Kjetil Holtmon Akø

Understanding the autonomy of autonomous technology

Master's thesis in Philosophy

Department of Philosphy and Religious studies

Faculty of Humanities

Norwegian University of Science and Technology (NTNU)

Spring 2016

Kjetil Holtmon Akø

Understanding the autonomy of autonomous technology

Supervisor: Rune Nydal

Co-supervisor: Ingrid Schjølberg

Master's thesis in Philosophy

Department of Philosphy and Religious studies

Faculty of Humanities

Norwegian University of Science and Technology (NTNU)

Spring 2016

Table of content

Acknowledgements

I am greatly indebted to my supervisor, Associate Professor Rune Nydal, for his insightful and guiding remarks throughout my writing of this thesis. His patience and thoughtful feedbacks have been indispensable during the process of completing this thesis, and without him the thesis would certainly lack in quality and consistency. I would also like to direct my gratitude towards my co-supervisor, Professor Ingrid Schjølberg, whose knowledge and directions have been greatly appreciated in order for me to get a better understanding of the technological aspects of this thesis. I am also grateful for the time I was able to spend at the AMOS center. To work alongside the technologists was highly motivational and allowed me ask questions and confirm other thoughts.

To all of my fellow students, both graduates and bachelor student, I want to express my sincere gratitude. Without you this year would surely been a bleak one. The social, friendly, yet academically stimulating atmosphere has contributed highly to making the process of writing an enjoyable one. I want to especially thank Rolf and Arnstein for their thoughtful remarks during our group sessions and our exchange of information and papers.

Lastly, I want to thank my family. Without them supporting me on this journey of philosophy I would never had made it. They have always support my choice of study and have always welcomed me home, making sure I was fit for fight when I went back to study. To my mother, my brother and my sister – I am truly grateful for your support.

Abstrakt

Nylig har autonom teknologi fått medieoppslag og det er visse bekymringer knyttet til denne teknologien. Denne oppgaven søker å svare på spørsmålet om hvordan man skal forstå autonomien til autonome teknologier. For å svare på dette har oppgaven tatt utgangspunkt i teknologenes forståelse og beskrivelse av teknologien, og dette har så blitt knyttet opp til en helhetlig teknologifilosofi. Dette innebærer at autonomien til autonom teknologi må forståes som automatisering av funksjoner og bruken av taksonomier klargjør hvilke funksjoner som er automatisert og hvordan dette forholder seg til en operatør. Dette betyr at teknologisk autonomi ikke er én idé, og at denne autonomien best beskrives stegvis. På basis av denne forståelsen blir så to argumenter kritisert for å ha ikke fullt ut forstå hva teknologisk autonomi er. Noe av kritikken her berører også vesentlige forskjeller mellom moralsk og teknologisk autonomi. Disse to begrepene betegner noe helt forskjellig og må ikke oppfattes som likestilte. Videre, ved hjelp av Ihde sin postfenomenologi belyses så hvordan teknologien påvirker mennesket forhold til verden med blikk på hvordan autonom teknologi opptrer for oss i vår opplevelse av omverden. Måten autonom teknologi opptrer for oss på gjør oss tilbøyelig til å bruke antropomorfe begreper og beskrivelser for å forstå og gjøre oss kjent med teknologien, samt å kommunisere denne kunnskapen til andre. Disse begrepene og beskrivelsene er dog ikke presise nok til å kunne stadfeste hva en autonom teknologisk gjenstand kan og ikke kan gjøre, og antropomorfismen kan i verste fall tillegge teknologien egenskaper den ikke har. Oppgaven konkluderer med at antropomorfismen har en funksjon ved å gjøre oss kjent med teknologien, men at taksonomier er bedre egnet til å beskrive og formidle hva den kan og ikke kan gjøre.

1. Introduction

This thesis seeks to answer the question of how we should understand the autonomy of autonomous technology. A satisfactory answer to this question is essential in order to discern and solve any potential dilemmas related to this technology, and moreover an answer might settle some of the worries related to the use of this technology. Important in answering this question is to acknowledge and include how the technologists views and describes autonomous technology, as well as identifying cases of misconceptions and inadequate perspectives. In the last decade autonomous technology have raised worries in both scientific fields and the public, and this thesis seeks to quell some of these worries by offering a perspective on technological autonomy and a method of describing it.

The public has from time to time organized and raised their concerns when introduced to new technologies. In the 1800s, the luddites reacted to the industrialized factories by sabotaging the new machines. Later, when electricity was introduced, people were worried that electricity would leak out of the electric sockets. In more modern times the public has also reacted to other technological innovations such as the genetic manipulation of organisms and nuclear plants. Common to these worries are that they depart from the disruptiveness of new technology and forces us to come to terms with the new situation.

Worries directed at new technologies is not only a matter of civic discussions, but can also be considered a cultural expression. The societal changes of technology has and continues to be a prominent theme in western film and literature. Classics such as Blade Runner, 2001: A space odyssey and Brave New World deals with the effects of new technologies on society. These works engages the reader in fascination and reflection over the novelty and safety of new and futuristic technology. The disruptiveness of technology continues to fascinate and frighten us. Techno-optimism and –pessimism are two sides of the same coin.

Importantly, these worries are not exclusively those of the nonprofessionals. Earlier this year, Nobel Prize winner Professor Edvard Moser raised concerns about the speed with which artificial intelligence is developed (Adressa, 2016). He feared that the ethical and social discussions could not keep up with the technological research and development. Similarly, in January 2015, Future life institute published an open letter on future research priorities for artificial intelligence and autonomous technology, where they voiced concerns regarding the potential pitfalls of such technologies. Various prominent researchers such as Professor Stephen

Hawking and Professor Nick Bostrom have signed this letter. The worries in the letter reflects the possibilities of this technology, as well as the prediction that it will become ordinary in the future. The worries directed towards autonomous technology and AI departs from a perceived lack of control we will have over the technological artefacts. Instead of us choosing what the artefact does, this new technology is able to decide on a course of action independently of an operator.

In a sense, this worry is both peculiar and understandable. It is peculiar because autonomous technology is made with the intention of us pulling back and not having the same type of control as we would with non-autonomous technology. Likewise, it is understandable to be worried since we might feel like we do not have the same amount of control. While autonomous artefacts do their work without you or me being explicitly in control of it and directing its every move, non-autonomous artefacts demands human dexterity and input to work. While peculiar, it is nevertheless fully understandable because autonomous technology implies a greater distance and less (explicit) control over the artefact. Instead of being the active user of an artefact, we adopt the role of supervising the artefact's operations. Rather than the pilot being present in the plane (the plane itself has now become a drone), the operator sits comfortably several thousand miles away, mainly supervising the operation and making operative decisions.

The same reflections, worries and skepticism are reflected in various scientific fields dedicated to the social and ethical impacts of technology. A hot topic recently has been autonomous technology and its consequences. In the last decade, technology of this sort has blown up in terms of research and development, and few could have foreseen the capabilities that current technology possesses – probably even more so for the years to come. According to the strategic research agenda for robotics in Europe (SPARC), it is believed that autonomous technology will increase its presence in everyday life in a variety of ways. Predicted uses range from autonomous processing software in corporations to robotic aid in health care. Military robots, self-driving cars, autonomous vehicles on land, air and water – these are all areas of predicted use. In fact, according to a recent report by the World Economic Forum, within 2020, over five million jobs will be lost worldwide due to implementation of autonomous technology (WEF, 2016). In general, autonomous technology is believed to be highly effective, cost-efficient, safe and precise tools that will enable a variety of possible uses within a range of fields (SPARC, 2014).

The predictions of autonomous technology and its presence has lead researches to investigate the potential consequences of this technology – be it widespread use or area-specific use (such as care robots or self-driving vehicles), or how this new technology fits with our current practices. Sparrow (2007) and Matthias (2004) have argued that autonomous technology is in conflict with our contemporary notion of allocating responsibility and liability. Sparrow maintains that autonomous technology is in conflict with our concept of just war, while Matthias claims this new technology prevents manufacturers and operators from being in control. Concerned with the moral status of AI, Bostrom and Yudkowsky (2014) states that artificial intelligence could one day meet the criteria of moral agency. Floridi and Sanders (2004) and Floridi (2014), in a similar manner, argues that autonomous technology should be considered moral agents due to their internal mechanisms and the moral consequences they produce.

Contrary to this, Johnson (2015), and Johnson and Noorman (2014) has argued that the people involved retain full responsibility due to the practices of responsibility within technological development. Grodzinsky et al. (2008), employing the same method as Floridi (2004), has argued that autonomous technology should not be considered moral agents due to essential differences between a human moral agent and the autonomous artificial agent.

My thesis concerns itself specifically with the concept of technological autonomy and how the concept is understood and used when attempting to draw out the impacts of the technology. The diversity of conclusions and claims in discussions concerning autonomous technology suggests that this type of technology is not fully understood. I propose the use of taxonomies in discourses concerning ethical and social implications of autonomous technology. As a method of describing autonomy, taxonomies are already in use by the engineers and programmers working in the field, so it comes as a surprise that this is not always the case in the social and ethical discussions of this technology.

Taxonomies offers a varied and differentiated understanding of technological autonomy, which in turn makes us less reliant on metaphorical descriptions of the technological capabilities. A lack of a taxonomy leaves the concept underdeveloped and in effect forces us to comprehend the technology by virtue of metaphorical use of language. In this thesis, anthropomorphic descriptions is considered a metaphorical use of language, and when not stated otherwise I will use “metaphorical descriptions” and “anthropomorphic descriptions” interchangeably (denoting the same). These descriptions contributes to and reinforces the worries associated with autonomous technology because they do very little to illuminate the extent of an artefact's capabilities. Anthropomorphic descriptions ambiguously reveals aspects by constructing implicit analogies between the artefact and its counterpart.

There is a way in which the concept of technological autonomy leaves too much open to interpretation. As Noorman and Johnson states, "…machine autonomy is not a single idea." (2014, 60), implying that this concept might be too vague to be useful for deriving ethical and social consequences. In this respect, taxonomies shifts attention to automatization of functions and away from the concept of "autonomy". The use of "automation" instead of "autonomy" is also the preferred choice of engineers and programmers as found in one review (Vagia et al. 2016). At one point Noorman and Johnson refers to machine autonomy as "…the high-end of an increasing scale of automation" (2014, 57), while the Department of Defence's 2011 Roadmap states that autonomous systems are "…self-directed toward a goal in that they do not require outside control, but rather are governed by laws and strategies that direct their behavior." (Department of Defense, 2011, 43). In contrast to autonomous systems, the DoD defines automatic systems as "fully preprogrammed" (Ibid. 43), meaning that autonomous systems are not fully preprogrammed. Here, autonomy also signifies independence from an operator.

1.1 Structure of the thesis

The first chapter is dedicated to provide a platform for understanding technology and autonomous technology. I start with a clarification of the two distinct concepts of autonomy found in the literature; *moral autonomy* and *technological autonomy* (autonomy-as-automatization). I follow up by providing an account of technology, where I emphasize the context of use and the social practices as imperative to understanding technological systems, and this provides the basis for the critique in the next chapter. Additionally I draw attention to the moral or normative aspect of technology. I then propose taxonomies as the proper way to understand technological autonomy. This entails that the autonomy of technology should be understood as the automation of functions. It also shows how technological autonomy could be understood as describing the artefact's state of independence from an operator. The taxonomy I am proposing is a combination of one taxonomy previously put forth by Endsley and Kaber (1999) and Professor Perri 6's ladder of machine autonomy (2001). This perspective enriches the concept of technological autonomy by conceptualizing it in *levels* and in *types*, and thereby avoids treating it as a single idea. The aim is to develop a taxonomy that could be used to describe levels of robotic autonomy. Simultaneously I believe that the same taxonomy, with some revision, is also useful in describing general levels of technological autonomy.

The choice of robots as the target of this taxonomy is motivated by two reasons. Firstly, many of the internal mechanisms of an autonomous robot will most likely have similarities to non-robotic autonomous technology. This means that the functions that constitute the autonomy of a robot will likely be found in other autonomous technologies. Secondly, autonomous robots are coming. Substantial amounts of money, energy, and time is put into research and development of autonomous robots and in all likeliness, robots will be a common sight in the future.

In the following chapter I criticize two arguments in light of the proposed taxonomy and understanding of technology. These two arguments are used as examples of under-developed notions of technological autonomy, and are meant to exemplify how an excessive focus on the technological artefact leads one to dismiss the context of use and the social practices. First I review the arguments and claims that Matthias puts forth in his article *The Responsibility Gap* (2004). I argue that he has not properly accounted for the limits of technological autonomy, which I find to be result of the way he views technology and technological development. I then turn to Floridi and Sanders' arguments concerning the moral status of autonomous artefact. While they draw attention to the morality of things, they employ an analogy between humans as moral agents and the autonomous artefact that I find untenable. There are important differences between moral autonomy and technological autonomy that their argumentation does not account for in satisfactory manner.

In chapter three I employ a phenomenological perspective in order to investigate why we are liable to use certain descriptions and concepts when commenting on autonomous technology. The phenomenological account provides a perspective on how technology affects our experience of the world, and in this respect it offers a way of investigating how autonomous technology appears to us experientially. I argue that autonomous technology takes part in a distinct relationship to us, and that this relationship can explain why there is a tendency to anthropomorphize the technology. From this I argue that an anthropomorphic language (as a subclass of metaphorical descriptions) might have the adverse effect of obscuring our understanding. While metaphors are useful in terms of familiarizing us with the unknown and conveying knowledge, they simultaneously invite interpretation and impreciseness.

2. Understanding the autonomy of technology

Since this thesis in its entirety is situated within discourses of the ethical and social implications of autonomous technology, I find it necessary to first draw out some essential differences regarding two concepts of autonomy. While the concept has recently become a popular way to denote a type of technology, the concept also has a long and rich tradition within philosophy. This means that two distinct concepts of autonomy takes part in these discourses.

2.1 Two concepts of autonomy

In philosophy, autonomy often denotes a capability that all humans possess and is something worthy of respect. Not respecting the autonomy of an individual is considered the same as not respecting the individual's right of self-determination. The word itself, autonomy, comes from the Greek word *autonomos*, which consists of the words *auto* (self) and *nomos* (law), and is meant to capture the notion that we give ourselves the rules by which we act. In other words, being an autonomous individual means that you are in charge of yourself and can decide what to do. Seeing as how we are individuals in groups, autonomy is also related in some way to how we perceive ourselves as individuals in a society where our actions have consequences for others and ourselves.

Fischer and Ravizza (1998) has argued that moral autonomy is a reflective attitude that we possess. In their view, this means that we as moral agents are *receptive* and *reactive* to reasons. We are receptive in the sense that we recognize available reasons that motivate action. We are reactive in the sense that we translate the reasons into choices and actions. This implies that autonomy is related to what we perceive the world to be, how it *should* be, and how our actions affect the surrounding world.

Taylor (1985) has argued that language and the ability to evaluate our desires and inclinations is what makes us moral and responsible beings. His notion of first- and second-order desires have similarities with being receptive and reactive to reasons, though differs in the sense that Taylor's notion is meant to capture the self-making of an individual, while Fischer and Ravizza is mainly occupied with moral responsibility. According to Taylor we have immediate desires and inclinations – this is something we share with rest of the animal kingdom. Particular to us humans is the ability to evaluate our immediate desires and inclinations in what Taylor calls a *language of worth*. This means that we assign values to our desires. We cannot help judge whether our desires is something we want to have and whether they are something we should act on. Accordingly, we have second-order desires – desires about the desires we have.

For Taylor this means that we have a particular responsibility not found elsewhere in the animal kingdom – morality is something only found in human beings. The notion that we evaluate our first-order desires means that the second-order desires comes about due to us willing them into existence through articulating descriptions about insights of what is important. In other words, we exercise our will against the first-order desires – we are self-evaluating beings. Since the insights by which we evaluate our desires can be distorted, we become responsible in the sense that we must continuously reflect on these insights.

The idea that we evaluate our desires in a language of worth also implies that our desires and actions are potentially open to dispute by others. A belief that something should be a certain way entails that the same belief can be questioned and disputed. Since reasons, beliefs and claims are most often expressed through language, and language is furthermore essential to any discussion of normativity, then it is hard to conceive of a moral agent that does not have the ability to understand and use language. Moral agency is therefore intimately tied to language and the claims expressed through it, which in turn means that moral agency is (currently) only found in humans.

Contrary to this, in this thesis I argue that technological autonomy means automation of functions. This point is further developed in the remainder for the chapter, and for now one should note that the autonomy of technology is not equal to the autonomy of humans. Johnson and Noorman (2014) summarizes the differences by stating that because of moral autonomy "Humans think, choose, decide, and then act. Humans act for reasons and their intentional behavior is outside the ordinary realm of material causality" (ibid. 151), while the non-moral autonomy describes "artefacts that operate independently from humans." (ibid. 151). Johnson and Noorman at times refers to moral autonomy as "the autonomy conception of agency" implying that the autonomy of agency is different from technological autonomy. The independent character of autonomous technology should be understood as "automatically executing functions". One should also take care to notice that the term "autonomous technology" is an general term that captures a variety of various technological artefacts and systems.

2.2 What is technology?

When answering a question like this, it is tempting to point at specific artefacts and say, “This microwave here is technology, and so is this refrigerator and that blender”. Answers like this are characteristic of a narrow understanding of technology. It refers to the materiality of technology and, as a consequence, it under-communicates important aspects such as the organization of people and knowledge, and the context of use. Artefacts are the material objects of technology – they are the products of us individuating entities and drawing ontological boundaries (Johnson and Noorman, 2014, 145). A technological artefact however, is always more than the object itself. In this thesis I adopt Winston’s notion of *technological systems*.

Winston distinguishes six interacting elements or aspects in modern technology (Winston, 1999, p. xii). The first aspect concerns techniques, activity-forms or practices. Artefacts and tools are material answers to challenges, and enables us to accomplish tasks and feats that were previously tedious or impossible. This means that technology is inherently a praxis-oriented endeavor where we employ techniques for doing a variety of tasks. Winston also calls this element for know-how. The second aspect refers to resources or basic materials. Technology is here understood as the manipulation or transformation of materials and resources. Through manipulation of basic materials we get artefacts, the third aspect. We give shape to the materials through our techniques and create artefacts to do our bidding. This leads to the fourth aspect, which is ends, functions or purposes. Artefacts have intended uses. This is not to say that an artefact cannot have multiple ends and functions different from the original intention, but refers rather to the notion that artefacts always have an original intended use. Winston refers to this as a double ambiguity since “the same artifacts can be used to achieve different ends, and different practices and their associated artifacts can be used to accomplish the same ends.” (ibid. p. xiii). What this aspect underscores, is the notion that artefacts should be understood by virtue its functions and potential ends, and that they were made with an intention at hand. The fifth aspect of technology is knowledge-that (theoretical knowledge). This aspect refers to the background of knowledge that works as a framework for technology. Knowledge is needed to know which resources are useful where, which techniques to employ, which ends and purposes techniques and objects serve, and how all these elements fit together. In modern technology, knowledge of these sorts are indispensable. The sixth aspect captures the social context of development and use. This aspect is important in order to realize that technology is not only the material artefacts, but also encompass the organization of work through division of labor, methods of working, and other cognitive techniques. In fact, Winston calls technology a social construction in order to highlight the social and historical element of technology.

Accordingly, technology should not be understood simply through references to the material artefacts. Winston summarizes and writes

> …when we speak of technology, we shall mean the complex of techniques, knowledge, and resources that are employed by human beings in the creation of material and social artifacts which typically serve certain functions perceived as useful or desirable in relation to human interests in various social contexts. (Ibid. p. xv)

This is similar to what Deborah Johnson writes:

> Artifacts (the products of human contrivance) do not exist without systems of knowledge, social practices, and human relationships. Artifacts are made, adopted, distributed, used and have meaning only in the context of social activity." (Johnson, 2006, 197).

The first sentence in this quote can seem strict, but I believe Johnson has subtle point. While artefacts can obviously exist outside of systems of knowledge, social practices, and human relationships, I believe her point is that they would be mere objects rather than actual artefacts. We recognize artefacts as artefacts because they serve some purpose or ends, but artefacts outside of a context of use seemingly loses this aspect, and therefore should not be referred to as artefacts. In a sense, Johnson is only emphasizing the necessity of a context of use to the meaning of an artefact. I can recognize the broken hammer as a hammer, but using the hammer is no longer an option. In a strict sense, the broken hammer is no longer equal to the functional hammer. Therefore, while the broken hammer exists as a hammer, there is still a way in which it is also not the same as a functional hammer.

The notion that artefacts always serve some purpose or end also implies that there are normative or moral aspects to technology. Verbeek (2011, 2014) argues that artefacts are morally significant things – they affect us in a variety of ways and help constitute our experience of the world. Verbeek's notion of technology being non-neutral is a different (theoretical) point that supplements Winston's account. One of his favorite examples is the obstetric ultrasound. The use of ultrasound in pregnancy has exposed the parents to-be to a whole range of questions and challenges that were previously non-existent. Rather than "just" being pregnant and having a baby, the parents are now facing decisions whether they should and want to have a baby in light of the medical status of the fetus. Because of the obstetric ultrasound, pregnancy has now expanded its moral dimension, and the parent to-be faces questions regarding the quality of life of their unborn child, as well as their own life. The technological artefact brings about a range of choices that we as moral agents must face. In this sense, artefacts are clearly not neutral as it offers us choices we did not have before. Not only do these artefacts offer us new choices, they also to a degree limit which choices we perceive as real – that is, they also have a normative element that suggests which choices to make.

Why is it so important to acknowledge and account for the social and normative aspects of technological artefacts? The short answer is because failure to do so might facilitate an inadequate ethics of technology. Not accounting for relevant aspects of an artefact might produce premises that does accurately depict the artefact, its functions, and its consequences or effects. In this way, we might believe we understand the artefact in question, while we in fact have not understood it properly. Additionally, when discussions departs from this (inaccurate) understanding, we run the danger of concluding inaccurately of the significance of the technology. This could in turn influence the development of technologies and shape the social practices and relations in which these artefacts are used and designed. Johnson (2014) writes that technological development

> …involves many different actors with interests that push development in a variety of directions. The many actors – scientists and engineers, funding agencies, regulatory bodies, manufacturers, the media, the public, and others – affect the direction of development. (ibid. 712).

This means that not acknowledging the social context might lead one to dismiss or fail to see how the various actors influence the final product, which in turn can facilitate a perspective on technology wherein the development follows a logic of its own. In the course of this paper I will refer to the positions of Winston and Johnson as the *contextualized perspective*. On the benefit of a contextualized view, Noorman writes

> A contextualized sociotechnical perspective that acknowledges the different interpretations and roles of metaphorical concepts highlights the multiplicity of human/technology configurations and the various ways in which these configurations take shape in different contexts.”(Noorman, 2009, 136).

The contrary perspective, which in this paper is exemplified in chapter two, is called the *decontextualized perspective*. In the latter perspective, the ontological boundaries are drawn at the object, that is, the artefact. Drawing the ontological boundaries around the artefact means to direct attention to the artefact, but this simultaneously directs attention away from all other aspects that made its functionality possible. These boundaries facilitate discussions. However, in the context of developing ethics and philosophies of technology, it becomes important to acknowledge that the boundaries are not inherent to the artefact. They are rather pragmatically drawn in order "to make sense of the world, to facilitate practices, to give meaning, to achieve tasks." (Johnson and Noorman, 2014, 147). This means that the boundaries are not natural, but rather discursive.

By viewing the artefact as decontextualized, some of its aspects are subtracted, and it changes the way we speak about the artefact. It objectifies the artefact by omitting a reference to the various social practices that is also a part of it. There is a sense in which the decontextualized perspective sees the artefact as closer to a mere object than as a technological artefact with ends, functions and users. In this way, viewing the artefact as decontextualized means to perceive it as having a different content of meaning. Decontextualizing the autonomy of artefacts should therefore be done with care as it has potential to distort and beguile any discussion that departs from it.

The contextualized perspective uncovers the various ways in which the social context co-constitute the production, development and use of artefacts. The distinction between contextualized and decontextualized perspective becomes important in the context of the thesis. Later I argue that the decontextualized view, when applied in arguments concerning autonomous technology, may reinforce a metaphorical use of language, whereby it invites misconceptions and impreciseness.

2.3 A proposed taxonomy of technological autonomy

The taxonomy I am proposing will potentially ease misconceptions regarding the concept of autonomy by outlining the central elements that could constitute the autonomy of an autonomous artefact as viewed from the technologists' point of view. It has already been suggested that a proper taxonomy of autonomy should consist both of levels varying from low to high and of different types of autonomy. The latter is important because it strikes at the core of technological autonomy; the types of autonomy tells us what kind of capabilities we are building into the artefact. By distinguishing between different types of autonomy it becomes clear to what extent an artefact can act on its environment, and simultaneously it will help ethicists, sociologists, lawyers, etc. to frame their worries in ways that correspond with technological development and the actual functions of artefacts. Distinguishing between levels of automation will also enable ways in which to frame worries and questions. The levels offers a way to describe the interaction and cooperation between the artefact and the operator, as well as which functions have been automated, thereby further clarifying the artefact's capacity to act and affect its environment.

The reason for including both levels of automation and types of autonomy is to attempt to nuance the concept and to show that technological autonomy is not all or nothing – it is rather something that moves across a continuum. Viewing it as such emphasizes that autonomy describes the independent character of a system by virtue of automating functions. It also emphasizes the notion that an artefact's operation in some way has a reference to an operator.

2.3.1 A 10-level taxonomy

Endsley and Kaber's taxonomy aims at describing how the machine increasingly automates the functions of monitoring, generating, selecting, and implementing options Endsley & Kaber, 1999, 464). Later I argue that these functions will be important in autonomous technology. Starting at manual control and moving through the levels to full automation, the taxonomy sheds light on how the operator is moving from being in the loop to being out of the loop of the systems operations. At the highest level, the control loop is somewhat closed and human intervention is limited, while at the lowest level the operator maintains all control. In their own words, they write that the taxonomy was developed to "…have applicability to a wide array of cognitive and psychomotor tasks requiring real-time control…" and is "…describing the way in which core functions can be divided between a human and a computer to achieve task performance" (Ibid. 464-5). While they are mainly concerned with teleoperations, advanced manufacturing and such, the taxonomy they offer is applicable to a number of human-machine relations, with the condition being whether the machine automates the functions mentioned above. The same four generic functions are applicable to a variety of artefacts and I suspect that these four functions will be essential to most autonomous artefacts.

The taxonomy tracks a delegation of functions and tasks between the system and the operator. At the lowest level, the operator is in charge of the four functions, and the system relies on human input to do its work. At the highest level, the system is in charge of the four functions, and the operator is completely out of the loop – at this level the system has automated all of its functions, and does not rely on human input to complete its tasks. Important to notice here, is that the system itself never decides which functions are automated. This decision lies with the various actors that took part in the development of the system.

It is important to keep in mind that what matters the most is which functions we automate. A function says something about what we enable an artefact do to, and there are different technical solutions that makes a function possible. For example, there are different ways to implement learning in a machine or a system, and different ways to implement movement in a robot. The technical solutions are less important in these contexts *unless* there is uncertainty tied to the solution itself. Certainty here means in agreement with established norms and rules of verification and validation. In cases where there is uncertainty tied to the technical solutions, this uncertainty will also transfer to the function, which in turn means that the artefact will have uncertainty tied to it. When the technical solutions are safe to implement, a question remains as to the scope or parameter of the function (e.g. the degree to what a function enables).

Autonomous technology, in general, will be near the end of Endsley and Kaber's scale – and I expect some autonomous artefacts will be at even higher levels than what their taxonomy accounts for, by including other functions and capabilities (such as machine learning). For example, Peter Asaro suggests that future autonomous robots might "…be capable of formulating their own moral principles, duties, and reasons, and thus make their own moral choices in the fullest sense of moral autonomy" (Asaro, 2007, 51-2). While Asaro is only entertaining the idea of future autonomous robots, it is illustrative of the gap between present and future robots, and highlights the possibility of further levels in the taxonomy. I am tempted to entertain Searle's Chinese room argument (1980), and I am subsequently suspicious of whether a robot can be said to understand language and normative claims. However, technological innovations tends to surprise us, and the innovators rarely predict the various uses of their new technology. This makes technology inherently hard to predict. Nevertheless, I am not convinced that the language-barrier and capacity for self-evaluation is something easily reproduced through automating functions. Subsequently, I cannot concur with Asaro on this point. I return to the notion of artificial moral agency in the next chapter.

Let us turn to Endsley and Kaber's levels of automation in table 1.

Level of automation	Description	Explanation
Level 1	Manual control	The human performs all tasks

Level 2	Action support	At this level, the system assists the operator with performance of the selected action
Level 3	Batch processing	The automation is primarily in terms of physical implementation of tasks
Level 4	Shared control	Both the human and the system generate possible decision options
Level 5	Decision support	The system generates a list of decision options that the human can select from or the operator may generate his or her own options
Level 6	Blended decision making	The system generates a list of decision options that it selects from and carries out if the human consents.
Level 7	Rigid system	The system presents a limited set of actions to the operator. The operator's role is to select from among this set.
Level 8	Automated decision making	The system selects the best option to implement and carry out that action, based upon a list of alternatives it generates.
Level 9	Supervisory control	The system generates options, selects the option to implement and carries out that action. The human mainly monitors the system and intervenes if necessary.
Level 10	Full automation	The system carries out all actions. The human is completely out of the control loop and cannot intervene.

Table 1 – Endlsey and Kaber's 10 levels of automation (Endsley & Kaber, 1999, 464-5)

At the first level, the machine does nothing without the operators' input – it is purely manual control. The operator performs all functions. At the next level, the machine assists the operator with executing the task, but the operator performs all other functions. At the third level, the machine alone executes the task, and so it goes gradually until the machine finally generates and selects the option, and implements the decision – all without human intervention, meaning it is autonomous. Remember that the autonomy of technology means automation of function, which in turn could also describe how the artefact is operating independently of an operator. Being autonomous in the technological sense also means to operate independently.

Consider level 7, the rigid system. Here the system has the ability present set of options that the operator must choose from – the operator cannot generate any options of her own. This means that the system has several information-related functions automated. It would need to recognize, process, and manipulate information in order to present it. At this level of automation, the system is said to have a degree of independence in the sense that it has fully automated certain functions. However, as we have seen earlier, this restricted independence is tied to what the system is intended to do. If it is a system involved with air traffic control, we can imagine that it will keep track of all the airplanes within a certain radius, and will offer new coordinates to the operator if two planes are on a collision course. Here we see that the system's capability to offer coordinates is limited by the task of preventing the collision planes – this task could in turn be contingent on other aspects, such as effective distribution of air space. We see that the degree of automation in a system is related to the functions and tasks it is perceived to accomplish.

Next, consider level 9, supervisory control. At this level the system has automated all four functions, and the operator's responsibility is to monitor the system, and intervene if necessary. In cases that revolve around systems at this level, the operator takes no part in the execution of functions unless something warrants an intervention. This means that if all things go as planned, the system will independently complete its task. Systems at this level are in use in big electrical grids, where the system monitors the grid and implements options, be it because there is a danger of blowing circuits or damage to infrastructure. This level of system can also be used to monitor credit cards, with the purpose of finding illicit use (fraud, stealing) and implement options towards the protecting the bank's customer. For example, the system could see that a card is being used at multiple cash dispensers in an area within a short time frame, and by virtue of that fact it could implement the option of closing that specific card because the usage resembles that of a thief.

The taxonomy gives us a way to see how the different functions (monitoring, generating, etc.) are assigned to the parties involved in the execution of a task. Assignment of functions are essential to understand technological autonomy because the assignments tells us, in a specific way, what the system can and cannot do – that is, it tells us where the autonomy begins and where it ends. It also emphasizes the notion that the systems have an intended purpose for which it was made, which in turn influences its capabilities to act. Both the assignment and the purpose goes to tell that the human praxis of technology is explicitly and/or implicitly present in the artefact. Explicitly present in the form of an operator or supervisor, implicitly present through the purpose and the capabilities to achieve said purpose – as such, there is always human practices present in technology. In the following section I turn to Professor Perri 6 and his ladder of autonomy.

2.3.2 The ladder of machine autonomy

In his paper on ethics and regulations of new artificial intelligence, Professor Perri 6 (2001) views technological autonomy as divided into 8 different categories. By doing so, it enabled him to answer some of the worries related to this new technology by framing the worries in ways that corresponds to the actual technology. Like Endsley and Kaber, Professor 6 suggests that autonomy is not all or nothing, but is rather a spectrum or a ladder (ibid. 407). This ladder, he suggests, goes from mundane artefacts like spoons all the way to highly sophisticated autonomous technology like intelligent weapon systems. He notes that as we move up the ladder, the artefact is less dependent on direct human input and has basic capacities for diagnosis and decision-making (ibid. 409). He always maintains throughout his examples that these artefacts are the product of humans, which means that what they do and how is circumscribed by those who made it. As such, he also embraces the contextualized perspective. Conceptualizing technological autonomy as consisting of different *types* of autonomy could prove highly useful in discourses concerning the ethical and social implications of autonomous technology. By doing so, the concept is nuanced and precise instead of general and underdeveloped, and this way of conceptualizing facilitates attempts at deriving the technological impact.

Professor 6 then goes on to describe some basic notions of artificial neural nets in order to draw attention to machine learning. While effective in finding solutions in complex and rich environments through induction (rather than deduction), neural nets have difficulties providing explanations of their decisions (ibid. 410). We can observe a neural net and what it does, but the process or route that leads from input to output is hard to decipher. Nevertheless, neural nets and machine learning opened a dimension in robotics in which a robot could now be equipped with capacities such that it could learn how to move its limbs. Now turn to table 2 to an overview of the types of autonomy. The reader is advised to pay closest attention to the first four types as these corresponds to the four functions in Endsley and Kaber's taxonomy. The other four types are included to draw attention to potential ways to conceptualize the autonomy of technological artefacts, but will not occur later in the text. Since the first four correspond to the functions in Endsley and Kaber they will occur in the proposed taxonomy later, while the other four will not.

Type of autonomy	Definition
Kinetic autonomy	Capability for making allocations of movement energy over a defined structure (a body) with purposive effect.
Cognitive autonomy	Capability for recognizing information, processing and manipulating it beyond merely routinely following a pre-programmed routine
Learning autonomy	Capability for developing models inductively of relationships between phenomena that could be expressed propositionally.
Decisional autonomy	Capability for using cognitive and learning autonomy to come to decisions to take action.
Classificatory autonomy	Capability for extending any set of semantic classifications provided in initial programmes specifying basic ground rules of operation, in order to further cognition, learning and communication
Second-order capabilities	Generic capabilities for learning additional specific capabilities, such as the above.
First-order institutional autonomy	Capabilities for selecting which of a range of available institutions in which to participate in order to solve trust problems.
Second-order institutional autonomy	Capabilities for innovating institutionally, to create new kinds of institutions as trustworthy decision environments.

Table 2 – Perry 6’s categories of technological autonomy (Perry 6, 2001, 413)

This ladder shows another way to conceptualize the autonomy of technology. Similar to Endsley and Kaber, this ladder describes autonomy in light of automated functions. These categories allows us to frame our worries in a more specific language by tracking the limits of an artefact’s autonomy. It also gives us a more nuanced concept of technological autonomy by listing different categories in which we can conceptualize the potential contents of an autonomous artefact. For example, we can say that Roomba, the vacuum cleaner, is autonomous by virtue of it possessing kinetic autonomy. Roomba has automated certain functions, such as avoiding obstacles and recharging, and can independently execute the task of cleaning an apartment. Another example is the intelligent thermostat, which we can now say have learning autonomy (since it learns patterns of temperature) and decisional autonomy (since it can change temperature).

I want to raise a point here in relation to decisional autonomy. While Professor 6 lists learning autonomy as necessary to decisional autonomy, I will maintain a less strict approach wherein learning autonomy is not necessary. While cognitive autonomy seems necessary due its relation to information, learning does not occupy a similar position. It is not vital to the concept of decisional autonomy that the artefact can learn, but processing and manipulating information is. This leaves us with the following definition

> Decisional autonomy: Capability for using cognitive autonomy to come to decisions to take action, which may or may not involve the use of kinetic efficacy, but without dependence in each case on human decision-making.

Instead of relying on the general and ambiguous concept that is technological autonomy, we can now discern technological autonomy in terms of the various types. This approach facilitates discussions concerning the autonomy of artefacts and stops it from being underspecified.

2.3.3 A proposed taxonomy

To recall, the aim of this chapter is to put forth a combination of these two perspectives on technological autonomy in order to create a richer conceptualization of the concept. The choice of robots as the target of this taxonomy is motivated by two reasons. Firstly, many of the internal mechanisms of an autonomous robot will most likely have similarities to non-robotic autonomous technology. This means that the functions that constitute the autonomy of a robot will likely be found in other autonomous artefacts. Secondly, autonomous robots are coming. Substantial amounts of money, energy, and time is put into research and development of autonomous robots and in all likeliness, robots will be an increasingly common sight in the future (SPARC, 2014).

In the case of autonomous robots, all eight types of machine autonomy could be included in a taxonomy. Nevertheless, in the taxonomy I am proposing, only the first four types are accounted for (kinetic, cognitive, learning and decisional). The reason for choosing these four types is that they correspond to the functions in Endsley and Kaber’s taxonomy. Since autonomy also describes the independent character of this type of technology, it is reasonable to assume that they will possess kinetic (using its body), cognitive (recognizing, processing, and manipulating information), learning (develop models inductively of relationships between phenomena), and decisional (making decisions to take action) autonomy. These four are capacities that will likely constitute a variety of autonomous robots, albeit there might be autonomous robots wherein other combinations of autonomy is found. For example, a robot might not have the capacity to learn, but still have kinetic, cognitive and decisional autonomy. Additionally, the robot might have one or more of the latter four types (classificatory, second-order, etc.). Which types of autonomy that constitutes a specific robot is circumscribed by its purpose. This implies that a variety of different combinations might see the light of day, which again means that a variety of taxonomies could prove useful.

Take as an example an autonomous submarine that inspects and repairs subaqueous structures. We can suppose that it would need kinetic autonomy in order to navigate properly in an ever-changing environment. It is also likely that the submarine needs cognitive autonomy in order to make sense of its environment. Lastly, it would need decisional autonomy in order to decide whether it can fix the damage itself or if it should call for help. On top of this it can also have learning autonomy in order to learn how to better navigate and map its environment.

Recall that Endsley and Kaber were occupied with describing general functions, i.e. monitoring, generating, selecting and implementing. These same four generic functions are transmittable to autonomous robots. Not possessing several of these functions is in contradiction with the independent character of autonomous robot. It will need monitoring in order to grasp a situation, it will need to generate options towards achieving its goal, it will need to decide on an option towards the goal, and it will need to implement its decision in order to achieve its goal. Removing two or more of these functions from the robot's operation means introducing a human operator into the process.

If we now view Endsley and Kaber's taxonomy in light of Professor 6's ladder of autonomy, it becomes clear that decisional autonomy is first found at level eight (automated decision making). Up until this level, the system is dependent on human decision-making. Also introduced at level eight is kinetic autonomy, whereby the system implements its decision into action. Cognitive autonomy is however present from level four (shared control), where it starts to generate options. In order to generate options, the system must first recognize and process information relevant to the execution of a task. Lacking, then, from this taxonomy is learning autonomy.

Combining Endsley and Kaber's taxonomy with four of the types of autonomy found in Professor 6's ladder gives is the following table. Take notice that learning autonomy is not accounted for (n/a). Learning is not accounted for because Endsley and Kaber have not accounted for this function in their taxonomy. The column is therefore left empty.

Level of automation	**Description**	**Kinetic autonomy**	**Cognitive autonomy**	**Decisional autonomy**	**Learning autonomy**
1	Manual control	None	None	None	n/a
2	Action support	None	None	None	n/a
3	Batch processing	None	None	None	n/a
4	Shared Control	None	Yes	None	n/a
5	Decision Support	None	Yes	None	n/a
6	Blended decision making	None	Yes	None	n/a
7	Rigid system	None	Yes	None	n/a
8	Automated decision making	Yes	Yes	Yes	n/a
9	Supervisory control	Yes	Yes	Yes	n/a
10	Full automation	Yes	Yes	Yes	n/a

Table 3 – My 1st combination of level of automation with types of autonomy.

Since Endsley and Kaber's taxonomy does not include machine learning, the row denoting learning autonomy is left empty. By viewing the taxonomy in types of autonomy, it now easier to see what the differences are and when a type of autonomy makes its entrance. Up until level eight, we see that the artefact is minimally autonomous – that is, it can only provide options to its operator, and the choice remains with the operator. To include learning autonomy in this taxonomy, there are at least two possibilities; either expand the number of levels, or, maintain the number of levels (10) and adjust. For the sake of ease and clarity, the taxonomy will remain at 10 levels. After the taxonomy, a discussion will follow wherein I review the various levels and show how some of the levels are susceptible to further division.

Level of autonomy	**Description**	**Kinetic autonomy**	**Cognitive autonomy**	**Decisional autonomy**	**Learning autonomy**
1	Manual control	None	None	None	None
2	Action support	None	None	None	None
3	Batch processing	None	None	None	None
4	Shared Control	None	Yes	None	Yes
5	Decision Support	None	Yes	None	Yes
6	Blended decision making	None	Yes	None	Yes
7	Rigid system	None	Yes	None	Yes
8	Automated decision making	Yes	Yes	Yes	Yes
9	Supervisory control	Yes	Yes	Yes	Yes
10	Full automation	Yes	Yes	Yes	Yes

Table 4 – My 2nd combination of level of automation with types of autonomy, this time with learning.

Learning autonomy makes its entrance at level four. It is possible to conceive of learning autonomy at lower levels, mainly the kind of learning associated with allocating movement energy over a defined structure. Consider as an example the autonomous thermostat. It possesses learning and decisional capabilities, but none of the others. Given time and internal storing, the thermostat could eventually map out variations in temperature, and use this information to maintain an ideal temperature. However, since the thermostat does not have a defined physical structure (a body) which it can move, it falls short of being a robot and is not described by this taxonomy. Consider instead Rumba, the vacuum cleaner. While Rumba does not possess any learning capabilities, it has a body and control over it. Rumba can navigate various obstacles on the ground in a room, and some versions can return to its base in order to recharge. Rumba therefore has some basic capacities for diagnosis, decision-making and movement. Accordingly, it fits the description of being an autonomous artefact – albeit a simple one.

Recall that kinetic autonomy was defined as the capability "…for making allocations of movement energy over a defined structure (body) with purposive effect…beyond an initial act of release (turning on electrical power)" (ibid. 413). In the table above, we see that the operator does more than an initial act of release up until level eight. From this level and on, the robot itself does not rely on the operator unless necessity warrants intervention – it can therefore be said to have kinetic autonomy. It is important to take notice of the fact that even if a robot possesses both learning and kinetic autonomy, these capabilities must not necessarily be intertwined and affect each other (although they can). That is, the capability to learn does not have to be directly related to the robot's movement, but could enable the robot to process information at a faster rate due to it learning from experience. The things a robot can learn is purpose-dependent and circumscribed by its designers. The fact that a robot can learn something does not mean that it can learn everything. It will have specific parameters programmed into it that will limit its area of learning, and furthermore, its ability to learn will have limits in the shape of its hardware and sensory equipment. Hence we see that machine learning is highly circumscribed, as is every robotic capability.

Returning to the relationship between the artefact and the operator, we can see that through level 1 to 7, the operator is in the loop in various ways and degrees, and the operator's presence is gradually smaller as the levels progress. At level 8 and 9 the operator is on the loop and the human functions as a supervisor. At level 10, the operator is off the loop. These levels are however susceptible to further division, and I will outline some of these possibilities.

At level 4 and 5 there is a possibility of differentiating between whether the user manually feeds the robot information or if the robot itself collects relevant information. If the latter, then it is also possible to differentiate whether the robot should ask for permission to collect certain types of information (for example private personal information) or not. At level 8 and 9 it is possible to differentiate whether the robot should ask or wait for permission to act or not, and if the robot should be able to override the human decision. It is also possible to differentiate at level 10 whether the human user can intervene or not. Common at all levels that includes learning is the possibility to differentiate levels based on how spacious the region of learning is (i.e. how much the artefact can learn)[1]. We see that most of these levels are susceptible to further division and can easily be revised to fit a specific type of robot or other autonomous artefacts.

[1] This sort of differentiation is applicable to most, if not all, of the capabilities. I have chosen to focus on learning at this time since this is the capability that prompts most worries.

3. Two arguments concerning the ethical and social consequences of autonomous technology

Departing from the theoretical background provided in chapter 1, I will criticize two different arguments; one put forth by Matthias (2004) in his article *The responsibility gap*, and one by Floridi and Sanders (2004) put forth in their article *On the morality of artificial agents*.

In the case of Matthias, I argue that his notion of technological autonomy is underdeveloped and that he has not fully accounted for the limits in the autonomy of machines. Moreover, when this underdeveloped notion is employed in the decontextualized perspective it facilitates further interpretation of the machine's capabilities. Lastly, I argue against his understanding of technological development by pointing at the various ways in which actors can take part in the development, and I take this notion as indicating the non-determinism of technological development.

In the case of Floridi and Sanders, I argue that there are important dissimilarities between the autonomy of artificial agents and the autonomy of human beings – these concepts do not describe the same phenomenon. Floridi and Sanders proposes an analogy (a sort of Turing Test) in order to establish the morality of artificial agents. I argue that this analogy does not hold because the two agents (human and artificial) are of two different types of agency – one provides a basis for moral agency while the other does not.

These two arguments are taken as examples of inadequate understandings of technological autonomy. In the context of autonomous technology, ethics and social consequences it is important to acknowledge how the engineers and programmers understand the autonomy of technology. Adopting a taxonomy provides a differentiated tool and language that enables us to analyze and understand what an artefact is capable of doing, thereby clarifying the delegation of tasks and where responsibility lies.

3.1 Matthias and the responsibility gap

In his paper Matthias is concerned with what he perceives to be an inevitable outcome of the development of autonomous technology – namely, the impossibility of holding manufacturers and operators morally responsible or liable for the actions of an autonomous artefact. In his text, Matthias does not distinguish between moral responsibility and liability, and I follow him in doing so. He follows Fischer and Ravizza, and maintains that being in control of a situation and the following consequences is necessary to hold someone responsible. This means that an individual must know enough of the facts pertaining to a situation and must be able to freely form a decision to act based on these (Matthias, 2004, 175). Alternatively, in the language of section 2.1 - the individual must be receptive (knowing facts) and reactive (form a decision) to the reasons involved in particular situation. Autonomous machines are accordingly in conflict with any human actor being in control over the situation, and thereby complicates ascriptions of responsibility.

Furthermore, according to Matthias this development is inevitable. This view proposes that technological development is determined by the nature of a particular technology – that the development has a logic of its own (independent of us). This perspective on technological development closes the discussion too soon and fails to acknowledge the various "behind the scenes"-activities that shapes the final product.

Matthias draws attention to usual practices concerning ascription of responsibility and machines, and notes that we usually assign responsibility to the operator depending on whether or not he uses the machine according to the manufacturer's specifications, while we assign responsibility to the manufacturer depending on whether or not the machine operates as specified (ibid. 175). These notions are in accordance with the control principle for responsibility. The operator commits himself to use the machine as prescribed by the manufacturer, thereby assuming a position of control over the use of the machine, which in turn makes him responsible for it. The manufacturer is committed to providing a functional machine that works as prescribed, thereby assuming control over the production, which in turn makes him responsible in the sense that he guaranties that the machine has no flaws in its construction. Matthias then states his case

> …There is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough *control* over the machine's actions to be able to assume the responsibility for them. These cases constitute what we will call the responsibility gap. (Ibid. 177)

Matthias claims that this new class of (autonomous) machine actions prevents both the manufacturer and the operator to be reasonably in control over the consequences produced by the machine. On this matter, Noorman and Johnson (2014, 52) notes that this idea states that as robots are becoming more autonomous, this suggests robots will be in control and human actors will therefore not be in control. Matthias believes this idea and backs his claim by referring to the learning capabilities that present and future machines possess, and writes

> …presently there are machines in development or already in use which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself*. (Matthias, 2004, 177)

Matthias is here arguing that machines that possess the capability to learn could turn the learning capability on itself, thereby altering the rules that determines its behavior. This would in effect entail that the machine becomes unpredictable and thereby complicating the manufacturers and operators having control. Nevertheless, this does not necessarily imply that a responsibility gap has arisen. Since artefacts are crafted, formed, and made entities that rely on human input in order to work, it simultaneously entails that there is at least a casual responsibility involved in both making and using an artefact. Casual responsibility means that someone or something takes part in the causal chain that leads to an event – for example, how the rain is responsible for the flood or how I am responsible for the vase I knocked over and broke. In the context of technology, this entails that someone is casually responsible for the existence of an artefact. Moreover, if artefacts are morally relevant things, then this could also entail a moral responsibility for the existence of an artefact. The leap from autonomous learning machines to a responsibility gap is therefore not as straight forward as Matthias implies. It might be, as Noorman and Johnson suggests, that autonomous machines entails that human actors have different kinds of control and responsibility (2014, 59).

Matthias' quote above reveals his perspective on technology and the quote lends support to an idea. Matthias is primarily occupied with the artefact itself and its functionality as a ready-made entity, which means he takes part in the decontextualized perspective. Moreover, the idea states that the machine is the locus of both decisional and learning authority – this is what "by the machine itself" indicates. This authority is two-folded. Firstly, it says that the machine enjoys a scope of action and the authority to choose within this scope. Secondly, the machine can (through learning) alter the rules that determines its behavior. The decontextualized perspective and the idea of "artefact-authority" makes it hard to see how we would stay in control over the machine. Not only are we dealing with an amoral entity (it has no concept of right and wrong), but also this entity can change how it acts (by itself). This endows the machine with an independence and capability to commit changes and execute actions that is in conflict with you or me being in control.

If we concur with this perspective and idea, then Matthias is right: there is a responsibility gap. However, as I will argue, this perspective fails to acknowledge the various ways in which actors negotiate and decide an artefact's capacity to act – it does not account for the "behind the scenes"-activity. Acknowledging the contextual perspective enables one to see how various actors are in control, thereby also suggesting where responsibility lies. Additionally, Matthias does not offer a definition of autonomy, which is problematic because we are then left with guessing what this autonomy consists off. He uses an undifferentiated language that invites interpretation. Viewing machine autonomy as automation of functions further shows how we stay in control and a taxonomy could describe the relationship of control between any artefact and its operator. Consequently, I argue that we stay in control over the technology and that there is no responsibility gap. Negotiations between various actors, the establishing of practices, and delegations of control and responsibility are ways in which we stay in control over technology.

Let us shortly review Matthias' conception of technological autonomy. While Matthias does not offer a definition of technological autonomy, he does work through the concept by use of examples. It is apparent that he uses the concept to denote a class of artefacts that operates independently of human supervision. By viewing the quote above in light of Professor 6's ladder we see that Matthias is mentioning decisional and learning autonomy. The machine can decide on a course of action (decisional autonomy) and can change the rules that determines its behavior (learning autonomy). This means that we have automated functions that allow the machine to this. Furthermore, his other examples indicated cognitive and kinetic autonomous as well. The problematic aspect of Matthias' presentation of this technology is that he employs concepts that does very little to illuminate the scope of these types of autonomy – the concepts are uninformative as to exactly what the machine can and cannot do. In this regard, Matthias' use of language invites interpretation and potentially misconceptions.

It is true that a machine is "…able to decide on a course of action..." but this miscommunicates certain aspects of the machine, and overstates other aspects. It is true by virtue of recognizing the causal part a machine plays in the creation of situations and consequences, but miscommunicates the part that the manufacturer and operator plays. In particular, it severely fails to convey how an artefact's ability to decide on a course of action is circumscribed by the designers and manufacturers. That is, the autonomous machine never really decides on its own because its choices and decisions are predetermined by the intentions and choices of those who creates and uses it. This same notion is applicable to all the capabilities of an autonomous artefact. Which capabilities that we endow an autonomous machine with will always be dependent on its intended and envisioned work and use, which in turn limits the scope of any capability.

Please note that this is not the same as saying that technology does things that the creators did not expect or foresee, but rather it highlights the need to conceptualize this technology as not "acting on its own". Drawing the boundaries such that we wind up talking of the machine as "acting on its own" contributes to the worries associated with this technology, and leads the discussion down a path where it becomes harder to see who is responsible other than the machine. There is a way in which these boundaries stops the causal chain at the machine rather than at the actors behind it. Unable to see the scope of the autonomous machine's capabilities (uninformative concepts), simultaneously as we see the machine doing things on its own, could facilitate an interpretation where the machine is seen as having a unquantified potential to act with no one responsible for its actions. A taxonomy could help alleviate this interpretation, and furthermore a taxonomy informs what functions are automated and how this relates to an operator, which in turn informs us as to what kind and level of autonomy this specific machine has and what it can and cannot do.

Consider Matthias' own example of the intelligent elevator with learning capabilities. In the example, the intelligent elevator leaves a high a ranking business executive waiting for a considerable amount of time, which leads to her company losing a contract and thereby suffering financial damage. Who is to blame, Matthias asks. The manufacturer can deny responsibility due to the elevator's capabilities being what they are, and more or less the same reason applies to the owners of the elevator (Matthias only mentions the manufacturers of the elevator in his example, but I suspect the same reasons will apply equally for the owners in this scenario).

Now it is obvious that both parts plays a part in the causal chain of events that lead to the elevator being used in that particular building, so there is casual responsibility to be found. What about moral responsibility? It is possible to argue that the manufacturer did not appropriately design the elevator in terms of human interaction with the elevator, which was what led to the business executive not getting the lift she was waiting for. Viewing the example in light of forward- and backward-looking responsibility might be illuminating.

In this context, forward-looking responsibility revolves around the division of tasks between the artefact and the human actor, while backward-responsibility revolves around discerning where something went wrong with the artefact (or operator). One might then argue, looking back at the mishap with the executive, that the manufacturer did not properly account for an aspect of the elevator's operation – namely, that it should provide lifts for those pressing the buttons. That is, by looking at what the elevator is intended to do, we might see what it should do in the future. Thus, one can argue that it is the responsibility of the manufacturer to provide elevators that does not enable situations as the one Matthias describes. This is an approach to analyzing the situation which does not wind up in a scenario where we conceptualize the elevator as "acting on its own", but we rather trace its capabilities back to the manufacturer and holds them responsible for the actions of the elevator.

Matthias holds as a position that since autonomous machines will not have all their behavior explicitly preprogrammed, they are in effect uncontrollable when the machines can learn. However, this is a coarse characterization of the machine's capabilities, as well as equally coarse regarding how these machines came to be. Firstly, not being explicitly pre-programmed does not entail not being pre-determined. Secondly, the machine's capabilities will have their limits and scopes determined by other factors than the codes in a software (such as hardware and envisaged tasks). Thirdly, it is reasonable to assume that the theoretical possibility of autonomous machines being uncontrollable will lead to stricter regulations concerning verification and validation of the machine's operation capabilities. On this topic, I agree with Johnson and Noorman (2014) that it is imperative to pay attention to what goes on inside the machine, as well as how its capabilities are a result of negotiations between various actors to fully understand the technology and its consequences.

3.1.2 A disadvantage of the decontextualized perspective

The decontextualized perspective that Matthias is employing certainly has its place in discourses concerning autonomous technology and its ethical and social consequences. Since autonomous machines are meant to operate *by themselves* in an independent manner, the decontextualized perspective can draw attention to important aspects. The perspective draws attention to the machine's regular operations, and can thereby facilitate practices and help us map out the machine's influence on its environment.

While the decontextualized perspective has its ways in which it is fruitful and meaningful, it also has its disadvantages. Firstly, it skews the picture. It fails to pay attention to the decisions made by the actors involved and how these decisions have had an impact on the development of the machine. Secondly, it transfers authority to the machine itself by omitting a reference to the actors involved in making and using it. In this perspective, the machine becomes locus of authority – not the human actors who determined the machine's framework and use. By neglecting the intentional history of the machine (why it was made) we simultaneously fail to see its limits. The decontextualized perspective of Matthias mainly accounts for the artefact, which was one of six aspects mentioned by Winston, but neglects to account for the other five. The artefact is (only) a component in a larger technological system.

As a result, the decontextualized perspective might facilitate a tendency to rely on a metaphorical use of language as a way of seizing the artefact's capabilities and potential to act. Johnson and Noorman states that drawing the ontological line such that it delineates the artefact simultaneously "…blinds us to all of the activity behind the scenes…" (Ibid. 146). The decontextualized perspective draws attention to certain aspects and parts of the technology, while also blinding us to other aspects and parts. In this case, our attention is drawn to the artefact itself, while drawn away from other aspects. Our explanation of the artefact departs from the artefact itself, and not from its makers and the actors that designed and negotiated its capabilities. Moreover, when the autonomy of this technology is left unspecified or underdeveloped we attempt to seize it through anthropomorphic descriptions. As these descriptions is one of the topics in the next chapter, I will only remark on the vagueness of such descriptions. I already mentioned that Matthias employs concepts that does little to illuminate the scope of a capability. In this thesis I take these concepts (learning, choosing etc.) as examples of anthropomorphic descriptions. That Matthias' elevator can learn conveys little information of how much it can actually learn. Can it learn theoretical quantum physics? How to play Led Zeppelin? Or is it limited to storing and mapping traffic patterns? The concept of learning (which is a metaphor in the sense that it draws an analogy between human learning and machine learning) does very little to inform me of exactly what the elevator can and cannot learn.

In the decontextualized view, the machine is seemingly operating entirely on its own - in manners that resembles life-like behavior and warrants the use of a certain language. We single out the machine, our attention is directed at the thing itself, and our explanations of it departs partly from the behavior of the machine as perceived by us. Should we however open the machine and be explained how it works, we would see how the machine's functionality is circumscribed by those who developed it, and we would see the scope of its capabilities.

3.1.3 Human constraints on machine autonomy

In relation to this, Noorman and Johnson (2014, 58-9) argues that the framework surrounding the machine's operations can be affected in three ways. Firstly, a machine has as mentioned above a specific problem or task that it is meant to do or solve. This means that the designers and developers work with this specific goal in mind, which in turn delimits and constrains the behavior of the machine (Ibid. 58). These limits to the artefact' capabilities show themselves in the software and hardware, where both aspects of the artefact constrain what it is able to do.

The autonomous vacuum cleaner in your house does not do much other than vacuum independently and occasionally bother the cat. While the latter should be considered a (unlucky) side effect, the former describes the intention of having an autonomous vacuum cleaner. What is important to note here, is that the capabilities of the vacuum cleaner is dependent on its (delegated) purpose (cleaning). Its ability to move back and forth, to turn and detect obstacles, to vacuum, to recharge when power levels are low – these are all tied to the purpose of the machine. Moreover, any capability or level of capability that interferes or could threaten to interfere with the work and purpose of the machine is either constrained or removed. For example, the vacuum cleaner is not powerful enough to suck your cat or carpet into itself, nor can it travel at such speeds that a collision would damage your walls or furniture.

Some might react to the notion that the vacuum cleaner above is being described as "autonomous". In the previous chapter I argued that autonomous technology must be understood as systems or machines that have automated functions and can on basis of this operate independently. The modern vacuum cleaner does exactly this. Functions such as movement, navigation, and recharging have been automated, and based on the automation the vacuum cleaner operates independently.

Secondly, human actors can affect technological autonomy through norms and rules that governs machine behavior (Ibid. 59). In this respect, Noorman and Johnson are concerned with a scenario in which a machine can interpret and apply laws and norms in different scenarios, but norms and rules can also exert influence through how they function as guidelines for development and research. While the former would be a significant achievement, as it demands a considerably sophisticated machine capable of understanding language, morality and complex situations (contextual clues), the latter is ever-present in and through the social context and background knowledge that constitutes the arena of research and development. Recall Winston's interacting aspects of technology in section 2.2 where the social context and background knowledge was mentioned as two of six elements of technological systems.

In relation to robotic behavior and morality, Asaro (2006) writes, "For the most part, the nature of robotic technology itself is not at issue, but rather the morality behind human actions and intentions exercised through the technology." (p.11). Asaro presents a view in favor of technology mediating morality and values, similar to that which Verbeek (2011) has argued elsewhere. In Verbeek's view, technological artefacts are morally significant in the sense that artefacts are active shapers of our experience and choices – technology affects situations and outcomes. This goes to show that there is also a way in which norms and rules are expressed through technological artefacts.

The third way in which human actors can affect technological autonomy is through predictability (Noorman & Johnson, 2014, 59). The idea that autonomous technology will function without direct human supervision puts extra demand that this type of technology will need emphasis on reliability and predictability. Even when an autonomous machine enjoys a scope of action, there must predictability within this scope that stops the machine from doing something unwanted. It must also be reliable in the sense that the human operator can trust that the machine does what it intended to do.

These three ways of affecting the autonomy of machines illuminates two important aspects. Firstly, it shows that the degree and types of machine autonomy is dependent on the negotiation between various actors (scientists and engineers, funding agencies, regulatory bodies, manufacturers, the media, the public, and others). Secondly, it shows that the machine's capabilities are directed towards its intended work or purpose. This facilitates a notion: A machine's capabilities (what it can and cannot do) is a product of human choices, choices that entail we are in control over the technology, and choices that could incur responsibility. The framework that constitutes a machine's operation is settled and decided upon by the various actors that participates in the development and negotiations. In a sense, this is obvious when you consider that machines and robots do not go around and do random stuff- rather, they do specific work in specific areas with specific goals at hand.

These three ways of exerting influence goes to tell that technological autonomy can never truly be detached from the social context of use and development. It goes to show that any mention of an autonomous machine as *choosing* by *itself* is potentially misleading if it does not acknowledge how human choices and decisions shape the framework of the machine. A mechanism of this sort does not allow for more decisional freedom than that which is bestowed by the involved actors. Furthermore, this influence also shows itself in a machine's ability to learn. As with the ability to choose, the ability to learn will always be limited or constrained by the task or problem that the machine is intended to do or solve.

Accordingly, Noorman and Johnson (2014) argues that we retain control over autonomous technology through deciding the behavior of the artefact. We can summarize this notion by referring to what Professor Perri 6 writes in his article. Artificial intelligence and autonomous technology "... raises questions as to which kind of decisions ought and which ought not to be delegated to such systems" (Perri 6, 2001, 406). The point here being that we, as the ones making technology, are the ones who decides what a machine can and cannot do. Returning to the vacuum cleaner, there is little or no point in it being able to learn how to solve complex moral dilemmas or provide mathematical proofs, but learning how to clean the apartment in the most efficient way is a desirable ability.

In conclusion, this leaves us with the notion that discussions concerning technological autonomy and its ethical and social consequences should strive to include a definition or taxonomy of autonomy. Failure to do so facilitates interpretation of the autonomy of the machine, which in turn reinforces the reliance on metaphorical descriptions as a way to seize the machine's capabilities. As stated earlier "…machine autonomy is not a single idea" (Noorman and Johnson, 2014, 60). Furthermore, I argued that this reliance is reinforced by Matthias' decontextualized perspective on technology. This perspective singles out the artefact from its history and context of use, and facilitates an impression that this machine truly acts on its own as a human would. Contrary to this, the contextualized perspective accounts for how various actors take part in negotiating the capabilities of a machine and shows how we retain control over the machine. The notion that we retain control provides a counterargument against Matthias' position of a responsibility gap as the control condition for responsibility is maintained in development and use.

3.2 Floridi and Sanders on autonomous technology and artificial moral agency

In their paper *On the morality of artificial agents*, Floridi and Sanders argue that artificial agents are "legitimate sources of im/moral actions" (2004, 351) and hence should be considered moral agents. They arrive at this conclusion by listing three criteria of agency and one moral qualifier, and through a method of abstraction they argue that artificial agents satisfies these. The method of abstraction enables them to pose an analogy between standard moral agents and artificial agents. The claim is that if there is a level of abstraction at which there is no discernible difference between a standard moral agent and an artificial agent, then the artificial agent should be considered a moral agent if it produces actions with moral consequences. It is Floridi and Sanders themselves that states humans as standard moral agents (Ibid. p. 357).

On the method of abstraction they employ, Floridi and Sanders writes "The Method of Abstraction comes from modelling in science where the variables in the model correspond to observables in reality, all other being abstracted" (ibid. 354). What counts as "observables" are not specified, but they mention intentions, meaning, wishing and wanting to act as "…psychological speculation" (ibid. 365). This indicates that they have a strict conception of observables where mental states and intentions are excluded – these "psychological speculations" cannot be observed and measured by their method, and are therefore seen as irrelevant – a conception that has similarities with methodological behaviorism. This raises the question of whether the method and the variables can properly account for necessary aspects of moral agency.

They list four variables necessary for moral agency: interactivity, autonomy, adaptability, and causing good or bad. The first three of these are necessary only of agency, while the last is the moral qualifier. Interactivity means that there is an exchange between the agent and the environment; autonomy means that the agent can change its states by internal transitions (without direct interaction); adaptability means that an agent can change the way (or rules by which) it changes between different states; and the moral qualifier means that the agent produces actions with good or evil consequences.

They then state the analogy. At a given level of abstraction we observe two entities, H and W. Both entities satisfies the three conditions of agency, and we can appropriately think of them as agents. We are then asked to suppose that H kills a patient while W saves a patient. Accordingly, we are now dealing with two moral agents since the agents in question have caused good (saving the patient) and bad (killing the patient). What is the twist? One of them is a human, the other is an artificial agent, but we cannot distinguish them from one another. Since humans are rightfully considered moral agents and we cannot distinguish between H and W, the artificial agent should be considered a moral agent as well.

This indicates that Floridi and Sanders are employing a version of the Turing Test. The original Turing Test (Turing, 1950) tests whether an interlocutor can distinguish between a human counterpart and a machine in a conversation. If the interlocutor cannot distinguish between them, the machine is said to be intelligent (it can think). This is similar to the scenario that Floridi and Sanders put forth. We are asked to view two agents, which we cannot distinguish, and are then told that one of them is a moral agent. Since we cannot reliably tell the difference between the agents, the artificial agent (in this case) passes the test of being a moral agent. The original Turing Test has been criticized (among others) for not testing for real intelligence, but rather tests for behavior that resembles intelligence (Saygin et al. 2000). I suspect a similar critique can be raised against Floridi and Sanders' proposed analogy. The artificial agents only resembles moral agents, but they cannot be appropriately thought of as having the autonomy usually connected to moral agency.

My contention here is that they are confusing the autonomy of technology with the autonomy of moral agents, and that as consequence their analogy is unsatisfactory. There are important dissimilarities between the autonomy of a human agent and an artificial agent that are not accounted for in their paper. They move from the use of metaphors as a way of understanding the significance artificial agents, to be using metaphors as a basis for attribution of moral agency. By relying on the different conceptions of agency as found in Johnson and Noorman (2014), I will show how they differ and how they must be kept apart. Furthermore, if moral agency is to be understood as they frame it, then this transforms the concept and it is hard to see how we can talk about "good" or "evil". If the only prerequisite of moral agency is an ability to produce moral good or evil, then how can Floridi and Sanders account for the role of reasoning, self-evaluation and normativity in ethics? It is hard to imagine any normative claims without the ability to reason, and consequently it becomes hard to imagine how anything could be good or bad. Additionally, their conclusion raises question concerning rights, duties, and the respect associated with moral agency. Would we consider an artificial moral agent as having the intrinsic worth that we consider a human moral agent as having?

3.2.1 A problematic analogy

Recall the start of the previous chapter where I maintained that moral autonomy consists of the ability to reflect on the reasons that motivate action, and that this is connected to a language of worth. Technological autonomy was here found as describing the automation of functions, and signifies the independence that arises due to the automation. Contrary to the concept of moral autonomy, the concept that Floridi and Sanders develops states the following:

> Autonomy means that the agent is able to change state without direct response to interaction. It can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment. (Floridi and Sanders, 2004, 357)

Now, it is possible to consider Floridi and Sanders' definition as partly capturing moral autonomy as it was understood above. Even if the language they use is influenced by what can be said to be *technical language*, they seem to agree that the autonomy of moral agents consists of an ability to perform internal changes. However, it is not entirely clear how Floridi and Sanders' definition could be seen as capturing the role of language and the reflexive attitude that I argued was essential to moral autonomy and moral agency. Due to the essentiality of language to moral autonomy, I instead regard their conception of autonomy as a specific type of technological autonomy. Recall section 2.3.2 where the types of autonomy were defined. In this section, decisional autonomy was defined as a capability to come to decisions to take action. This definition bears resemblance to the one offered by Floridi and Sanders. That the system can "perform internal transitions to change its state" means that the engineers have automated functions that enables the system to come to such decisions. We can therefore describe their conception of autonomy as decisional autonomy. Additionally, we also see that Floridi and Sanders views autonomy as a concept describing the independent character of autonomous technology.

On these grounds (the lack of moral autonomy), it is now possible to reject the thesis defended by Floridi and Sanders. However, in light of this, Floridi and Sanders may argue that the conception of moral autonomy that I propose is not necessary for ascription of moral agency. What is instead necessary is their characterization of agency and the moral qualifier they put forth.

However, as shown by Grodzinsky et al (2008), the method of abstraction can lead to a conclusion wherein the artificial agent is not considered a moral agent. Floridi and Sanders are claiming that if there is a level of abstraction where you cannot differentiate between the human and the artificial agent, then you should ascribe moral agency to the latter by virtue of the former being a moral agent. Grodzinsky et al. shows that at a given level of abstraction, the behavior of the artefact must be attributed to the designer, which means that the analogy is flawed – they are not analogous as previously claimed. If you are an (healthy) adult, then you will not attribute your behavior to your parents.

There is a difference between the artificial agent and the standard moral agent in terms of capabilities (reasoning, self-evaluation, feeling of sensations etc.). Another difference shows itself when we consider that the engineers of an artefact (its makers) are not equal to parents of humans. The most essential difference between these two levels of abstraction is that Floridi and Sanders are departing from the perspective of the user, whereas Grodzinsky et al. departs from the perspective of the designer. The latter perspective pays attention to what goes on inside the artefact, as well as how its functions are circumscribed and pre-determined, while the former perspective only pays attention to the external influences of the artefact and views it as decontextualized.

That the same method results in two opposing conclusion points to the possibility that Floridi and Sanders might be confusing various types of agency. According to Johnson and Noorman (2014), there are three different conceptions of artefactual agency usually found in discoursers concerned with the moral status of technical artefacts. The first conception of agency refers to artefacts as being things that brings about states of affairs (causal efficacy); the second refers to artefacts as things that act on behalf of humans (acting for); and the last refers to autonomy – autonomy as independence from human actors or as moral autonomy (ibid. 148-152). The first conception becomes important in this context. Floridi and Sanders are claiming that artificial agents are moral agents if the artificial agent produces moral consequences – that is, they bring about moral state of affairs (a moral efficacy if you like).

Conceptualizing artefacts as having agency can be highly informative with regards to what they do and how they affect us and the environment, which in turn can be useful when designing, developing and using technology. Attributions of agency is a way of understanding through metaphors. It draws attention the role that artefacts plays and reveals aspects of their significance. However, metaphorical use of language is not entirely innocent. Johnson and Noorman writes

> …in thinking metaphorically, we may be directed to think that the two things have more in common than they do. Important and relevant dissimilarities between the compared entities may be pushed to the background by making a particular analogy between the two entities. Moreover, analogies can leads us to believe we understand something when in fact the

> thing used in the analogy is very poorly understood. (Johnson and Noorman, 2014, 150-1)

When employing anthropomorphic descriptions, it then becomes important to realize these potential pitfalls that Johnson and Noorman notes. The concept of agency is colored by consciousness, and the potential danger in conceiving artefacts as having agency lies in whether or not anthropomorphic properties of agency are perceived as also belonging to the artefact. Therefore, while we can meaningfully ascribe agency to artefacts in order to understand how they affect moral actions, we must take care not to "…claim that humans and artefacts are interchangeable components in moral actions" (ibid. 153). Moral autonomy means responding to reasons, while causality steers the behavior of artefacts. The former provides a basis for rights and moral agency, whereas the latter does not. At best, the agency of artefacts must currently be understood as amoral agents with moral relevance.

In light of the three conceptions of artefactual agency, consider this example used by Floridi and Sanders. In their paper, they argue that a thermostat is moral agent if it senses the temperature in the room, changes the temperature based on sensory information and past treatments, and finally saves the patient by doing so. When explicating their example it becomes clear that they also conceptualize the artefact as decontextualized. As with Matthias, they are also mainly concerned with the artefact itself, and is thereby only accounting for one of six aspects of a technological system. Perceiving the thermostat in isolation draws attention to the fact that the thermostat is casually responsible for changing the temperature in the room and that it did so on the basis of stored information and its internal decision mechanism, but the perspective cannot properly account for how its capabilities are a product of negotiations between actors. Since the thermostat is not only bringing about states of affairs and is acting instead of a human, but also does this in an independent manner, we can conceive of the thermostat as being an autonomous agent. According to Floridi and Sanders, considering the thermostat as being a moral agent is warranted by its actions having moral consequences.

A problem in this argument is the move from the use of metaphors in order to understand the significance of autonomous artificial agents, to the use of metaphors as a basis to attribute moral agency. The three conceptions of agency are meant to illuminate aspects of an artefact, and could be seen as an epistemological and communicative tool. Important in this respect, is to keep in mind that these three conceptions of agency are not equal to human agency. The three conceptions allows us to see important aspects and elements of technology, and facilitates communication and practices. Nevertheless, humans and artefacts are not analogous in the respect that Floridi and Sanders claim. The agency of humans is different from the three conceptions mentioned. To see this from another angle, let us view human-to-artefacts delegation of tasks and responsibility in light of the thermostat-example.

The thermostat gets delegated the task of maintaining a healthy room temperature – we can say that the responsibility of maintaining the room temperature now lies with the thermostat. However, this delegation of task does not imply that the thermostat is responsible for maintaining the temperature other than in a casual manner. In other words, we delegate casual responsibility to the thermostat – not moral responsibility. Johnson and Noorman argues in this respect that we never delegate moral responsibility to artefacts – in their view moral responsibility is only part of human-to-human delegations (Johnson & Noorman, 2014, 153-4). That delegation of tasks to artefacts does not include delegation of responsibility points to the idea that artefacts are not perceived as appropriate recipients of moral appraisal and responsibility, and in this respect it also points to the idea that artefacts cannot reason about potential state-of-affairs. Their ability to decide and to act is the result of causality, programming and hardware. Since artefacts cannot reason about normative claims and future state-of-affairs it makes little sense to consider them moral agents and responsible in the same manner as human agents. We do not expect or believe that artificial agents evaluate their motivations and desires. This is what we may consider to be a relevant difference between machine and man.

The distinction between moral and causal responsibility offers a way to show the differences concerning human-to-human and human-to-artefact delegations, and can help in framing the worries in a language that reduces the chance of being allured. In this respect, it should be mentioned that denying artefactual moral responsibility and agency is not the same as denying that artefacts play a role in moral action.

Floridi and Sanders' decontextualized perspective on technology contributes to their conclusion. They too, just as Matthias, adopts a perspective wherein the artefact is perceived as decontextualized from the human praxis. We saw above that this type of perspective on technology could complicate our understanding of the artefacts and how they work by neglecting to account for other essential aspects of the technology. By viewing the artefact in isolation, we blind ourselves to how the activity behind the scenes has given shape to the functionality of the artefact. The decisional autonomy of the thermostat is limited to changing the temperature in the room, but this decisional autonomy is in itself further limited by what is thought to be a satisfactory and healthy temperature. In other words, the thermostat has a target temperature that was circumscribed by its designer. This shows that a predetermined framework limits the options available to the thermostat and that the thermostat's choice is in some way also predetermined. Professor Perri 6 (2001) explicitly mentions thermostats of this kind, and notes that the

> …abilities of the system to learn are limited to the capacity of a simple sensor to detect changes in or levels of temperature…The definitions of appropriateness of targets are set by human actions, and the diagnostic capacities are limited to those for which it was designed. (Ibid. 408).

This also points to the notion that artefacts and humans are not identical parts in moral action – in the case of the thermostat there is a difference regarding the agent's ability to choose. Viewing the artefact as decontextualized could lead one to not properly acknowledge how the human actors circumscribe the artefact's functions and capabilities, thereby endowing the artefact with a capability to act that it does not have.

Nevertheless, while viewing the artefact in isolation and employing metaphors to understand its aspects might lead to one to not properly seize what the artefact is and its moral status, this line of thought also draws attention to the moral relevance of artefacts. I agree with Floridi and Sanders in that an action qualifies as moral if it can cause good or evil – that is, an action is morally relevant if it has good or evil consequences. Now, autonomous technology do hold great potential to do just this. In general, technology co-constitutes aspects of reality[2]. Artefacts shape our choices and perceptions of the world. Recall Verbeek's example of the obstetric ultrasound in section 2.2 where we saw that technology is not neutral. In Verbeek's own words, stating the constitutive role that artefacts play, he writes "Ethics is about the question of how to act, and technologies appear to be able to give material answers to this question by inviting or even exacting specific forms of action when they are used" (Verbeek, 2006, 377). Conceptualizing the artefacts within the human context therefore seems to draw attention to how artefacts can affect moral choices and situations, and this means that Floridi and Sanders are correct in concerning themselves with the moral consequences of technological artefacts.

In the case of autonomous technology, this notion of non-neutrality becomes important when viewed through the decontextualized view. In contrast with non-autonomous technology, autonomous technology does not need the explicit input of a present operator, which means that

[2] The co-constitution of experience with technology will be explained in the next chapter

it will carry out actions regardless of whether we are observing or not. The independent character of autonomous technology then entails that there might be no one present to stop potential accidents or wrongdoings from happening. Now it easy to see why Matthias, Floridi and Sanders have adopted the decontextualized view. By adopting this view, they are concerning themselves with the artefact's operative presence – with the normal work state of the artefact – thereby drawing attention to the morality of autonomous technology. If an action is considered a moral action if it produces good or evil, then this implies that technological artefacts could be considered moral entities if their actions have moral consequences. The category of "moral entities" are meant to capture those entities or beings that are relevant to moral decisions or situations.

Since autonomous technology could be relevant to moral decisions or situations, and even produce actions with moral consequences, we can appropriately think of them as moral entities (they are by no means neutral and completely amoral). By viewing artefacts as moral entities, we can reserve the category of moral agency to those who contemplate and reflect of normative claims and future state-of-affairs, thereby avoiding obscuring important aspects of the standard moral agent. This line of thought draws attention to how we can view artefacts as mediating intentions and actions without having to equate artificial agents and moral agents.

During this subchapter I have argued that technological autonomy and moral autonomy must be understood as two distinct concepts that describes two different phenomenon. The latter provides a basis for attribution of rights, moral responsibility and moral agency. The former describes a particular type of technological artefact that has an array of functions automated so that it can operate independently of a human operator. These two concepts must not be conflated, and it is important to keep in mind how the engineers and programmers understand technological autonomy. Essentially, I have criticized the method that Floridi and Sanders employ. This method is unable to capture important dissimilarities between humans and artificial agents when it comes to moral agency.

Furthermore, while ascriptions of agency to artefacts is a metaphorical use of language meant to highlight aspects of technology and facilitates understanding and communication, these ascriptions must be kept separate from the conception of human agency due to important dissimilarities. I agree with Floridi and Sanders that there is moral relevance to be found in artificial agents, but this relevance only amount to them being seen as moral entities – not moral agents.

Also important is the notion that the autonomy of technology is not a single idea, but can rather be conceptualized as a system of concepts as I showed in the previous chapter. Floridi and Sanders' autonomy conception is similar to Professor 6's conception of decisional autonomy. However, as stated, there are at least seven other variants of autonomy that can be found as descriptive of an autonomous system's capabilities. Their conception deals with an artificial agent's capability to perform internal transitions without direct response to interaction, thereby endowing the agent with an independence from its environment. While this is correct in the sense that it describes an important aspect of autonomous technology, it fails to account for other important aspects, and is a narrow understanding of technological autonomy.

4. A perspective on autonomous technology and metaphorical descriptions

The purpose of this chapter is to investigate why we are liable to use certain descriptions and concepts when commenting on autonomous technology, and furthermore to show how these descriptions can obscure our understanding. I start by laying out the basic notions of human-technology relations as found in Ihde (1990). His post-phenomenological account provides a perspective on how technology affects our experience of the world, and important in this respect is his notion of *alterity relations*. In alterity relations we relate to an artefact as something more than a mere object, and this appearance gives rise to anthropomorphism (or zoomorphism). I argue that autonomous technology is a case where the artefacts appears to us as more than a mere object due to their independent and interactive character. In this respect, Ihde's perspective can clarify why we employ certain descriptions and concepts when commenting on autonomous technology.

I then supplement this account with research concerning anthropomorphism and robots/software, that autonomous technology reinforces the use of anthropomorphic descriptions. Having done so, I argue that these descriptions are not suitable to affirm the significance of an artefact's capability, and that they rather can have the adverse effect of obscuring our understanding.

4.1 Ihde's phenomenology of human-technology relations

Phenomenology in general takes as its point of departure the structures of experiences and can be described as a study of human-world relations (Sokolowski, 2000, 2; Gallagher, 2012, 8-9). A central idea in the phenomenological tradition is the notion that humans and the world constitute each other through the relationship between them. This is made possible due to the intentional structure of our experience – our experience is always directed at something – experience is here seen as referential (Ihde, 1990, 22). In the phenomenological tradition this entails that the world and the perceiver (human) cannot be thoroughly separated and understood apart of each other. The world is "in" us as much as we are "in" the world, and any separation of the perceiver and the perceived is discursive and entails a distortion of the phenomenon. In the act of perceiving, the world and the perceiver cannot be thought of as independent of each other, but must rather be seen as constituting each other reciprocally (ibid. 23).

Developing this fundamental notion further, Ihde argues that we have become so intertwined with technology that our relationship to the world has become technologically mediated. We live in, with and by technological artefacts, and they pervade our everyday lives in various ways and manners. We seldom relate to the world without our experience being in some way mediated by or directed at artefacts. We wear clothes to protect against weather, we live in houses and apartments for the similar reasons, we work, write and read on computers, and the list goes on. Substantial amounts of our lives revolves around technology and it affects our lives in a variety of ways.

To show how technology can have an influence on our perception and relation to the world, Ihde identifies two interrelated dimensions of perception: *microperception* and *macroperception* (Ihde, 1990, 29). The former is the type of perception that comes from the sensory capacities of the body, such as feeling, hearing and seeing. The latter is the type of perception that comes from a shared community – it is the cultural and interpretive perception that comes from history and tradition. While these dimensions can be distinguished thematically, they are never truly independent of each other and they constitute the whole that is perception. As Ihde himself writes on this relationship "There is no microperception (sensory-bodily) without its location within a field of macroperception and no macroperception without its microperceptual foci" (Ibid. 29). This is another way to frame the phenomenological notion that when perceiving, we perceive something as something. We see a tree because we know what a tree is and because the thing we see fits our understanding of what a tree is. The knowledge of what a tree is, is given to us by virtue of language, history and culture – we are born into a tradition and history of interpretation that through language makes sense of the world we inhabit.

The point of discerning between micro- and macroperception is to suggest how perception is both sensory and cultural (or hermeneutical), how the interplay between these are essential in our relation to the world, and to show how technology influences our experiential relation to the world. According to Ihde, human-technology relations can be constituted in several different ways that must be understood as elements along a continuum.

The first of these includes what Ihde calls *embodied technics* (Ibid. 72) and it denotes those instances where technological artefacts affect the microperceptual field. Characteristic in this human-technology relation is the notion that a technological artefact *mediates* my sensory experience of the world. In this instance, it means that I perceive the world *through* the artefact by virtue of it drawing attention to aspects of the world, while the artefact *withdraws* from my perception and becomes transparent.

Consider reading glasses as an example. When in use, the glasses mediate my perception of the book I am reading, while also withdrawing from my perception - I am not explicitly aware of the glasses when I read, and the pages in the book becomes the center of my experience. Consider another example, this time borrowed from Merleau-Ponty ([1945] 1970, 143). A blind man uses his cane to navigate and perceive the world around him. In this context, the cane withdraws as an object and extends his bodily presence (he and the cane co-constitutes the experience), mediating his perception of the world when he is navigating the streets – he *feels* the ground in front of himself *through* his cane.

Embodying technology requires a familiarity with the artefact in question – in order to embody an artefact, you must know how to use it. The better you are at using an embodied technology, the more it withdraws and becomes experientially transparent. Schematically we can present this relation as

(human-technology)→ world

Here the human actor and the technological artefact co-constitutes the experience (or the relationship to the world) – conversely, if the artefact is removed, then the perception differs as well. The subject (me) with the artefact (the clothes) becomes a changed subject, which due to the co-constituted "nature" can do and experience different things and tasks than the un-clothed subject. The arrow in this representation denotes the directedness of the experience and the brackets denotes the co-constituted subject.

A central notion in this respect is the idea that technological mediation *amplifies* and *reduces* aspects of the world. This amplification or reduction is relative to what Ihde calls *naked perception*, by which he means technologically unmediated microperception (Ihde, 1990, 43-9). For example, binoculars amplify your ability to see, while simultaneously other sensory perceptions are reduced in relation to the object of perception– you do not smell or feel the wind blowing at the island you are looking at through the binoculars.

In a modern, technological society, it is hard to see if naked perception is something that we experience. From birth and onward, technology surrounds us at every moment and mediates our experience and perception of the world. In light of the understanding of technology offered in chapter 1, both macro- and microperception is intertwined with technology to such a degree that naked perception never occurs; our experience is (in a Heideggerian way) always already technologically mediated. I therefore propose to view Ihde's notion of naked perception as an imaginary condition or state that is meant to highlight the various ways in which technology can affect our experience of the world.

Within this first way of human-technological relations, Ihde discerns another way of relating to the world. It differs from embodied technics by virtue of us not relating to the world *through* the artefact, but rather by *means* of it. The artefact does not become transparent and withdraws, but offers instead a representation of the world that we must interpret (Ibid. 80). For example, infrared pictures reveals aspects of the world previously unavailable to our perception, and we must interpret what these pictures reveal before it makes any sense to us. Even the words in this paper could (in a wide sense) be taken as an example. Ihde calls this relation *hermeneutic technics* (Ibid. 81) and it differs from embodied technics by being *read* rather than *felt*. This means that we relate directly to the artefact, and indirectly to the world. Artefacts of this kind have become indispensable in modern science and technology, allowing us to perceive and manipulate new aspects of the world, but they are just as indispensable in everyday life (the speedometer offers a representation of how fast we travel, and the thermostat represent temperature). We can depict this relation as

Human →(technology-world)

Here, as opposed to embodied technics, the world with the technological artefact becomes the object of perception. A perception in hermeneutic technic is radically different from that of naked perception. As with embodied technics, hermeneutic technics also amplify and reduce aspects of the world differing from the former by having a high contrast to the naked perception. "High contrast" in this context means that the perception is radically different from naked perception. For example, glasses have a low degree of contrast since perception-with-glasses strongly resembles perception-without-glasses.

In general, embodied technics have low contrast, while hermeneutic technics have high. The higher the contrast, the more we need to interpret. Recall that hermeneutic technics reveals aspects of the world that we otherwise would not experience. The obstetric ultrasound reveals the unborn child still in its mother's womb – something we would not see were it not for the equipment. The picture itself conveys little information unless a skilled technician interprets it.

The second way which human-technological relation can transpire is when the technology keeps in the background and shapes our experience. Fittingly, and in line with the phenomenological tradition of Heidegger, Ihde calls it *background relations* (Ibid. 108). Central to this type of relation is the notion that these technological artefacts are not explicitly present in our experience, but rather provides a frame surrounding and shaping our experiential relation to the world. Ihde mentions large technological systems like electrical grids, heating and cooling systems as examples. These systems are rarely explicit objects of perception and remain unthematized in the background, mostly drawing attention to themselves when they stop functioning. Schematically, we get the following

Human- (technology\world)

As opposed to both embodied and hermeneutic technics, there is in background relation no explicit division between the world and the technological artefact. Rather, both the technology and the world together form what could be called a technological world, which from perspective of the perceiver remains unthematized until the background calls attention to itself. There is a sense in which these technologies are both present and absent in our experience – presently shaping our relationship, but absently experienced. The lack of an arrow in this representation captures the notion that the experience is not directed at the background.

The last way of human-technological relations, and in the context of this paper the most important one, Ihde calls *alterity relations* (ibid. 97). Ihde borrows and redefines the concept of alterity from Emmanuel Levinas ([1969] 1979). In Levinas' work, alterity denotes the otherness that another human has to me. "The other" in this sense is fundamentally anthropocentric, with the "ultimate other" denoting God. Differing from the other human-technology relations, in alterity relations we relate to or with the artefact itself – not necessarily to the world through or by means of an artefact. That we relate directly with the artefact endows it with a sense of *otherness* – it is perceived as what Ihde calls *quasi-other*. The artefacts are perceived as something external and independent of us. Ihde writes of the quasi-otherness as "…stronger than mere objectness but weaker than the otherness within the animal kingdom or the human one." (Ihde, 1990, 100). That the artefact is not a mere object, while simultaneously not an animal, highlights the idea that we experience these artefacts as independent and interactive entities. In contrast to this, embodied technics are perceived as quasi-me seeing as they withdraw from perception, becomes transparent, and rely on bodily use. Artefacts that co-constitute alterity relations, on the other hand, cannot become transparent.

According to Ihde, the perception of artefacts as quasi-others almost automatically carries with it whiffs of anthropomorphism or animation (imbuing with life) (Ibid. 98-100). The quasi-otherness of technological artefacts comes about due to our experience of their abilities – that is, they exhibit certain traits and behaviors that are usually found in living entities, which in our perception endows them with a sense of spiritedness. This experience of these artefacts invite the use of a certain language by virtue of how what these artefacts do and how this is presented to us. For example, the fondness people develop to their beloved things, like the dad who caringly washes and polishes the car every weekend, or conversely blames the car when it breaks down, implying that the car is a fit receiver for such language. In this respect, Ihde consistently mentions all reactive attitudes aimed at quasi-others as quasi-attitudes.

Ihde mentions the anthropomorphizing of artefacts as a "…problematic interpretation of technologies… [That] can reach from serious artifact-human analogues to trivial and harmless affections for artifacts." (ibid. 98). His example of problematic interpretation is directed at AI research, where he writes that any characterization of "…computer 'intelligence' as human-like is to fall into a peculiarly contemporary species of anthropomorphism…" (ibid. 98). While Ihde does not explain in depth why and how anthropomorphism is unwarranted, he shows an aversion to characterize AI and computers as intelligent. From my reading of Ihde, anthropomorphism as an interpretation of technology is unwarranted since it is unprecise and at worst makes false claims regarding the artefact's capabilities.

Schematically we get

Human→technology-(world)

Here we relate directly to the technological artefact, but differs from hermeneutic technics since it does not assume a relation to the world through the artefact. In alterity relations, our experience revolves around the artefact that appears as a quasi-other. We interact directly with the artefact and the world remains in the background of our perception – the focus of our perception is the artefact. Ihde does mention that in alterity relations we might relate to the world, but this is not a necessary condition in this human-technology relation. It might be, as he writes, that the usefulness of an artefact presupposes a reference to the world (ibid. 107). Ihde writes on this relation

> I have placed the parentheses thusly to indicate that in alterity relations there may be, but not need be, a relation through the technology to the world…The world, in this case, may remain context and background, and the technology may emerge as the foreground and focal quasi-other with which I momentarily engage. (ibid. 107)

Ihde's postphenomenological account provides a perspective on the various ways in which technology can take part in our experience in and of the world. In the following subchapter, I argue that autonomous technology possesses capabilities and appearances that places it in alterity relations as quasi-others. This human-technology relationship provides a platform for explaining why certain concepts and descriptions are in use when commenting on autonomous technology.

4.2 Robots and autonomous technology as quasi-others

While many artefacts that we relate directly to or with can in principle qualify as quasi-others, there are some types of technological artefacts that have a stronger appearance as quasi-others in alterity relations. Robots and software are taken as cases of autonomous technology in this section – note that neither must be autonomous in order to take part in an alterity relation. What matters the most in alterity relations is the appearance of the artefact as an independent other different from me. However, autonomous technology (in general) will likely appear as a quasi-other due to the automation of functions and independent appearance that comes from the automated functions.

The physical structures of a robot (its body) makes for a stronger appearance than that of a software on a screen. Additionally, robots are able to move energy across their physical structure in order to execute functions and tasks, and they can do this in independent manners. This makes robots appear animate. From the perspective of a nonprofessional, a robot can be viewed as moving purposely and independently in manners that resembles living beings. Objectification has turned into animation. Moreover, robots are at times constructed as to induce a familiarity with the perceiver through humanoid appearances. They are dressed in skin-like fabrics, have human voices, or they have human-like limbs. Humanoid aspects in robots are an effective way to promote human-robot interaction (Zlotowski et al. 2014).

What of autonomous, non-robotic technology? In the previous chapters I argued that the autonomy of technological artefacts should be understood as the automation of functions that enables an artefact or system to operate independently. As functions are automated, the artefact in question will appear as doing work *on its own*. No operator is present and directing the artefact. It will appear independent and interactive with its environment. Autonomous robots, as a subcase of autonomous technology, are likely to be a class of artefacts that particularly gives rise to the appearance of quasi-otherness. However, autonomous technology does not have to come in the shape of robots. Technology of this sort also comes in the shape of software.

Now, while software lacks the stronger or immediate presence of robots, they can exercise dynamic interactions in a variety of ways that engages an individual in social or intelligent activity. Furthermore, when software is made to replace people or improve efficiency they are often modelled after how people work – software can therefore be seen as extending human cognition (assisting with mental activities such as calculating). Due to this, software is at times described as "intelligent" if it exhibits certain behavioral traits. They appear as resembling behavior that we find in other humans. Ihde writes of computer technologies "…that, while failing quite strongly to mimic bodily incarnations, nevertheless display a quasi-otherness within the limits of linguistics, and more particularly, of logical behaviors." (Ihde, 1990, 106). The more strongly a software resembles human behavior, the more prone we are to treat it as an other. For example, in 2014, a company in Hong Kong elected an AI as a member of the board that is also treated as a human board member –it is eligible to cast a vote as a member of the board (BBC, 2014).

Note that robots also extend human activity. However, the tasks and assignments delegated to robots are different from what is delegated to computers and software. When I stated that software extends human cognition, this was to emphasize that software exhibits capabilities associated with cognitive intelligence and that they assist humans in performing cognitive tasks. Robots on the other hand, are often employed in physically demanding tasks or dangerous environments. In this way, robotic behavior does not necessarily exhibit capabilities that is associated with cognitive intelligence. Sophisticated autonomous robots are cases on point that could have capabilities associated with cognitive intelligence.

Now it easier to see why Ihde said that these artefacts appear as stronger than mere objectness, but weaker than the otherness of living things. There is then a two-fold sense in which robots appear as an other; they have a physical presence and a functional presence. They appear both visually and casually as something more than mere objects. The quasi-otherness of robots resembles even stronger the otherness of people than the quasi-otherness of a car. In contrast, software does not have the physical presence, but makes up for this with the various tasks it can complete and how its functionality is presented to the user. Robots have a stronger appearance as independent entities, while software has a stronger appearance as intelligent. Both types of technological artefacts can appear as interactive, and both appears as quasi-others in our dealings with them.

Furthermore, both robots and software are in a sense multilayered, meaning that a majority of its internal workings are hidden from the operator. This further facilitates the appearance of independence since the artefact in question does not show how it works other than through its actions and external influences. There is a sense in which these artefacts appear to us as a blackbox. The operator will likely know how to use and deploy the artefact, but cannot necessarily explain how and why it works as it does. Lack of knowledge of an artefact's internal functions could in turn facilitate our experience of the artefact as having capabilities resembling capabilities of living beings – it could invite us to anthropomorphize (or zoomorphize) the artefact in question.

4.3 Anthropomorphism and autonomous technology

Anthropomorphism is a phenomenon that describes a tendency to see human-like characteristics or shapes in an environment (Zlotowski et al. 2014). Anthropomorphism in this respect is a method in which we attempt to seize and convey aspects of something non-human by virtue of what we already know of ourselves. Epley et al. (2007) have suggested that three psychological factors influence when and why we anthropomorphize:

> The accessibility and applicability of anthropocentric knowledge (elicited agent knowledge), the motivation to explain and understand the behavior of other agents (effectance motivation), and the desire for social contact and affiliation (sociality motivation). (Epley et al. 2007, 1)

The first two of these three are in accordance with the suggestion that a metaphor is an epistemological and communicative tool. Anthropomorphism shows itself as a metaphorical description that draws an analogy between the object of description and ourselves (as humans). I argued earlier that metaphorical descriptions are not innocent – they obscure as well as illuminate. Important aspects may be pushed to the background when stating an analogy. Another potential pitfall when using (human-x) analogies is that they invite interpretation of the non-human constituent. This could in turn facilitate readings where capabilities or levels of capabilities are interpreted as belonging to x, when this is in fact not true. The un-innocent character of anthropomorphism should therefore remain in the forefront of one's mind when one uses anthropomorphic descriptions of technology. These descriptions can have the adverse effect of making one believe that the artefact's capabilities are on level with a human's capabilities. In this respect, McDermott (1976) has argued that scientists should use colorless technical descriptions instead of potentially misleading psychological expressions when describing artificial intelligence.

A question of interest then becomes whether anthropomorphizing is justified or not as a method of understanding technology and conveying this knowledge, which if it is justified, raises the question to what degree should we rely on anthropomorphic descriptions.

A full answer to the first question is outside the scope of this thesis, but a short answer is at least warranted. There is justification to be found for anthropomorphizing if one views it as an evolutionary mechanism. Viewed like this, anthropomorphizing becomes something we do unintentionally solely by virtue of being humans. However, this does not provide an adequate answer to whether anthropomorphism is justified – it merely argues how it came to existence and which role it has played during humanity's evolution.

A more satisfactory answer is found if one views anthropomorphizing as an epistemological and communicative tool. By viewing it as such, the question becomes not "whether anthropomorphizing is justified or not?" but rather "to what degree?" This shifts the question from being about the phenomenon to the use associated with the phenomenon. This question is of interest when one comments on technology and especially autonomous technology, and is part of the topic of the next subchapter.

In the two preceding sections we saw that some artefacts appears to us as more than mere objects, while still not belonging to the animal kingdom. This makes them prone as targets of metaphorical descriptions, and particularly as targets of anthropomorphism. The automation of functions endows the artefact with an appearance of life (through its ability to act) and self-sufficiency (independence, no operator in sight). The artefacts are seen as doing things and executing tasks in independent manners usually associated with living beings. That certain technological artefacts appear to us as such facilitates further use of anthropomorphic descriptions of their capabilities, and enables attitudes towards the artefacts that were previously directed at humans or animals. Artefacts of this kind *appear* to us as quasi-living entities and the language we employ to describe them reinforces this appearance.

Furthermore, there is yet another manner in which autonomous technology itself reinforces this tendency. Due to the independent character of this technology, we separate the artefact out from the surroundings, observing it as producing effects, moving around seemingly on its own. Johnson and Noorman (2014) writes of this separation as drawing ontological lines and delineating the object, and by doing this blinding "…us to all the activity behind the scenes" (Ibid.

146). By drawing the ontological lines as such, some aspects of the artefact are brought to the forefront, while other aspects are pushed to the background. As mentioned, perspectives wherein we isolate the artefact draws attention to vital aspects of the artefact, and furthermore it facilitates communication and gives meaning to the artefact. However, when the technology also functions autonomously (moving and navigating independently, choosing, learning etc.), this perspective facilitates the tendency to use anthropomorphic language when describing the artefact.

Consider Paro, the therapeutic robot seal, as an instance of an artefact that takes part in an alterity relation and gives rise to anthropomorphism. Paro is shaped as a baby seal, it has motoric functions that mimic lifelike behavior (moving eyelids, tracking movement with its head and eyes, making seal like noises), and goes on standby at night (mimicking sleep). Paro has these characteristics because as a tool, Paro is meant to soothe and calm agitated patients. What is interesting in this case is that research shows that elderly patients with dementia treats Paro more like an animal than a lifeless object (Turke et al., 2006). It has been hypothesized that a lack of social connections lead people to compensate this lack by anthropomorphize entities in their surroundings (Epley et al., 2007). While some might have reservations about considering Paro an autonomous artefact, it has automated an array of functions that would allow it to be described in a similar taxonomy as proposed earlier – albeit it would not be placed on the high end of a scale.

Nevertheless, research shows a tendency to describe robotic capabilities anthropomorphically and behave towards the artefact as more than just an object (Zlotowski et al. 2014). Research also shows that how a robot behaves is important for it to be treated as a companion (Turkle, 2010), and that a robot's embodiment affects our perception of intelligence and intentionality in robots – on a neurophysiological and behavioral level (Hegel et al. 2008). Furthermore, animated robots attracts more visual attention (Bae & Kim, 2011), and robotic performance was found to be more important than embodiment for anthropomorphism (Hancock et al. 2011).

The research mentioned here gives support to the thesis that anthropomorphizing might be a "natural" tendency for humans. Moreover, it also supports Ihde's notion that the activity and interactivity of some artefacts (robots in this case) inclines us towards a certain usage of words and concepts. The research tells that robotic embodiment is important, but simultaneously that performance is even more important for anthropomorphism. If robotic performance is important, then one might suppose that this also counts for non-robotic technology. In fact, research into human-computer interaction has shown that people also anthropomorphize software and computers based on visual appearance and performance (Culley & Madhavan 2013; Headland et al. 2015).

While research shows that anthropomorphism happens in interaction with various technological artefacts, a question remains as to what degree metaphorical descriptions are justified as a method of understanding the technology and conveying this knowledge. In chapter 3 we saw two instances where central concepts were underdeveloped, and I argued that this underdevelopment lead to misconceptions regarding the technology in question by inviting interpretation. These instances are cases on point where metaphorical descriptions take the form of anthropomorphism, which in turn obscured the discussions that followed.

4.4 How metaphorical descriptions might obscure our understanding

Earlier in this thesis I stated that anthropomorphism could be understood as a type of metaphorical use of language, and that this specific use draws an analogy between a human and something non-human. This metaphorical use not only allows us to become familiar with the object of description, but also enables us to communicate aspects of the object. In this sense, metaphors could be considered an epistemological and communicative tool. However, as I will argue, metaphors and analogies might have the adverse effect of obscuring our understanding of these technologies. They obscure because they invite interpretation or impreciseness. That they can obscure while also have a function as communicating knowledge raises a question as to the role metaphors and analogies play in discussions concerning technology and society.

In the context of this thesis, analogies that take the shape of human-technology are important as they occur in both the media and in academic papers. In the previous chapter I provided an example of such an analogy in a scientific paper. There I showed how Floridi and Sanders treated the autonomy of artificial agents as analogous to the autonomy of humans and argued that these concepts are not analogous at all. Earlier in this chapter I argued that autonomous technology appear to us in ways which invite the use of certain words and concepts. This adds weight to question as to the role of metaphors.

Metaphorical use of language could allow us to seize aspects of the non-human and to convey these aspects. For example, I could tell you that this elevator in front of us has the ability to learn the traffic patterns in this building in order to maximize efficiency. Now neither of us really knows what machine learning is or how it actually works, but both of us now know that this elevator has a capability that allows it to learn. We know (in all likelihood) that machine learning is different from our own learning, but we cannot say how. In this sense, metaphors have a role and a function. One might say that metaphors, instead of being descriptive (in a strict sense), are indicative – they point in a direction without providing a thorough explanation.

While we can consider the metaphorical use of language as indicative, it is simultaneously so that it is not precise enough for us to properly understand the significance of the object we are describing. In this thesis, I consider terms such as "learning" and "choosing" as implicitly relying on our own familiarity with them. I understand machine learning because I have grip on what it means to learn, and similarly with choosing and talking. In the previous subchapter I argued that terms like these draws an implicit analogy between the object of description and the human. When these terms are not defined explicitly as technical terms, there is a danger that they might convey more than intended, and they might facilitate misconceptions.

To see how these descriptions can obscure, consider what they convey in terms of potential actions. From a human point of view, we know that the ability to learn has a wide area of applicability. My potential for learning has an unquantifiable scope; potentially, I could learn how to do fine arts, how to play a variety of games and sports, or amass knowledge in any theoretical study. The same potential shows itself in any decision or choice I make. Potentially, I could have chosen otherwise, and I could have decided not to go on a walk instead of taking the bus. When this potentiality (which lies hidden and implicit in the analogies) is attributed to artefacts, the artefacts appear as having a potential to do more than they actually can – there is an unexpressed potentiality that follows from the human constituent of the analogy (unless a technical definition is offered).

This potentiality could contribute to the disruptiveness and worries associated with autonomous technology. When we understand this technology through metaphorical descriptions (especially through anthropomorphism), the technology is not only seen as assisting with tasks, but also as equal to or surpassing us in the task at hand. It is common knowledge that technology already has surpassed us in some ways. A robotic arm can lift heavier weights than any person can, a software can process information at speeds that boggles our minds, and they can continue working without any fatigue. While this is true, one must also keep in mind that the manufacturers circumscribe these technological feats, and that the applicability of these capabilities to other areas of use are limited. However, when the knowledge of technology's superiority is coupled with the potentiality conveyed by analogies, the artefact in question could appear as an uncontrollable and superior other.

There is a two-fold sense in which the metaphorical use of language might obscure an understanding of the technology. Firstly, this use establishes analogies that draws attention to some aspects, while pushing others to the background. Analogies, I stated earlier, are not innocent in this manner. Since analogies push aspects to the background, it might facilitate misconceptions regarding the object being described. Secondly, this use of language is poorly suited to affirm the extent of an artefact's capability. When not defined as technical terms, descriptions of the artefact "as learning" does not illuminate the scope of learning, and is potentially misleading. In this respect, we might get more than we bargained for when we use metaphorical descriptions and anthropocentric concepts. The artefact's potential to act appears to us as greater than it actually is because we seize the artefact by virtue of metaphorical descriptions.

This two-folded way that metaphorical use might obscure becomes important to acknowledge in discussions concerning technology and society, and perhaps especially so when discussing autonomous technology. These systems, machines, robots and computers are described as "learning", "talking", and "choosing", and moreover the general term "autonomous technology" also is a concept that invites interpretation. I argued earlier that moral autonomy and technological autonomy are two distinct concepts; one is connected to being a human, the other denotes the artefact's independent operation. If one chooses to denote this technology as autonomous, then one must take care to separate and distinguish them properly as I did in the beginning of this thesis.

One might follow McDermott and argue that the concept of autonomy, when applied to technology, is misleading and implies an anthropomorphizing of the technology. McDermott would arguably insist on the use "automation" instead of "autonomy", as the former is neutral and the latter is not. This approach intends to lessen the potential conflation of the two concepts of autonomy by denoting the technology as automated – not as autonomous.

Another approach would be to properly define technological autonomy in terms of a taxonomy or otherwise technical language. Employing a taxonomy as a description of an artefact's capabilities highlights the various ways in which its functions are automated and how the artefact relates to an operator. In this respect, taxonomies provides a perspective on autonomous technology that is more differentiated and precise than a metaphorical use of language – there is less need for interpretation when using taxonomies. This could in turn facilitate discussions of the ethical and social consequences by enabling us to analyze the various effects of this technology. We are better equipped to see an artefact’s capabilities and the scope of its capabilities when describing the autonomous artefact in a taxonomy.

In conclusion, while anthropomorphic descriptions and metaphorical uses of language facilitate communication and familiarizes us with the object of description, they also have the adverse effect of inviting impreciseness and interpretation. As such, they are unsuitable to convey the full scope or significance of an artefact and its capabilities. In this respect, taxonomies offer a better way to describe and communicate the artefact and its significance. Taxonomies tells us the which functions are automated and how the artefact relates to the operator. As such, taxonomies also shows us what an artefact can and cannot do, and is a helpful tool that enables us to more clearly see potential dilemmas and dangers.

5. Summary

The aim of this thesis was to answer the question of we should understand the concept of technological autonomy. I argued that an answer to this question should acknowledge the technologists and how they view the technology they are making. As a result, technological autonomy should be seen as the automation of functions, and the concept describes the independence that arises due to the automation. Borrowing from the technologist, I proposed a taxonomy that accounts for four functions; monitoring, generating, selecting and implementing options. This taxonomy was then combined with Professor 6's ladder of autonomy in order to create a better tool to view technological autonomy. By doing so, I avoided treating the concept of technological autonomy as a single idea, and instead treated it as something along a spectrum. The taxonomy I proposed describes an autonomous robot, but I argued that the four functions would likely constitute other autonomous artefacts. This is due to the applicability of these functions.

In the beginning of the thesis I argued that moral autonomy and technological autonomy are two distinct concepts. One provides the basis for rights, duties, respect and moral agency, while the other does not. This was done to avoid any conflation of these concepts and in order to highlight important dissimilarities between them. Having outlined important elements to moral autonomy, I went on to offer an understanding of what technology is. Here I referred to Winston's theory of technological systems. I named this perspective as the contextualized perspective. By naming it so, the contextual nature of technology were brought to the forefront, and I avoided treating technology solely as material artefacts. This understanding of technology highlights the various aspects or elements that technology consists of, and became important later when I criticized some authors of not acknowledging these aspects. The contrary perspective, which in this thesis was primarily occupied with the materiality of the artefact, I named the decontextualized perspective.

Having argued for a philosophy of technology and taxonomies as a method of describing autonomous artefacts, I went on to criticize two arguments; one put forth by Matthias, and one put forth by Floridi and Sanders. I argued that Matthias employed an underdeveloped concept of technological autonomy that invites interpretation and impreciseness due to a lack of definition and taxonomy. I went on to argue that the interpretation reinforces the reliance on a metaphorical use of language as way to seize the artefact's capabilities. Additionally, the decontextualized perspective that Matthias also employs reinforces this reliance. I argued that there is no responsibility gap to be found when one acknowledges the processes of technological development and how various actors can affect the functions and capabilities of an artefact. Furthermore, by analyzing his example in terms of forward- and backward-responsibility I argued that this example is not indicative of a responsibility gap.

In the case of Floridi and Sanders, I argued that the concepts of moral autonomy and technological autonomy is analogous as they claim. There are important dissimilarities between these concepts – dissimilarities that their method and argumentation does not account for. Moreover, I criticized them for employing a metaphorical use of language as a way of establishing moral standing. I argued that while ascriptions of agency to artefacts is a metaphorical use of language meant to highlight aspects of technology and facilitates understanding and communication, these ascriptions must be kept separate from the conception of human agency. I also argued that their conception of technological autonomy is but a small part of a much larger idea (recall that they only mentioned decisional autonomy, one out of eight). Lastly, I agreed that artificial agents could be morally significant things, and that one could therefore view them as moral entities. Moral agency, however, is intimately tied to self-evaluation, reflection and language.

In the following chapter, I employed a postphenomenological perspective in order to investigate how autonomous robots and software appears to us. I argued that these autonomous artefacts appear to us as more than mere objects, while still not belonging to the animal kingdom. They appear to us as such because of the automation of functions and the independence that arises due to this. Additionally, some of these automated functions will give rise to interactivity, which was found to promote anthropomorphic attitudes towards the artefacts, which in turn invites anthropomorphism and anthropomorphic use of language. I then argued that anthropomorphic descriptions of technology is unsuited to convey the full scope and significance of the technology. These descriptions could have the adverse effect of obscuring our understanding since they invite interpretation and impreciseness. In this regard, I argued that there is an unquantified potential that is transferred by virtue of the implicit analogies in anthropomorphic descriptions.

I concluded in the last chapter that taxonomies are more suited than anthropomorphic uses of language to describe the significance of an artefact. Taxonomies presents an overview of the various functions, whether they are automated or not, and how the operator relates to the artefact. In this respect, taxonomies can be helpful in mapping out potential dilemmas, worries or dangers that are associated with autonomous technology. By employing a taxonomy we can see what an artefact can and cannot do by virtue of examining which functions are automated.

Bibliography

Asaro, P.M. (2007) 'Robots and responsibility from a legal perspective', *Proceedings of the IEEE 2007*, 20-4

Asaro, P.M. (2006) 'What should we want from a robot ethic?', *International review of information ethics*, 6(12), 9-16

Bae, J. E., & Kim, M. S. (2011, May). Selective visual attention occurred in change detection derived by animacy of robot's appearance. In *Collaboration Technologies and Systems (CTS), 2011 International Conference on* (pp. 190-193). IEEE.

BBC (2014) *Algorithm appointed board member*. Accessible from: http://www.bbc.com/news/technology-27426942 (retrieved: 21.04.2016)

Bostrom, N. & Yudkowsky, E. (2014) 'The ethics of artificial intelligence', Frankish, K. & Ramsey, W.M. (ed.) *The Cambridge handbook of artificial intelligence*. Cambridge: Cambridge university press, 316-34

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior, 29*(3), 577-579.

Endsley, M.R. & Kaber, D.B (1999) 'Level of Automation effects on performance, situation awareness and workload in a dynamic control task', *Ergonomics*, 42(3), 462-92

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864.

Fischer, J.M. & Ravizza, M.S.J. (1998) *Responsibility and Control – a theory of moral responsibility*. 2nd edition. New York: Cambridge University Press.

Floridi, L. & Sanders, J.W. (2004) 'On the morality of artificial agents', *Minds and Machines*, 14(3), 349-79

Floridi, L. (2014) 'Artificial agents and their moral nature' Kroes, P. & Verbeek, P.-P. (ed.) *The Moral Status of Technical Artefacts*, 1th edition. London: The University of Chicago Press, 185-212

Future life institute (2015) *Research priorities for robust and beneficial artificial intelligence*. Accessible from: http://futureoflife.org/ai-open-letter/ (retrieved: 14.03.2016)

Gallagher, S. (2012). What Is Phenomenology?. In *Phenomenology* (7-18). Palgrave Macmillan UK.

Grodzinsky, F.S., Miller, K.W. & Wolf, M.J. (2008) ‘The ethics of designing artificial agents’, *Ethics and information technology*, 10 (2), 115-21

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517-527.

Headleand, C. J., Ap Cenydd, L., Priday, L., Ritsos, P. D., Roberts, J. C., & Teahan, W. (2015). Anthropomorphisation of Software Agents as a Persuasive Tool

Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008, August). Understanding social robots: A user study on anthropomorphism. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 574-579). IEEE.

Idhe, D. (1990) *Technology and the Lifeworld*, Indiana University Press

Johnson, D.G (2006) ‘Computer systems: moral entities but not moral agents’, *Ethics and information technology*, 8(4), 195-204

Johnson, D.G (2014) ‘Technology with no human responsibility?’, *Journal of Business Ethics*, 127(4), 707-15

Johnson, D.G. & Noorman, M. (2014) ‘Artefactual agency and Artefactual Moral Agency’, Kroes, P. & Verbeek, P.-P. (ed.) *The Moral Status of Technical Artefacts*, 1th edition. London: The University of Chicago Press**,** 143-58

Levinas, E. (1969[1979]). *Totality and infinity: An essay on exteriority* (Vol. 1). Springer Science & Business Media.

Matthias, A. (2004) The Responsibility gap: Ascribing responsibility for the actions of learning automata, *Ethics and Information Technology*, 6(3), 175-83

Meland, S. V. (2016) ‘Må sette grenser for kunstig intelligens’, *Adresseavisa*, 15.02.2016

Merleau-Ponty, M., & Smith, C. (1945[1970]). *Phenomenology of perception*. Motilal Banarsidass Publishe

McDermott, D. (1976). Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*, (57), 4-9.

Noorman, M. & Johnson, D.G. (2014) ‘Negotiating autonomy and responsibility in military robots’, *Ethics and Information Technology*, 16(1), 51-62

Noorman, M. (2009) *Mind the gap: a critique of human/technology analogies in artificial agents discourse*. PhD-thesis. Maastricht university, Maastricht.

Saygin, A.P., Cicekli, I. & Akman, V. (2000) Turing Test: 50 Years Later, *Minds and Machines*, 10, 463-518

Searle, J. R. (1980) Minds, brains, and programs, *Behavioral and Brain Sciences* 3 (3), 417-457

Sokolowski, R. (2000). *Introduction to phenomenology*. Cambridge University Press.

Sparrow, R. (2007) 'Killer Robots, *Journal of Applied Philosophy*, 24(1), 62-77

The European Robotics Public Private Partnership (2014) *Strategic Research Agenda for Robotics in Europe 2014-2020*. 108 sider (Accessed 29.02.2016)

Taylor, C. (1985) *Human Agency and Language – Philosophical papers 1*. Bath: Cambridge University Press

A. M. Turing (1950) Computing Machinery and Intelligence, *Mind*, 49, 433-460.

Turkle, S. (2007) ‘Authenticity in the age of digital companions’, Anderson, M. & Anderson, S.L (Ed.) *Machine Ethics*, New York: Cambridge university press, 62-76

Turkle, S., Taggart, W., Kidd, C. D. & Dasté, O. (2006) ‘Relational artifacts with children and elders: the complexities of cybercompanionship’, *Connection Science*, 18(4), 347-61

Turkle, S. (2010). In good company? On the threshold of robotic companions.*Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues. Amsterdam, The Netherlands: John Benjamins Publishing Company*, 3-10.

U.S. Department of Defence (2011) *FY2013-2038 Unmanned Systems Integrated Roadmap*. Accessible from http://www.defense.gov/Portals/1/Documents/pubs/DOD-USRM-2013.pdf (retrieved 07.05.2016)

Vagia, M., Transet, A.A. & Fjerdingen, S. (2016) ‘A literature review of the levels of automation during the years. What are the different taxonomies that have been proposed?’, *Applied Ergonomics*, 53 (a), 190-202

Verbeek, P.-P., 2006. Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology & Human Values*, 31(3), pp.361–380

Verbeek, P.P (2011) *Moralizing Technology – Understanding and Designing the Morality of Things*. Chicago: The University of Chicago Press

Verbeek, P.P (2014) ‘Some misunderstandings about the moral significance of technology’, Kroes, P. & Verbeek, P.-P. (ed.) *The Moral Status of Technical Artefacts*, 1th edition. London: The University of Chicago Press, 75-88

Winston, M. & Edelbach, R (1999) *Society, Ethics and Technology*. Belmont, CA:Wadsworth.

World Economic Forum (2016) *The future of jobs – Employment, skills and workforce strategy for the fourth industrial revolution*. (Accessed: 08.02.2016)

World Health Organization (2010) *Health Topics: Ageing.* Available at http://www.who.int/topics/ageing/en/ (Accessed 15.02.2016)

Zlotowski, J., Proudfoot, D., Yogeeswaran, K. & Bartneck, C. (2015) ‘Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction’, *International journal of social robotics*, 7 (3), 347-60

6, P. (2001), ‘Ethics, regulation and the new artificial intelligence, part II: autonomy and liability’, *Information, Communication & Society*, 4(3), 406-34