfunctioning in this type of moral agent. It can be caused by designer error, a software virus or ascribed to an unavoidable accident. Depending on the cause of an unethical outcome, responsibility can be ascribed to the designer, the user or to no-one if the unethical outcome really was unavoidable. The artificial agents themselves, since they are limited to perform morally acceptable actions, cannot be held responsible for any unethical outcome.

The third type of moral agents Moor identifies are explicit ethical agents. These agents are able to operate more autonomously and have the ability to make ethical judgements. However, they do not develop their own ethical framework. They have a particular pre-programmed ethical framework which they use in their decision making processes. An example is a robot in warfare that

is deployed to detect and assess potentially dangerous individuals. They consult their ethical framework and use it to determine if it is justifiable to attack this individual based on an assessment of how likely it is that this individual poses a threat. Another example is an artificial agent that is deployed after a natural disaster such as an earthquake. Its function is to search for survivors and assess which survivors need help the most and for which survivors help is most effective and likely to succeed. Explicit ethical agents use ethical considerations for their decision making. However, they are not themselves responsible for the nature and content of their ethical framework.

Finally, full moral agents develop their own ethical framework. Human beings are considered to be full moral agents. Capabilities such as consciousness, intentionality and free will are often considered to be the basis of being a full moral agent. Currently, artificial full moral agents have yet to be developed. There are arguments for and against the possibility of the existence of such artificial agents (Tonkers, 2009). In the following section, I will evaluate these arguments and I will argue what level of moral agency would be suitable and sufficient to allow the type of autonomous agent discussed in this thesis to be implemented responsibly in society.

The artificial agents discussed in this thesis require a degree of moral agency that is capable of making ethical judgements. The level of independence these artificial agents have in their employment, which include unfamiliarity and uncertainty, calls for a certain level of ethical reasoning. The artificial agent is required to assess the different moral outcomes of their actions, which include not taking action. They have to be able to have ethical reasoning processes that extent beyond the specific function they are developed for. According to Moor's model, this will require moral agency on the level of explicit ethical agents or full moral agents. I will continue this chapter with analysing what level of moral agency is desirable and will discuss the possibilities of this level of autonomy.

Full Moral Agency in Artificial Agents

One of the main differences between an explicit moral agent and a full moral agent is how the ethical framework is determined and developed. The ethical framework of explicit moral agents is decided *top-down*, as it is programmed and implemented by the designer. In a top-down approach, the ethical framework is decided upon by the designers of the agent and is explicitly specified in theoretical terms. The framework consists of ethical rules and principles that determine how to act in any situation. From the framework, which may be described in fine detail, the artificial agent should be able to derive how to act in specific situations. This framework determines what the moral values of the artificial agent are and how it should act ethically in certain circumstances. This creates a huge responsibility for the designers of artificial agents since they are

in charge of providing appropriate moral values to the artificial agent. The top-down approach is opposite to with full moral agents and a *bottom-up* approach. With a bottom-up approach, artificial agents are supposed to develop their ethical framework themselves from learning and experience. Human agency is often understood as full moral agency with a bottom-up approach. However, there are important factors that reduce the strength of this comparison.

Human beings do not shape their ethical framework entirely by themselves. They are heavily influenced by family, society, religion, role models and circumstances in their lives. For example, children are punished by their parents or teachers when they are believed to act unethically and there are numerous stories and books that are supposed to help shape moral values in children. However, human beings are able to reflect upon their ethical beliefs and are able to alter them. Emotions are important for human beings to determine what their goals are and play an important role in the ethical considerations of human beings (Ziemke, 2008). In human beings, strong emotions and desires can play a role in both undermining and underlining their ethical beliefs. Another example is a human being who acts out of sympathy to help another living being. How someone imagines others will respond to one's actions and how he or she imagines the perception and image others have of him or her can contribute to how someone acts as well. Artificial agents do not possess these emotions and therefore cannot act upon them. This tension between emotions and desires and ethical beliefs is absent in artificial agents. The way in which artificial agents shape their ethical framework is therefore to a large extent not comparable with how human beings shape their ethical framework. Furthermore, the actions of artificial agents are not determined by self-preservation. Their moral values are not challenged by their own desires and needs. As artificial agents do not posses any desire, the development of their ethical framework is to be solely based on computations. Furthermore, human beings place a certain value in themselves and in their own survival. This value has the capacity to extend to other human beings, living beings and the environment. In artificial agents, this value needs to be programmed or learned through a pre-programmed model. It is questionable whether artificial agents with pre-programmed ethical rules can still be considered to be full moral agents, since a requirement for full moral agents is to develop one's own ethical framework.

In addition to this, ethical frameworks of human beings differ amongst each other. Actions accepted by some individuals may be considered unethical by others. For artificial agents, the goal is to keep diversity in ethical beliefs to a minimum as it is undesirable to have artificial agents in society with unusual ethical beliefs. Human beings would want them to follow what are considered to be acceptable ethical beliefs within their own society. Furthermore, if all artificial agents possess the same ethical framework, it will be easier for human beings to place a certain amount of trust in

any artificial agent they might encounter. If a person knows what the ethical framework of another person is, he or she is generally more inclined to place trust in this person compared to a stranger. This argument can be extended to artificial agents. If their ethical framework is known, it is easier to trust them.

Ideally, human beings would want the ethical frameworks of artificial agents to be comparable with general ethical understandings in society. This is not straightforward because ethical understandings differ between groups and individuals within the same society. However, because artificial agents operate in particular, restricted functions, some consensus can be reached on how they should act. For example, that a human life has value is a generally accepted belief and certainly something we would want artificial agents to agree with. I will further examine why the conception of generalized values is complicated when they are implemented in artificial agents in the following section. For this part, I wish to focus on some of the difficulties that can arise when an artificial agent has the freedom to develop its own ethical framework, in the sense of having full moral agency. For example, a military robot could develop an ethical framework that prohibits it from harming any human being. It has developed an ethical understanding that it is wrong to harm human beings and refuses to do so. Even thought this may be a justifiable understanding, such artificial agents are unlikely to be employed for military purposes by any nation. Another example would be that an artificial agent decides that human beings are most harmful to the environment. It can therefore decide to stop aiding human beings or even start to work against them.

It is unlikely that robots with full moral agency, if they existed, would be deployed in any society. It is unlikely that human beings would allow this type of artificial agent because of the fear of losing control over them. The ethical framework of such artificial agents might diverge too far from the existing ethics in society for human beings to accept. Depending on an artificial agent's conception of what is important, they might perform actions that are not considered desirable by human beings. Additionally, it might be a possibility that the ethical framework of artificial agents becomes incomprehensible.

Another major complication with full moral agency is that there is the possibility to choose to act immorally. This is undesirable in artificial agents that are developed to perform tasks in society in order to aid human beings. One of the problems with full moral agency in artificial agents is that their ethical framework becomes opaque. This results in a lower level of trust that human beings place in artificial agents compared to trust placed in other human beings. Even though human beings are considered to be full moral agents as well, they are three reasons why placing trust in the framework of a human being is easier than in an artificial agent with full moral agency. First, we are able to trust other human beings because of the relationship we have with them or the

role they play in society. For example, someone knows one of his family members and therefore trusts him or her, and one knows that a police officer has to represent a certain ethical framework in performing their role. Secondly, we can make a comparison between other human beings to ourselves, including their ethical frameworks. In general, human beings have the ability to feel empathy, which is recognised by another person and this result in a certain level of trust. We are able to get an idea of someone's thoughts and intentions by communicating on different levels, such as verbally and through facial and bodily expressions. There are exceptions in which these mechanisms are misguided, either intentionally or unintentionally. However, this can be viewed as misguided trust in another person. Finally, it is possible to penalise human beings for unethical behaviour, such as a monetary fine or a prison sentence, for example. This makes it possible for people to place some trust in strangers. It is very complicated to punish artificial agents, since they would be required to be able to have emotions. These three reasons underline the differences in trusting an artificial agent compared to trusting a human being.

Apart from practical complications that may result from artificial agents with full moral agency, it is questionable whether full moral agency is realistically possible in artificial agents that are deployed to execute specific tasks. The discrepancy in artificial agents between their level of autonomous functioning and their level of autonomy in determining their own goals is problematic when one wishes to equip these artificial agents with full moral agency. The ability of full moral agency might have a negative impact on the functioning of artificial agents that are designed to fulfil specific tasks. In contrast with human beings, goals in artificial agents are externally determined. We, as human beings, would want them to complete the task we order them to do. Full moral agency is inefficient for artificial agents with goals that are externally determined. It appears problematic to have artificial agents develop their own ethical framework and at the same time let their tasks be determined by others. Such an agent would have its goals determined externally, and would only use its ability of full moral agency to determine what actions are ethically most desirable in order to reach that goal.

It is very complicated to ascribe responsibility for the actions of an artificial agent that has a high level of autonomy, such as artificial agents with full moral agency. If an artificial agent operates autonomously and has its own ethical framework, it becomes difficult to ascribe responsibility to the designer of the artificial agent. Additionally, it is very complicated to hold the agent itself responsible. The artificial agents needs to have certain capabilities that are currently not developed. These capabilities include, but are not limited to, a will of its own and being able to have emotions. As discussed previously in this section, it is undesirable to deploy artificial agents with their own free will in society. Emotions are required for an artificial agent to feel responsibility and

for human beings to be able to hold it to its responsibility by imposing penalties on it when it breaks the rules.

Programmable Ethics

Instead of full moral agency, a top-down approach to moral agency may be more appropriate for artificial agents. In this case, a top-down approach means that the moral laws of an artificial agents have been imposed upon it. This approach would result in an ethical framework in artificial agents that resembles that of an explicit ethical agent in Moor's categorization of ethical agents. Even though this approach may be less complicated on a technical level, several problems of a different nature occur. It is difficult to come up with an ethical framework that represents the moral values of society. Not every member of the same society has the same moral values. However, despite differences in ethical beliefs, there are generally sufficient shared values between members that allow a society to function properly. For example, virtually all members of a society would agree that murder is wrong. This resembles the idea of overlapping consensus by John Rawls (Rawls, 1987). According to Rawls, the goal of political philosophy is to formulate a conception of justice based on ideas of the citizens of society. Overlapping consensus means that citizens with different normative beliefs can come to an agreement about the principles of justice for their society, despite the different justifications they may have for accepting those principles

Another complication with a top-down approach in ethics is that it demands a very detailed delineation of moral values and ethical rules to be programmed into artificial agents. Because the ethical framework needs to be programmed beforehand, it needs to be more concrete and visible than ethical beliefs usually are. Differences in ethical beliefs between members of society come to stand out more when an agreement has to be reached for the ethical rules and beliefs of artificial agents. Not only is there disagreement about what artificial agents should do in particular situations, people's opinions change depending on their role in specific scenarios. In a recent study by Bonnefon, Shariff, and Rahwan, participants where asked about their opinion regarding whether an autonomous car should be programmed to sacrifice its passengers to save the lives of a larger amount of pedestrians (Bonnefon, Shariff & Rahwan, 2016). The results showed that participants were more willing to let the car sacrifice its passengers when they themselves were not an owner of an autonomous car. Moreover, they results depended on how many people were in the car and what the relation was to the participants. For example, people were less willing to let a car sacrifice its passengers to save others when their children were passengers in the car as well. I will discuss variations of these ethical dilemmas in more detail later on in this chapter.

Even if a common consensus would be reached about ethics or if just a single person would

have the authority and responsibility for developing an ethical framework for artificial agents, it would still be very complicated to reduce ethical beliefs to algorithms. There are many different understandings of algorithms (Moschovakis, 2001). In general, an algorithm is a set of rules that determines operations in order to create a desirable output based on input. They are most commonly used for computer programs, however, they have a variety of applications. What is most important for the argument in this chapter is that moral decisions need to be determined beforehand, and need to be made based on a general formula. These two factors make it very complex to decide upon how an algorithm should be constructed. The complexity of deciding upon algorithms that are most suitable in a wide variety of situations is comparable with how laws are constructed and implemented. Laws are generally carefully defined and often agreed upon by citizens. However, judges are still required to interpret the laws. They take specific circumstances into account when they are required to make a judgement in cases where there is uncertainty over whether (or how) the law applies.

Considerations about ethics in artificial agents are not just necessary when it comes to dangerous situations. It is commonly understood that artificial agents that are deployed in warfare or are used after natural disasters have to make moral decisions. However, ethical decisions need to be made in situations that are not necessarily considered dangerous as well. Artificial agents that operate in every day situations are also required to make ethical decisions in potentially dangerous situations that require rapid decision making to avoid or reduce damage (Goodall, 2014). In the case of cars, human drivers make decisions that are in principle against traffic laws in order to make manoeuvring through traffic easier or even safer, such as, swerving over traffic lines to let pedestrians or cyclists pass easier, for example. Therefore, ethical considerations are common in artificial agents that do not necessarily operate in dangerous situations on a daily basis.

If artificial agents were to operate according to algorithms that are considered the most optimal representation of moral values, difficult decisions about priorities have to be made. W.D. Ross argued in favour of ethical intuitionism in his influential work *The Right and the Good* (Ross, 1930). Ethical intuitionism argues that that human beings can rely on ethical intuition to make a decision in the case of conflicting ethical principles. According to ethical intuitionism, intuition comprises the basis of ethical judgement by human beings. Artificial agents do not posses these intuitive capabilities.

An example of an artificial agent that is required to set priorities is an artificial agent that is developed to provide help to victims of natural disasters. The envisioned job of such a robot would be to locate, provide first aid and transport victims to safer locations and medical treatment facilities. In large natural disasters, where the number of victims exceeds the amount of help

available, not every victim can receive the help he or she needs. In order to be as effective as it can be, the artificial agent needs to be able to identify which victims most require medical aid and for which victims medical aid is most effective. This would include leaving victims behind with minor injuries and leaving victims behind for which medical aid might be futile. These decisions are morally very difficult, especially when one considers that the artificial agent is probably incapable of making perfect diagnoses. This argument counts for human beings as well, especially in emergency situations. However, if these artificial agents are deployed, and their ethical framework has to be decided beforehand, how the agent operates in these scenarios need to be decided beforehand as well. It requires certain algorithms that allows it to operate in these situations. Setting priorities for medical aid during disasters is not just a problem for artificial agents. There is an ongoing ethical discussion about the allocation of health resources in situations with an imbalance between needs and supplies. Triage is used to categorize victims solely based on the victims' medical condition (Manger, Domres, Koch & Becker, 2001). It is developed to allocate medical resources as efficiently as possible. On a certain level, this type of categorization resembles an algorithm, as the course of action is determined based on a system that is constructed beforehand.

In military robots, equally difficult moral dilemmas can be conceived. Artificial agents deployed in warfare need to be able to assess how dangerous an individual is. If it decides that the individual is a threat, the artificial agent is probably programmed to attack the individual, with possible lethal consequences. If the artificial agent decides that the individual is not dangerous, it is probably programmed in such a way that it will not take action, except for remaining cautious. There are severe consequences if the assessment of the artificial agent is wrong. The artificial agent needs to be able to make solid evaluations of dangerous individuals and even if it has that capability, it would still be very complicated to reduce moral decisions that involve the lives of human beings to algorithms.

These problems emerge for autonomous vehicles as well. One of the suggested solutions for the avoidance of crashes by automated vehicles is that a human driver should take over control over the vehicle in situations that can lead to a collision. This would require that the human driver pays constant attention to what is happening on the road. Research has shown that human drivers lose focus when they are not required to operate the car constantly (Llaneras, Salinger & Green, 2013) (Jamson, Merat, Carsten & Lai, (2013)). Even thought this research has not been conclusive, it seems unlikely that this kind of semi-autonomous vehicle is a safe alternative for completely autonomous vehicles. Moreover, this would reduce one of the benefits of autonomous car, namely that people inside the car do not need to pay attention to what is happening outside of the car.

To pre-program an ethical framework in artificial agents means that hypothetical scenarios

become real in the sense that artificial agents need an answer to them in order for them to act in any potential scenario that was previously merely hypothetical. In the case of autonomous vehicles, attempting to limit damage in unavoidable crash scenarios means that the car has to prioritize between different objects to avoid colliding with one or more particular objects. This process is referred to as *targeting* (Lin, 2015).

A famous example of a hypothetical, ethical problem is the trolley problem and its many variations (Thomson, 1976). The trolley problem is a hypothetical scenario in which a person is asked whether he or she would pull a switch on a train track to save five persons from being hit by a tram at the cost of another person dying from being hit by it. This specific scenario is unlikely to happen nor is it likely that an artificial agent will be asked to act in this specific case. However, there are numerous imaginable situations that can occur in real life that would demand an artificial agent to make decisions of a similar nature. For example, in the case of an unavoidable crash, an autonomous car has the option to swerve around a child and hit an adult. In many situations, people seem to prioritize the lives of children over the lives of adults. However, it is questionable whether society would want autonomous cars that automatically target adults in order to avoid children.

Another problem that needs to be solved is whether the car should prioritize the safety of its own passengers or that of the human beings outside of the car. Should a car drive off a cliff to avoid colliding with another car or should it take the risk of hitting the other car, which contains a family? (Lin, 2015, p. 76) A similar dilemma is whether a car should attempt to hit safer cars, in the event of an unavoidable crash, in order to reduce damage to passengers of the car (Lin, 2015, p. 72). Targeting a bigger, safer car is relatively better for the passengers in the car that is getting hit, but targeting a smaller car is safer for the passengers in the car that is deciding what car to crash into. This problem extends beyond these particular situations. If it would be agreed upon that safer cars should be targeted, one of the reasons for buying such a car, the safety it provides, would be mitigated because passengers of such a car would be more likely to get hit in situations in which a crash cannot be avoided. It seems unethical to diminish the safety of human beings because they bought a car with the particular motive that the car is safe. Another option is to agree that smaller cars should be targeted by autonomous cars in order to protect the passengers of the car that does the targeting. This would result in the obvious problem that smaller, less resilient cars would be targeted even more and these cars would therefore be incredibly unsafe. This option is unethical as well because passengers of cars that are most susceptible to damage are being targeted specifically because their car offers less protection and is therefore safer to hit from the standpoint of passengers of the car its collides with. Targeting is a very important and complex problem for which a decision needs to be made when top-down approaches are considered. Take for example an autonomous car

that is facing an imminent crash with either one of two motorcyclists and has to decide who to target (Lin, 2015, p. 73). One motorcyclist is wearing a helmet and the other one is not. For the autonomous car and its passengers it probably not does not matter who to hit in terms of impact. However, both of the motorcyclists will suffer from severe damage when they get hit by the car. The chances to survive the crash is higher for the one that is wearing a helmet compared to the motorcyclist that is not wearing a helmet. Therefore, it seems to make sense to target the motorcyclist that is wearing the helmet in order to cause as little damage as possible. However, this is unfair to motorcyclist who follow the laws and make the responsible decision to wear a helmet. Not only is this unfair from a justice point of view, eventually it may reduce safety in general for motorcyclists since less of them will wear a helmet. These examples are just a few of the numerous possible scenarios in which an artificial agents needs to make a solution. These scenarios exemplify the difficult decisions that designers and policy makers need to make for explicit ethical agents that are employed in an open, real-world, environment.

To summarize, deciding upon a top down approach for artificial agents results in a variety or problems and difficult decisions. Most of the problems derive from the fact that a very detailed description of moral values is required, and these detailed descriptions should be applicable and appropriate in many situations. Ethical values and decisions need to be reduced to algorithms, which makes previously hypothetical scenarios into real problems that need to be solved. Additionally, it will be very difficult to come up with a framework that is accepted by everyone.

A Hybrid Approach for Moral Agency in Artificial Agents

As discussed in the previous sections, both top-down and bottom-up approaches seem far from ideal for modelling moral agency in artificial agents. However, it is difficult to think of an alternative, since explicit moral agency represents the minimum requirements of moral agency for artificial agents with a high level of responsibility and full moral agency is the maximum of what is possible for moral agents in general. Therefore, hybrid approaches are suggested (Allen, Smit, & Wallach, 2005). The general idea is that values are implemented in an artificial agent using a top-down approach. Additionally, the agent is in the possession of computational systems which are capable of taking into account many different inputs, including the implemented values. Methods for machine learning, such as neural networks and learning algorithms, should aid in acquiring moral intelligence in an artificial system. This approach resembles how ethical beliefs are shaped in human beings. A child has genetic information and learns principles that contribute in formulating a basic ethical framework. She keeps learning, is able to reflect on her own ethical beliefs and can therefore alter these beliefs based on experiences. This approach has complications on its own. For

example, emotions and consciousness play a role in constructing ethical beliefs in human beings. Moreover, a framework that involves learning mechanisms in combination with general ethical principles requires elaborate reasoning capacities. The agent must have knowledge about the ethical consequences of its actions. The ethical consequences of these actions are measured by the ethical values that are implemented. At the same time, however, the artificial agents has to interpreted the ethical consequences of its actions and this interpretation can alter the ethical values that are implemented.

An agent needs to be able to reflect on its own actions based on the ethical consequences of those actions. The main problem of a top-down approach is that it requires an extremely detailed description of an ethical framework. Even though the hybrid approach circumvents this problem, it still requires a moral model that allows the system to reason about the consequences of its own actions based on its implemented ethical values while the end product of this reasoning can alter these implemented ethical values as well. Therefore, it needs a method for evaluation and ethical reasoning. Two of the main general methods for ethical evaluation are deontology and consequentialism.

In deontological ethics, actions are to be assessed based on a set of rules or principles. These principles and rules should guide actions and choices based on whether they are morally required, forbidden or permitted. There are two main approaches for the implementation of deontological ethics in artificial agents. One is the direct implementation of a list of actions that an artificial agent is never allowed to perform. A famous accounts of this type of deontological ethics that is relevant to the topic of this thesis are the three laws of robotics by Isaac Asimov (Weld & Etzioni, 1994). The first law of robotics is that an artificial agent is never allowed to harm a human being or allow a human being to be harmed through inaction. This is a rule that would be generally agreed upon. However, this will result in a model that is very inflexible. In some cases, such as targeting, an artificial agent has to make a decision that are impermissible according to the principle. Moreover, directly implementing deontological principles will eventually result in a variation of a top-down approach. Another approach for deontology is to use a more abstract approach, such as Kant's categorical imperative (Johnson, & Cureton, 2016). According to the categorical imperative, all agents should act according to an unconditional moral law. This law applies to all agents and is independent of desires or personal motives. In any situation of moral choice, agents should act according to a principle that they would want everyone to act according to and that this principle becomes a universal law. This type of approach attempts to guide all action based on a single principle. This is such an abstract approach that it will result in computational difficulties, because it should be able to take into account the goal of its own actions and should be able to assess the

behaviour of others who may or may not act based on that same principle. For example, if this system is implemented, all artificial agents will act according to this principle, however, human beings may not. It will become extremely complex for an artificial agent to anticipate behaviour of other agents, as well as the effects of the behaviour of other agents on the actions and goal of the agent itself, based on one abstract principle.

In consequentialism, the consequences of an action determine whether an act is morally right (Sinnott-Armstrong, 2015). The problem with consequentialism is that there is an endless sequence of consequences that have to be computed by the artificial agent. There are three aspects to this problem. First, it is impossible to foresee all the direct and indirect effects of an action. Secondly, it will take immense computational power to calculate all the possible effects of an action, especially since an action can have an endless string of consequences. Thirdly, all possible effects should have a certain value that should be represented within the artificial agent.

I suggest a combination of deontology and consequentialism. In this approach, certain deontological values are implemented in an artificial agent. These include general principles regarding what are wrong outcomes, such as harming human beings. These principles are supposed to guide the agent in its consequentialist computations regarding the outcomes of anticipated actions. This will limit the required computations because certain actions are not allowed and therefore its possible outcomes do not have to be calculated any further. However, the deontological principles can be overruled if, and only if, the only possible actions are against these principles, such as in the case of targeting, for example. The artificial agents should then fall back to its consequentialist computations to assess what outcomes, from those that require actions that go against the deontological principles, have the least negative consequences. This approach will avoid endless computations by artificial agents by initially prohibiting actions that conflict with deontological principles, thereby limiting possible actions and consequently limit the number or calculations about the outcome of those actions. At the same time, it avoids the inflexibility that deontological approaches impose on artificial agents. This will allow an artificial agent to have capacity to learn, through consequentialism, and at the same time can be controlled and guided by deontological principles. Moreover, this approach allows machine learning on two different levels. One for normal situations and one for crisis situations in which the deontological principles are overruled.

For example, an autonomous car has as deontological principles: never harm a human being and never collide with a trash can. This will limit the consequentialist calculations by not having to calculate the consequences of performing any of those actions. Additionally, an artificial agent can learn how to operate in normal situations and learn how to avoid ending up in situations in which

deontological principles are overruled. However, when the only available options go against the deontological principles, we would want the autonomous car to hit the trash can. In these crisis situations, the artificial agent uses different decision making processes that are based on a separate learning mechanism. This separate consequentialist learning mechanism will guide the artificial agent when deontological principles are overruled. In short, the deontological principles should prohibit certain actions and limit the amount of calculations while consequentialist aspects allow the artificial agent to develop its moral evaluations by machine learning mechanisms. In the previous chapter, I have discussed how anticipation is important for autonomous decision making. The evaluation regarding whether a goal is reached and moral evaluations of the outcomes should both play an important role in the the selection of the most desirable future state of artificial agents.

Conclusion

In this chapter I argued that more artificial agents are operating in real world environments, perform increasingly complex tasks and have more responsibility. Therefore, they should be able to make justifiable moral decisions. However, designing artificial agents with these capabilities is extremely complex.

There are two main approaches available for designing a model for moral decision making in artificial agents. One is a bottom-up approach, in which artificial agents learn to make moral decisions based on experience. This is the approach that can result in full moral agents. There are a number of problems that make it difficult if not impossible to deploy this type of artificial agent in society. It is currently not possible, from a technical perspective, to design these artificial agents. Moreover, these artificial agents are probably undesirable for several reasons: it can be difficult or impossible to control these agents, their ethical framework might deviate to far from what is common or accepted in a society and it can become complex or impossible to have artificial agents with full moral agency fulfil specific tasks, among other reasons.

The second approach a top-down approach, would result in designing explicit ethical agents. The problem of this approach is that a complete all-encompassing ethical framework needs to be developed and implemented. This will require specific ethical situations to be resolved by algorithms. Pre-determined ethical rules will determine the moral decisions of an artificial agent. One of the problems with this is to reach an agreement about the ethical framework. A second problem is that algorithms for moral decisions need to be made beforehand. This will result in the necessity of a very detailed description regarding what an artificial agent has to do in a specific situation. This requires decisions regarding scenarios with ethical dilemmas that were previously mere hypothetical scenarios for which a proper ethical response is still debated.

Instead, I suggest the use of a hybrid approach for morality in artificial agents. This combines a top-down approach in terms of deontological principles and a bottom-up approach by consequentialist learning mechanisms. This should result in artificial agents that are both controllable and adaptable.

A Model of Autonomy for Artificial Agents

Overview of the Conclusions of Previous Chapters

The goal of this thesis is to construct a adequate model of autonomy for artificial agents that fulfils the designer's moral responsibility. I have argued in the chapter about moral responsibility of designers that artificial agents need to have decision making processes that is at least up to the standards of human decision making. This is because the moral principles that designers should follow state that new methods to fulfil a certain function cannot be more harmful to society than the methods it is replacing. In this case, artificial agents are replacing human beings. Therefore, the decision making processes of artificial agents should be at least up to the standards of human beings.

In the second chapter I argued that behavioural autonomy is the most important aspect of autonomy for artificial agents. The requirements of autonomy for artificial agents differ from those of biological organisms such as human beings. What is most important for artificial agents is that they are able to autonomously complete tasks that are determined by human beings. Therefore, they require a limited type of autonomy. Artificial agents need decision making processes that allow them to make justifiable decisions.

I have described the importance of anticipation and representation in the third chapter. It is important for artificial agents that operate in an open and uncontrolled environment to be able to represent moving objects and multiple objects at the same time. Event calculus is a promising type of logic for this type of representation and, additionally, might be a partial solution for the frame problem. This type of representation should allow for the anticipation of moving objects. An artificial agent should be hyperincursive because it should anticipate multiple future states and select the most desirable state. From there, an agent can plan its actions based on calculations of probability and utility of certain states.

In the fourth chapter I described what is required from the moral autonomy of artificial agents. I have argued that both full moral agency for artificial agents in a bottom-up approach and a top-down approach are not suitable. Instead, I suggested a hybrid approach to combine the positive aspects while attempting to bypass the negative aspects. This approach combines deontological principles to guide and control the artificial agent with consequentialist learning mechanisms in order to have a flexible framework for artificial agents.

A Model of Autonomy for Artificial Agents

The intended role of artificial agents in society is to complete tasks in an uncontrolled

environment. Moral principles of designers state that the implementation of a new method or technology cannot be more harmful to society than the method or technology it is replacing. Artificial agents discussed in this thesis are meant to perform tasks that are currently performed by human beings. According to the moral principles of designers, artificial agents are required to execute those tasks as least as well as human beings. The artificial agents discussed in this thesis operate in an open an uncontrolled environment. They need to be able to make justifiable decisions autonomously in order to reach their goals, which are determined by human beings. Therefore, to fulfil the moral responsibility of designers of artificial agents, an adequate model of autonomy for artificial agents should allow them to make justifiable decisions.

The goals of artificial agents are determined by human beings, therefore, artificial agents are not required to authorise their own actions. This means that a limited account of autonomy is required and desirable. Artificial agents need an behavioural account of autonomy which allows them to make justifiable decisions in an open and uncontrolled environment. In order to to that, an artificial agent needs to anticipate multiple future states. From these states, it should attempt to reach the most desirable one and plan its actions based on reaching that state. Therefore, I suggest a hyperincursive system that anticipates multiple future states based on representations by event calculus. Calculating the probability and utility of certain outcomes allow an artificial agent to discriminate between different outcomes and select the most desirable one.

The decision making of artificial agents includes moral considerations. A combination between top-down and bottom-up approaches allows for the combination of the positive aspects of both approaches. A hybrid approach will keep the artificial agent controllable by utilizing deontological principles. At the same time, consequentialist machine learning mechanisms avoid the complications of programming an ethical framework in which every decision needs to be determined in detail beforehand. The artificial agent is limited to following the deontological principles in standard situations, in which machine learning based on consequentalism will further guide the decision making of artificial agents. In situations in which the only possible actions are against the deontological principles, these principles will be overruled and the agent will make decisions based on a separate consequentialist machine learning mechanism that will take over control in crisis situations. I suggest this model of autonomy for artificial agents in order to meet the demands of autonomous decision making in open and uncontrolled environments. I discussed that starting with analysing the role of autonomy of artificial agents can contribute to developing an adequate model of autonomy for artificial agents that fulfils the moral responsibility of designers. However, this model needs to be applied in order to determine whether it results in an adequate model of autonomy for artificial agents. Since it is specifically aimed at real-world scenarios, it is

difficult to determine outside of real-world situations whether this approach will indeed result artificial agents that can execute their tasks at least up to the standards of human beings.

Next to having a adequate model for autonomy, there should be methods to determine the decision making processes of artificial agents, especially in the case when there are undesirable outcomes. This will help determine which party is responsible and can contribute to the development of improved artificial agents. For autonomous cars, a suggestion is to record the autonomous behaviour of the car (Gitlin, 2016). These recording devises can aid in determining if a crash was avoidable or not. Moreover, when the crash could have been avoided by the car, they can provide insight regarding what aspect of the autonomous behaviour was the cause of the crash.

Shared representation and communication between autonomous car can also contribute to better decision making and safety. If autonomous cars share their position to other nearby cars, they are aware or each others position without the need of visual representation. Moreover, they can share other relevant factors of the environment, such as oncoming pedestrians, for example.

References

Allen, C., & Wallach, W. (2012). Moral machines: Contradiction in terms or abdication of human responsibility. In Robot Ethics: The Ethical and Social Implications of Robotics, (pp. 55-68). MIT Press, Cambridge (MA).

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3), (pp. 149-155).

Artikis, A., Sergot, M., & Paliouras, G. (2015). An event calculus for event recognition. *IEEE Transactions on Knowledge and Data Engineering*, 27(4), (pp. 895-908).

Association of Computing Machinery (ACM) (1992). "ACM Code of Ethics and Professional Conduct."

Astor, E., Ekdahl, B., Davidsson, P., & Gustavsson, R. (1991). Anticipatory planning.

Benčin, R., Strle, G. & Gurzawska, A. (2015). Principles and approaches in ethics assessment, social responsibility in science and engineering. European Commission's Seventh Framework Programme.

Bickhard, M. H. (2000). Information and representation in autonomous agents. *Cognitive Systems Research*, 1(2), (pp. 66-75).

Boden, M.A. (1996). Autonomy and Artificiality. In The Philosophy of Artificial Life, (pp. 95-108). Oxford University Press.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), (pp. 1573-1576).

Brandano, S. (2013). The event calculus assessed.

Brooks, R. A. (1991). Traditional Approaches.

Buss, S (2014). Personal Autonomy. <http://plato.stanford.edu/archives/win2014/entries/personal-autonomy/>

Charlesworth, A. (2014). The comprehensibility theorem and the foundations of artificial intelligence. *Minds and Machines*, 24(4), (pp. 439-476).

Chrisley, R. (2002). Some foundational issues concerning anticipatory systems. *International journal of computing anticipatory systems*, 11, (pp. 3-18).

Christensen, W., & Hooker, C. (2000). Anticipation in autonomous systems: foundations for a theory of embodied agents. *Int J Comput Anticip Syst*, 5, (pp. 135-154).

Collier, J. (2006). Conditions for fully autonomous anticipation. *In Computing Anticipatory Systems, CASYS'05(AIP Conference Proceedings)*, 839, (pp. 282-289).

Collier, J. (2008). Simulating autonomous anticipation: The importance of Dubois' conjecture. *BioSystems*, 91(2), (pp. 346-354).

Davidsson, P., Astor, E., & Ekdahl, B. (1994). A framework for autonomous agents based on the concept of anticipatory systems. *Cybernetics and Systems*, 94, (pp. 1427-1434).

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation?. *AI magazine*, 14(1), (pp. 17-33).

Dennett, D. C. (2006). The Frame Problem of AI. *Philosophy of Psychology: Contemporary Readings*, , (pp. 433-454).

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 68(4), (pp. 87-106).

Di Paolo, E. A., & Iizuka, H. (2008). How (not) to model autonomous behaviour. *BioSystems*, 91(2), (pp. 409-423).

Domres, B., Koch, M., Manger, A., & Becker, H. D. (2001). Ethics and triage. *Prehospital and disaster medicine*, 16(01), (pp. 53-58).

Dubois, D. M. (2003). Mathematical foundations of discrete and functional systems with strong and weak anticipations. *In Anticipatory behavior in adaptive learning systems*, , (pp. 110-132).

Ekstrom, L. W. (1993). A coherence theory of autonomy. *Philosophy and Phenomenological Research*, 53(3), (pp. 599-616).

Eshleman, A. (2014). Moral Responsibility. http://plato.stanford.edu/archives/sum2014/entries/moral-responsibility.
Evans, D. (2002). The search hypothesis of emotion. *The British Journal for the Philosophy of Science*, 53(4), (pp. 497-509).
Floridi, L. (2013). The Morality of Artificial Agents. In The Ethics of Information, (pp. 134-160). Oxford University Press.
Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, 14(3), (pp. 349-379).
Forge, J. (2010). A note on the definition of "dual use". *Science and Engineering Ethics*, 16(1), (pp. 111-118).
Forge, J. (2000). Moral responsibility and the 'ignorant scientist'. *Science and engineering ethics*, 6(3), (pp. 341-349).
Franklin, S., & Graesser, A. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In International Workshop on Agent Theories, Architectures, and Languages, (pp. 21-35). Springer Berlin Heidelberg.
Froese, T., Virgo, N., & Izquierdo, E. (2007). Autonomy: a review and a reappraisal. In European Conference on Artificial Life, (pp. 455-464). Springer Berlin Heidelberg.
Garriga, E. & Melé, D. (2004). Corporate social responsibility theories: Mapping the territory. *Journal of business ethics*, 53(1-2), (pp. 51-71).
Georgeff, M. P., & Lansky, A. L. (1987). Reactive reasoning and planning. *AAAI*, 87, (pp. 677-682).
Gillon, R. (1994). Medical ethics: four principles plus attention to scope. *Bmj*, 309(6948), (pp. 184-188).
Gitlin, J. M. (2016). Germany wants black boxes in self-driving cars. http://arstechnica.com/cars/2016/07/germany-wants-black-boxes-in-self-driving-cars/
Goodall, N. J. (2014). Machine ethics and automated vehicles. Springer International Publishing.
Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology*, 10(2-3), (pp. 115-121).
Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and information technology*, 15(2), (pp. 99-107).
Hersh, M. (2014). Science, technology and values: promoting ethics and social responsibility. *AI & society*, , (pp. 167-183).
Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: an ethical analysis. *Science and engineering ethics*, 21(3), (pp. 619-630).
Hookway, C (2016). Pragmatism. \url{http://plato.stanford.edu/archives/sum2016/entries/pragmatism/}.
Jamson, A. H., Merat, N., Carsten, O. M., & Lai, F. C. (2013). Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions. *Transportation research part C: emerging technologies*, 30, (pp. 116-125).
Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and information technology*, 8(4), (pp. 195-204).
Johnson, R. & Cureton, A. (2016). "Kant's Moral Philosophy". <http://plato.stanford.edu/archives/fall2016/entries/kant-moral/>
Kirsh, D. (1991). Today the earwig, tomorrow man?. *Artificial intelligence*, 47(1-3), (pp. 161-184).
Kowalski, R., & Sergot, M. (1989). Springer Berlin Heidelberg. In Foundations of knowledge base management, (pp. 23-55). .
Lavigne, F., & Lavigne, (2000). Anticipatory semantic processes. *International Journal of Computing Anticipatory Systems*, 7, (pp. 3-31).
Lin, P (2015). Why Ethics Matters for Autonomous Cars. In Autonomes Fahren, (pp. 69-85). Springer Berlin Heidelberg.
Llaneras, R. E., Salinger, J., & Green, C. A. (2013). Human factors issues associated with limited

ability autonomous driving systems: Drivers' allocation of visual attention to the forward roadway. *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, , (pp. 92-98).
Maturana, H., Varela, F., (1980). Autopoiesis and Cognition: the Realization of the Living. Reidel, Dordecht.
McKenna, M. S. (2000). Assessing reasons-responsive compatibilism. *International Journal of Philosophical Studies*, 8(1), (pp. 89-114).
Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media..
Moor, J. M. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems, IEEE*, 21(4), (pp. 18-21).
Moreno, A., Etxeberria, A., & Umerez, J. (2008). The autonomy of biological individuals and artificial models. *BioSystems*, 91(2), (pp. 309-319).
Moschovakis, Y. N. (2001). What is an algorithm. *Mathematics unlimited–2001 and beyond*, 2, (pp. 919-936).
Nilsson, N. J. (1984). Shakey the robot. SRI INTERNATIONAL MENLO PARK CA.
Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, 16(1), (pp. 51-62).
Powers, T. M. (2013). On the Moral Agency of Computers. *Topoi*, 32(2), (pp. 227-236).
Rawls, J. (1987). The idea of an overlapping consensus. *Oxford journal of legal studies*, 7(1), (pp. 1-25).
Reath, A. (2013). Kant's Conception of Autonomy of the Will. In Kant on Moral Autonomy, (pp. 33-52). Cambridge University Press.
Rohde, M., & Stewart, J. (2008). Ascriptional and 'genuine'autonomy. *BioSystems*, 91(2), (pp. 424-433).
Rosen, R (1991). Life itself: a comprehensive inquiry into the nature, origin, and fabrication of life. Columbia University Press.
Rosen, R. (1985). Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations. Pergamon Press.
Ross, W. D. (1930). The right and the good. Clarendon Press Oxford.
Rosslenbroich, B. (2009). The theory of increasing autonomy in evolution: a proposal for understanding macroevolutionary innovations. *Biology & Philosophy*, 24(5), (pp. 623-644).
Russell, S., Norvig, P. (2014). Artificial intelligence: a modern approach (Vol. 3). Pearson.
Schmidt, C. T. A., & Kraemer, F. (2006). Robots, Dennett and the autonomous: a terminological investigation.. *Minds and Machines*, 16(1), (pp. 73-80).
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(03), (pp. 417-424).
Shamoo, A. E., & Resnik, D. B. (2009). Responsible conduct of research. Oxford University Press.
Shanahan, M. (1999). The event calculus explained. In Artificial intelligence today, (pp. 409-430). Springer Berlin Heidelberg.
Sinnott-Armstrong, W (2015). "Consequentialism". <http://plato.stanford.edu/archives/win2015/entries/consequentialism/>
Skarlatidis, Paliouras, Artikis, & Vouros (2015). Probabilistic event calculus for event recognition. *ACM Transactions on Computational Logic (TOCL)*, 16(2), (pp. 11-46).
Sullins, J. P. (2006). When is a robot a moral agent?.
Thompson, K. (2014). Computer chess strength. *Advances in computer chess*, 3, (pp. 55-56).
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59(2), (pp. 204-217).
Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines*, 19(3), (pp. 421-438).
Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), (pp. 433-460).

Van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and engineering ethics*, 19(2), (pp. 407-433).
Vernon, D., Lowe, R., Thill, S., & Ziemke, T. (2015). Embodied cognition and circular causality: on the role of constitutive autonomy in the reciprocal coupling of perception and action. *Frontiers in psychology*, 6, (pp. 1-9).
Walton, D. (2016). Some Artificial Intelligence Tools for Argument Evaluation: An Introduction. *Argumentation*, 30(3), (pp. 317–340).
Wardziński, A. (2006). The role of situation awareness in assuring safety of autonomous vehicles. Springer Berlin Heidelberg.
Weld, D., & Etzioni, O. (1994). The first law of robotics (a call to arms). *AAAI*, 94, (pp. 1042-1047).
Yadron, D. & Tynan, D. (2016). Tesla driver dies in first fatal crash while using autopilot mode. https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk
Ziemke, T. (2008). On the role of emotion in biological and robotic autonomy. *BioSystems*, 91(2), (pp. 401-408).