University of Bergen

Department of Linguistic, Literary and Aesthetic Studies

DIKULT350

Master's Thesis in Digital Culture

Autumn 2016

Charting Artificial Intelligence in Reality and Fiction

A Study of How the Act of Fictionalizing

Influences Human Perception of Technology

Camilla Kottum Elmar

Abstract

This study strives to understand the act of fictionalizing and how it affects humans' perception of artificial intelligence both with and without a bodily representation. As these technologies becomes more intrinsic to our lives, and more complexity is added to them, it is important to have some understanding of how these technologies are being perceived by the public. Not everyone has knowledge of how technologies are assembled and how they function, and therefore they might not understand what the technology is intended for either. And so, information about the artefact has to be derived from somewhere, which is why this thesis is looking into the act fictionalizing as a method for creating a meaning and understanding where there is none. Accompanied by theories from literary studies, psychology, computer science, and the social and natural sciences, this thesis explores the boundaries between the real, the fictive and the imaginary in order to explain the function of fiction and how it may affect humans' observation of reality and empirical entities. It is a cross-disciplinary study with basis in contemporary research in artificial intelligence and robotics, and a collection of fictional films that portrays artificially intelligent agents differently. The thesis will defend the view that science fictions may both be modes of revelation and concealment, and that they may be viewed as possible futures, although they should never be considered definite ones. But then again, the thesis also argues that progress in any field is dependent on the act of fictionalizing, as one cannot create that which has never been envisioned and one cannot envision what has never existed without the use of imagination. In this sense, fictionalization is an important part of the human condition.

Sammendrag

Intensjonen med denne studien er å forstå fiksjonalisering og hvordan det påvirker menneskers oppfatning av kunstig intelligens, både med og uten en kroppslig representasjon. Ettersom disse teknologiene bare blir mer fremtredende i livene våre, og mer komplekse, er det viktig å ha et bilde på hvordan disse teknologiene blir oppfattet av allmennheten. Ikke alle har en god forståelse av hvordan teknologier er satt sammen eller hvordan de fungerer, og derfor kan det også hende at de heller ikke forstår hva de er ment for. Derfor må de hente inn informasjon om dette fra ett eller annet sted for å kunne bygge sin egen forståelse av den, og det er derfor denne oppgaven vil se på fiksjonalisering som en metode for å skape en mening og forståelse der det ikke er noen. Akkompagnert av teorier fra litteraturstudier, psykologi, informatikk, samfunnsvitenskap og naturvitenskap, utforsker denne avhandlingen grensene mellom det virkelige, det fiktive og det imaginære for å forklare funksjonen til fiksjon og hvordan det kan påvirke menneskers observasjon av virkeligheten og empiriske enheter. Dette er en tverrfaglig studie, som tar grunnlag i aktuell forskning innen kunstig intelligens og robotikk, samt en samling av filmer innen vitenskapsfiksjon som skildrer kunstig intelligente agenter i forskjellige lys. Avhandlingen vil forsvare oppfatningen om at vitenskapsfantasi kan både være moduser for åpenbaring og tildekking, og at de kan bli sett på som mulige fremtider, selv om de aldri bør bli oppfattet som absolutte muligheter. Det argumenteres også at framdriften i alle felt er avhengig av fiksjonalisering, ettersom ingen kan lage noe som aldri har vært tenkt, og at ingen kan forestille seg hva som har aldri har eksistert uten bruk av fantasi. I den forstand er fiksjonalisering en viktig del av det å være menneske.

Acknowledgements

This thesis has been carried out as a part of the Master's Programme in Digital Culture at the Department of Linguistic, Literary and Aesthetic Studies, Faculty of Humanities, University of Bergen, since August 2014.

First of all, I would like to express my deepest gratitude to my supervisor Daniel Apollon, Associate Professor of Digital Culture, who has given me professional guidance, challenges and useful discussions. Thank you for inspiring me to write about this topic, for the hard times you put me through while I wrote this thesis and for seeing it through with me.

I would also like to offer a special thanks to my sister, Christina Kottum Elmar, who inspired, supported and encouraged me before I even started. Thanks for the feedback, the brainstorming and for always being there for me. And the same goes for my parents; thanks for listening, thanks for reading, and thanks for asking questions.

Finally, I would like to thank all the people who contributed, big or small, in one way or another, in finishing this thesis: Daniel Croles Fitjar, Magnus Låhne, Jonas Bøe Laastad, Fay Tveranger, Mauro Candeloro, Trond Einar Snekvik and Åsne V. Brede. And also, a big thanks to the rest of my fellow master's students, who have gone through this process with me and kept my spirit high through all of it.

Bergen, November 2016

Camilla Kottum Elmar

Table of Content

Keywords

Science fiction, artificial intelligence, computer science, robotics, consciousness, philosophy, anthropomorphism, dehumanization, mental representation, act of fictionalizing, the real, the fictive, the imaginary,

1 Introduction

1.1 Topic to be Adressed

Science fiction and technological development have a mutually beneficial and influential relationship as they evolve with and around each other, not always fluently together, but in different stages. According to David Seed, science fiction can be seen as a thought experiment where aspects of a familiar reality is transformed to explore the 'what if?' scenarios of technological development and scientific discoveries, placing the narrative somewhere between the possible and the impossible (Seed 2011). Combining this view with the thoughts of Wolfgang Iser (1993), fiction may arguably be possible realities in a future we have yet to experience. Although most people associate fictions with the story-telling branch of literature (as falsehoods or lies), it is actually a process of overstepping; of exceeding reality in order to talk of that which does not exist. Fiction is therefore not the opposite of reality, but an extension of reality. And it is this process of extending the possibilities of our own reality by inflicting imaginative features upon it that is the act of fictionalizing, or simply referred to as fictionalizing (Iser 1993).

In a similar manner, scientists, engineers, programmers and researchers, as well as the general public, can become inspired by the representations projected in fictional narratives. To realize what has never been created, thought of or experienced, it has to be imagined. Throughout the centuries, science fiction have expressed humans' desire to conquer space and ocean depths, and even if mankind would have landed on the moon and reached the bottom of the Mariana Trench without them, these stories inspire the scientists and adventurers who make such giant leaps (Reisz 2015). And so, all fields of study, as well as the general process of envisioning the future, have to be governed by fictionalization.

But because the act of fictionalizing is entangled in the creation of both fictional and real possibilities, it may lead to misconceptions about the technologies that surrounds us in our everyday lives. The construction of a reliable understanding of a technology depends on people's ability to segregate real possibilities from the phantasms of fictional narratives. Technologies

that were once just a figment of the imagination of some science fiction writers have turned into technologies people use in their everyday lives, and so, the line between the possible and the impossible only becomes more and more obscured. As a result, it is not always obvious which realm technological features belong. People are bombarded with tales of the possibilities technology may bring, and what dangers may lie ahead, through fictional narratives and media, but people often neglect to ask the questions of how and why it will be so. Because of this, a lot of information about technological progress is left to be constructed by imagination, resulting in many of the characteristics and possibilities of technological devices to be derived from sources that is not representative of the empirical entity.

Humans have attempted to create machines in the image of their own minds and bodies for centuries, and even though robots have not yet overrun humanity, artificial intelligence (AI) have made a profound impact in more subtle ways. Finance, hospital and medicine, heavy industry, online and telephone customer service, transportation, aviation, toys and games, music, and journalism are some of the fields where AI applications, today, are highly prominent. From the algorithms that push buys and sells on the stock market to self-driving cars to Google's search engine, they are all run by artificially intelligent software, which makes the field include a broad specter of technologies.

Gaining an understanding of what the field of AI constituted can hence be confusing, especially if technologies from movies and television programs are taken into consideration. Works, such as *Star Wars*, *Terminator*, *Blade Runner*, *Battlestar Galactica*, *I, Robot* and *A.I. Artificial Intelligence* portray AI differently, but they may convey the impression that AI is roughly equivalent to robots. Even though this might be a faulty assumption, it is understandable that features of the real and the fictive entity get muddled. When faced with the unfamiliar, people look to similar situations or objects that they have experience with to rationalize the future outcome or behavior. If no similar experience can be derived from their real lives, fictional narratives become the second-best source as they describe scenarios that are usually not encountered in real life. This can transform technologies into entities with features and characteristics that are not present in reality, but they are assigned to the technology in an attempt to create a logical solution for understanding it. And so, the process of fictionalizing is

an important aspect of human cognitive abilities that is relevant to all acts of creating an understanding of both the familiar and unfamiliar. It is therefore important to gain knowledge of this process to understand how it affects both human perception and the construction of real possibilities.

1.2 Research Goals

The goal of this thesis is not to predict the future, nor to tear down the imagined possibilities of fictions, but rather to gain understanding of how fictionalizing, governed by theories derived from literary studies and psychology, may shape the public's perception of reality. By looking into research being done within the field of AI, supported by philosophical standpoints and hypothetic solutions, this thesis will examine the relationship between the real, the fictive and the imaginary, with the hopes of highlighting the purpose and motivations of AI research and development as it is today.

Uncertainties of what the field is currently working on and how far they have come in the process only increases fictionalizing, which is why knowledge about the fields intent has to be obtained in order to restrict faulty perceptions. As technologies become even more present in our lives, and as more jobs, tasks and information will be handled by technologies in the future, it is important to understand how they are being perceived by the people who are intended to interact with them. By being aware of the cognitive processes that takes place inside a subject when examining an entity, both the creators and users may become better at handling the changes in society.

As Digital Culture is the study of the relationship between culture and technology, and as one of the focus areas at the University of Bergen is on critical and historical approaches to technology and society, the main theme of this thesis is therefore the process of fictionalizing. It will be explored through different cognitive strategies derived from both psychology and literary studies to explain how the general public may experience technologies, using AI as the example of choice.

It is important to keep in mind that this study is not mainly intended for the AI expert, but for people engaged in the digital humanities. Because of this, the writing will not go into depths when discussing engineering and programming, as these are fields that are highly specialized and complex, and giving an in-depth explanation would be regarded as a research topic of its own. General knowledge about the processes and goals within these fields will be considered sufficient to understand how the act of fictionalizing may have an effect on their progression. The focus will be to give an elementary introduction to the scientific fields of AI and robotics, and also to the philosophical discussions that are prominent, so that the act of fictionalizing can be better explained with relation to them.

To summarize, this thesis is intended to (i) present the content and basis of current theories and discussions within the field of artificial intelligence, (ii) map key aspects of the processes of creating fictional realities and explore how understanding is revealed through the act of fictionalizing supported by theories on mental representations and anthropomorphism, (iii) clarify which types and instances of operational functions attributed to technological entities are activated during the act of fictionalizing, (iv) investigate how the act of fictionalizing may, in return, effect the public's understanding of what AI constitutes and (v) explore how science fiction may affect the field of AI by examining examples of fictive AI, contrasting their features with the technologies that exists today.

1.3 Structure of the Thesis

Following this introduction, a chapter with background information and introduction to relevant theories will be presented. Since this thesis is adopting an interdisciplinary approach, theories from literature analysis, sociology, psychology, philosophy, AI, engineering and computer science will be pursued. These theories will be considered relevant to the understanding of fictional narratives and technologies, and of humans' relations to them. What is presented in Chapter 2 will therefore lay the foundation for the upcoming discussion.

In Chapter 3, the methodological framework is introduced. This chapter will elaborate on the inventory of theories presented in Chapter 2, and argue for the relevance of an interdisciplinary approach, which may offer valuable insights on human perception and technological development.

In Chapter 4 fictional technologies will be explored and examined in light of fictionalizing with basis in real-life AI research and development. It will investigate how fictional characters, objects and narratives are being created, and the inventory will be exploited to highlight, where possible, how schemas and social constructions of AI may affect the general perception of the field. Additionally, the impact of different methods and ranges of fictionalizing will be explored.

Chapter 5 will close this thesis by adding final considerations and conclusions. It will summarize the previous chapters, the main ideas and arguments, and link them to key insights. Prospects for future research will also be discussed here.

2 Terminology, History and Theories

2.1 Explaining Artificial Intelligence

Artificial intelligence (AI) is the branch of computer science that is concerned with understanding, modeling and replicating human intelligence and cognitive processes. It is a cross-disciplinary approach aimed at making machines behave like humans, using various computational, mathematical, logical, mechanical and biological principles and devices (Frankish and Ramsey 2014, 1). The research can be highly abstract and theoretical, dealing with theories to understand natural cognition, or purely pragmatic, focusing on engineering smart applications. Some of the specializations within the field are programming computers to make decisions in real-life situations, understand natural human language, play games against human opponents, and see, hear and react to sensory stimuli.

The emphasis of AI is aimed at being artificially created, meaning that these intelligent systems are made by human beings rather than occurring naturally. Artificial components are often replications of something natural, and in the instance of AI, human's cognitive abilities are the desired target of replication. According to this definition, any program that is able to complete a task that would require intelligence for a biological organism to perform can be considered AI. How the program does this is not an issue to be considered with as the circumstance in which a constructed device is able to perform such task at all allows it to be categorized as AI. In this sense, an artificially intelligent agent does not need to understand how and why it is performing a task, it does not need to be aware of its processes, and it does not need to be sentient or conscious to be categorized as an AI system.

The field of AI is often divided into three camps with widely different goals and approaches. The first camp of AI creators is working on building and coding systems that can perform simple tasks. These AI systems are called 'narrow', 'weak', 'limited' or 'light' AI (LAI), as they are not intended to have humanlike intelligence, but rather to simulate and mimic human behavior. LAI simply acts upon and is bound by the rules imposed on it, and it cannot go beyond those rules. Therefore, LAI is very good at doing one specific task, but cannot perform any task other than that specific one (Frankish and Ramsey 2014, 16)

A good example of a LAI is Apple's Siri, as it seems to be an intelligent agent. It is able to communicate in both written and oral language, and it even gives snide remarks and tells a few jokes, but it actually operates in a narrow domain that has been predefined. Because of this, Siri will use similar methods to find answers to your questions, give the same response more than once and interpret attempts to engage in conversations as questions to be answered. The narrowness of its functions can easily be seen in its inability to understand difficult accents, and inaccuracy in results for specific search inquiries. These are 'problems' that occur as it has not been programmed to respond to a specific type of conversation.

The second works of AI creators are intended to compute humanlike intelligence in AI applications, often called 'strong AI' or 'Good Old-Fashion AI' (GOFAI). This means that they are trying to create computers that are not mimicking, but genuinely intelligent (Frankish and Ramsey 2014, 89-103). This means that a computer should be able to perform any task a human mind can; such tasks include reason, represent knowledge, make judgements, solve puzzles, plan, learn and communicate. The goal is to make GOFAI cognitively indistinguishable from humans, and because of this, it is also important that the computer has consciousness (subjective experiences and thoughts), sentience (the ability to feel emotions and perceptions subjectively) and self-awareness (capacity for introspection). It is highly debated what the terms consciousness, sentience and self-awareness constitutes, and whether or not it is possible to prove other encompasses these abilities, which is why they are not further explored in this chapter.

Examples of these types of AI cannot be found in the realm of the real, as these technologies have yet to be realized. The only examples of such a technology can be derived from science fiction. They are usually presented as sentient beings that are superior to humans in intelligence, precision and speed, and if they have an android body, they often exceed human physical strength and flexibility as well.

The works of the third camp are motivated by ideas taken from both of the previous camps, and so, it may be referred to as 'in-between' AI. These are systems that are guided by human reasoning to solve tasks, but they are not driven by the goal of perfectly modelling human cognition. This means that these kinds of computers will be intended for a wider area of use than LAI, as they are not restricted to one specific task, but can apply a set of skills in different

contexts. These computers are therefore able to solve more general puzzles and derive information from unorganized materials. They may also be able to ‘understand’ the intended meaning of a sentence by reasoning according to the rules of grammar and context, and learn through experience and communication with human experts on a given field.

One such computer is IBM’s Watson, which started out as one of their game-playing agents (IBM 2015). Watson has over time become something which its designers calls a technology platform that uses natural language in its communication with human interactors. Watson has through machine learning and communication with humans, been taught how to evaluate possible meanings from sentences and determine what is being asked. Based on its interpretation of the question, Watson presents answers and solutions backed by supporting evidence and quality information found in its database, which consists of a large body of Word documents, PDF-files and web pages, related to a given subject. By using a scoring algorithm to rate the quality of the collected evidence, it ranks different answers and presents more than one option. Because of this, Watson is able to give more accurate answers and find more relevant information for its users than any other ‘search engine’.

The above description is very limited and does only present a small aspect of what AI is and what it is capable of. It also mainly reflects the philosophical standpoint of AI programmers and engineers. The technological, computational and mechanical aspects of AI are much more diverse, as they are not only working with the computers of today, but also inventing new technologies to run software on. That aspect will be further examined in Chapter 4, along with robotics and the programming of minds.

What may be of more importance is to highlight what AI does not constitute in order to exclude certain technologies which are related to the field, but not a part of it: As a contrast to intelligent software, techno-organic enhancements (TOE) are biological entities (humans or animals) fused with artificial components or technologies to either restore functions or enhance abilities. The goal of this transition is for humans to exceed their limitations by fusing with technology and is usually associated with multiple posthumanistic views. This transition will allow humanity to proceed into the next stage of humanity (humanity+ or humanity 2.0) by becoming creatures that can mainly be found in fictional narratives, like cyborgs and Whole Brain Emulations

(WBE)(Bostrom 2009). Even though the field of TOE and AI are related to each other, and much of the same technology is present in these entities, they are not to be confused with each other. Artificially intelligent software operated entities, like computers, robots and androids, may have features that resemble human behavior and appearance, but AI puts its emphasis on artificial, meaning a computer program has to be the foundation of its intelligence. TOE on the other hand, are humans that has become technologies.

The term ‘cyborg’ is a contraction of ‘cybernetic organism’, and it is used to describe a human being whose physical or mental capabilities have been regained or enhanced beyond normal abilities by machine components or other artificial modifications (Clynes and Kline 1960). Examples of cyborgs are RoboCop, Darth Vader and Inspector Gadget, who were all saved from terrible injuries by fusing their damaged bodies with artificial components.

Whole Brain Emulation, often informally called ‘uploads’ or ‘downloads’, is another term used to describe an intelligence that is based directly on the human mind. The basic idea is to scan the structure of a subject’s brain in detail and construct a software model of it. As the human brain is a construction of different neuronal patterns, it is believed that our individual identity can be replicated through a computational design. The copy will be a direct replica of the original individual, meaning when it is run on the appropriate hardware (or wetware), it will behave exactly the same as the person which it is copied from (Sandberg and Bostrom 2008). For a software emulation to be the exact replica of the original human mind, it needs the same information in its memory and would have to process it in the same way as the original person. Dr. Will Caster in *Transcendence* and Dr. Arnim Zola in *Captain America: The Winter Soldier* are examples of human uploads, and one of the main concepts of *Avatar* revolves around the notion that humans’ subjective consciousness can be transferred between organisms.

As cyborgs and downloads were originally human beings that later acquired technological features, they do not fall under the category of AI. This distinction is important to take notice of, not only to gain a better understanding of the field of AI itself, but also because the terminology in science fiction is sometimes erroneously used compare d to that which are utilized in real-life science. Entities that fall under the definition of TOE, as explained above, can be referred to as AI and vice versa, either deliberately (artistic license) or because the writers themselves are not aware of the different distinctions. As an example, the term ‘cyborg’ is used when referring to

Arnold Schwarzenegger's character in *The Terminator* (Cameron 1984), the Terminator or T-800 Model 101. He is described as a humanoid robot running intelligent software, where his complete interior is mechanical and electronic, and the source of his intelligence is a computer program, while the exterior appears to be organic and impossible to differentiate from a real human. Even though the interior is covered by an organic exterior, the intelligence of the machine is not fundamentally human in any way. In this sense, the Terminator has been wrongfully categorized as a cyborg when it is actually an artificially intelligent humanoid robot.

In this sense, science fiction does not use the terminology in contemporary AI similarly and this, in turn, can make it harder for people to understand the concepts of each term. Perceptions of technologies are influenced by science fiction, and therefore, terminology which stray from the one applied in reality may, indeed, complicate the matter. Although science fiction is supposed to have some basis in science, one cannot really rely on fictional narratives to give a true image of authentic research, prospects and possibilities, and it is not in their intent to do so either.

This study will be concerned only with technologies that are fully artificial, excluding both uploads and cyborgs. Questions related to the two latter technologies have some of their bases in AI research, but they are more concerned with preserving and augmenting humanity, rather than the creation of artificially intelligent agents.

2.2 The Quest for Intelligence

It seems like both science and science fiction have always been fascinated by the urge to create life and to extend human capabilities by surpassing their own limitations. Augmentation is generally the purpose behind any tool or technology; to make life better in a way that unaided human effort cannot. But, as illustrated by the term machine-men (e.g., Maschinenmensch in Fritz Lang's *Metropolis*), these lifelike entities have not just been an extension of the human being itself, but a novel contribution to life. And so, humans have through tales and stories explored powers that is mainly attributed to gods by presenting humans as the creators of artificial life.

The earliest reference to an artificial man found in European literature can be traced back to ancient Greece and the 3rd century BC. According to Greek mythology, Talos was a giant man

made by the Olympian god Hephaestus to guard the island of Crete. Most of what is known about Talos is gathered from *Argonautica* (Rhodius 2008), the epic poem of Jason and his Argonauts. Talos was made of bronze, and his source of life was a vein filled with ichor (life-blood) that stretched from his heel to his neck. A peg in his foot kept the ichor from leaving his body. He never slept, and made a trip around the island three times a day to keep the peace among the Cretans.

Many similar creations have made their way onto the pages of literature to protect and save humans and their society from what they could not battle themselves. They are autonomous and humanlike artifacts imbued with life from either a divine power or fantastic clockworks. The further back in history one travels, the less is articulated about their functions and appearance, and most of their features are left to the imagination to fill in. The few lines of description mentioned in the section above is all the information *Argonautica* gives with regard to Talos, and it is baked into the telling of the story. It seems like the mentality was that if it is not important for driving the narrative forth, it is not worth mentioning.

In Jewish folklore, a similar creature, the golem, has been presented. It was an "artificial man of clay, animated by a ritual incantation" (Gelbin 2011, 1) and first appeared in the mystical interpretation of the Torah, Kabbalah, in the Middle Ages. There are many different tales of how one can bring life to the golem, but most of them agrees that the human creator has to have a divine connection with God in order to call upon his powers to animate the golem. In some stories, the golem came to life when its creator walked or danced around it while announcing letters from the alphabet and saying the secret name of God, and it was stopped by walking in the opposite direction while saying the words backwards. In other tales, the golem was animated when God's name was written on a piece of paper and stuck into its mouth, and stopped when the paper was removed. And a third option was to write three letters, *Aleph Mem Tav*, on the golem's forehead or on an amulet. This sequence of letters is *emet*, which means 'truth', and it would bring life to the golem. By removing the first letter, *Aleph*, the word *met* would be left, which means 'death', and by doing so the golem would become de-animated (Gelbin 2011, 109). *Aleph* is the first letter of the Hebrew alphabet, and it is usually associated with God, and so this sentiment signifies that with God there is truth and without Him there is death.

As with Talos, the golem came to life through the influence of a divine power. Humans did not assemble clockworks which would allow it artificial life, and in this sense, although the stories were being written and the desire to create life were present, humans had yet to learn those skills themselves. With the progression of science and general knowledge about the world, the artifacts became more complicated in their descriptions. The first descriptions of automata that owed their source of life to engineering, and not a divine god, started to appear during the Middle Ages. It should be mentioned that the sources (although connected to people which have been proven to have existed) are very few and unconcise, and there is no other significant evidence to support the claims that automata were already functional in medieval times, which is why they are mainly regarded as fictional (Truitt 2015). Although science has a long history with roots in ancient Egypt and Mesopotamia, it is indisputable that modern science emerged many centuries later. As a marker for the emergence of modern science, the scientific revolution began in Europe at the end of the Renaissance period and continued throughout the late 18th century, influencing the intellectual and social movement known as the Enlightenment (Grant 1997). During this period, large leaps were made within astronomy, cosmology, anatomy, physics, chemistry and biology. New theories, thoughts and ideas were tested and proven, and many scientific discoveries and achievements were accomplished in this period. As a part of the scientific revolution, real-life automata did most certainly become a source of entertainment and amazement in European societies. Jacques de Vaucanson was called the 'new Prometheus' for his power to create (artificial) life from new materials, which was displayed in many countries (Wood 2002). One of his most famous inventions is the digesting duck, which was a mechanical automaton with the appearance of a duck that seemed to have the ability to eat, metabolize and excrete kernels of grain. Although this was only trickery, the process was bewildering to most. He also made automata that played instruments, like the flute and tambourine.

In the 18th century, mechanical creatures were a trend in Europe and the creations were recognized as being revolutionary, although most of them could not be regarded as anything more than toys today, like a wind-up music box. Automata were metallic dolls driven by clockwork, completely automated, without any form of intelligence, which is often reflected in the writings about them that emerged in the 19th century. An example of this can be found in "The Sandman", a short story written by E. T. A. Hoffmann in 1816, where Nathanel is being charmed by Olympia, an automaton that is posing as the daughter of a professor at his university.

> Olympia appeared dressed with great richness and taste. Her beautifully shaped face and her figure roused general admiration. The somewhat strange arch of her back and the wasp-like thinness of her waist seemed to be produced by too tight lacing. In her step and deportment there was something measured and stiff, which struck many as unpleasant, but it was ascribed to the constraint produced by the company. The concert began. Olympia played the harpsichord with great dexterity, and sang a virtuoso piece, with a voice like the sound of a glass bell, clear and almost piercing. (Hoffmann 1816)

Nathanel was the only one who believed that Olympia's weird features were stunningly beautiful. "They were particularly hard upon the dumb, stiff Olympia whom, in spite of her beautiful exterior, they considered to be completely stupid, and they were delighted to find in her stupidity the reason why Spalanzani had kept her so long concealed" (Hoffmann 1816). Apart from her stiffness and doll-like behavior, Olympia is not well described in the story (like Talos), and the discovery of her automation is easily missed as it is presented in a rather short quarrel between her co-creators. What sets Olympia apart from both Talos and the golem is that her source of life is a clockwork of mechanical components invented and assembled by humans without the help of a godly power. Even though her features are not explained in detail, one understood her functionality as automata were actualized at the time of writing.

Today, we are more aware that a system of mechanical components is not enough in themselves for a human artefact to simulate life; they also need a power source just like other machines. This idea also came to Edward S. Ellis when he wrote his 1865 novel *The Steam Man of the Prairies.* The Steam Man was steam-driven and so it had a boiler in its body, which in turn made it tall and somewhat misshaped. It was a combination of a locomotive, a man and a horse, as it was used to drag a carriage on various adventures. Whatever relationship Ellis had to science and technology, he realized that if a steam engine could power factory machines and other vessels, it could also be used to power a mechanical man, and so he imagined a new technology. The steam man is a great example of using one's imagination to create fictional artefacts by fusing known technologies with each other and imbuing them with imagined life. As a result, a fictional machine, inspired by real science was created, which in turn could inspire science.

Another way fiction has inspired to real-life changes is in language. The word ‘robot’ was introduced to the English language with the translation of Karel Čapek’s play *R.U.R. (Rossum’s Universal Robots)* in 1923 (originally from 1921). It is derived from the Czech word ‘robota’, which means ‘forced worker’, making it clear that Čapek intended his drama to comment on the harsh use of labor forces and slavery (Bostrom 2005).

> So young Rossum said to himself: “A man is something that feels happy, plays the piano, likes going for a walk, and in fact, wants to do a whole lot of things that are really unnecessary.”
>
> […] But a working machine must not play the piano, must not feel happy, must not do a whole lot of other things. A gasoline motor must not have tassels or ornaments, Miss Glory. And to manufacture artificial workers is the same thing as to manufacture gasoline motors.
>
> […] Young Rossum invented a worker with the minimum amount of requirements. He had to simplify him. He rejected everything that did not contribute directly to the progress of work!— everything that makes man more expensive. In fact, he rejected man and made the Robot. My dear Miss Glory, the Robots are not people. Mechanically they are more perfect than we are, they have an enormously developed intelligence, but they have no soul. (Čapek 2014)

The word ‘robot’ in the English language today is used when referring to machines that can perform a complicated series of tasks automatically, like industrial robots do when assembling cars, and has in many ways become exactly what was described in *R.U.R.*

The topic portrayed in the play highlights the cruelty of slavery and the force of oppressed beings, and it seems like the intention of the play was not only concerned with humans’ desire to create new life and that humans should be careful in this process, but also to think about how humans treat each other. It questions the use of robots as laborers, and asks if robots can be provided (or can acquire on its own) humanlike awareness, and if they, because of that, have become more than the robot was intended to be. Can robots become humans in the sense that their awareness and conscious being should be entitled to the same rights as humans? And if humans deem them not to be so, will robots revolt against their human makers and enslave them instead?

It has been 95 years since Karel Čapek wrote *R.U.R.*, but his play has not lost its relevance. In some sense, it becomes more and more applicable to our everyday situations as technologies similar to those fictionalized by Čapek are being realized. Fields like cybernetics, computer science, neuroscience and psychology have been working together for several decades to develop artificial intelligences, and with an exponential progress within these fields some believe that the autonomous and sentient beings from science fiction will soon manifest (Bostrom 2005).

As stated earlier, artificial intelligence is a branch of computer science. Although the term was not coined until 1956 by John McCarthy during the Dartmouth Workshop (Frankish and Ramsey 2014, 18), thoughts about computing intelligent machines and how to do this had been present from the very beginning of the computer era. Widely considered the father of theoretical computer science and AI, Alan Turing developed and formalized the concept of algorithm and computation with the Turing machine, which he wrote about in his doctoral thesis in the 1930's. A Turing machine was an infinite class of machines that manipulated symbols on a strip of tape according to a table of rules. With such an abstract machine, Turing claimed that a Turing machine could compute anything that is computational, and more significantly, Turing proved that another abstract machine, the 'universal' Turing machine, would be capable of imitating any one of the Turing machines.

With this as his backdrop, he wrote "Intelligent Machinery", which makes a comparison between the universal Turing machine and the human brain. Turing argues that it would be possible to create, or grow, a computer that behaved intelligent, because the human body functions much like technological components.

> A great positive reason for believing in the possibility of making thinking machinery is the fact that it is possible to make machinery to imitate any small part of a man. That the microphone does this for the ears, and the television camera for the eye are commonplaces [*sic*]. One can also produce remote-controlled robots whose limbs balance the body with the aid of servo-mechanics. (Turing 1969, 12)

He further explains that an adult human has a mind that operates similar to a universal Turing machine; it is a multi-purpose machine, which has been 'modified' to reach its optimal potential. In its infancy, the human mind is not particularly intelligent, but over time its interference with

the outside world through stimuli changes the patterns of the brain, resulting in modification of its operation, knowledge and behavior. Humans are shaped into universal machines, rather than being born as one. Based on this, Turing believed that a similar process could be applied to machinery, teaching them to exhibit intelligent behavior.

At the end of the 1930's and through the Second World War, a series of electronic computers were made, including Z3, Colossus and ENIAC. Referred to as the 'Giant Brain', ENIAC weighed 30 tons, occupied 167 m^2 and had 18,000 vacuum tubes that were used for processing (Beev and Jung 2014). The invention of a functional computer furthered the school of ideas that the human brain was essentially an electronic computer itself, and so the quest for intelligent life was transferred from the fictional realm into reality. The creations that had only been described on paper, like the Tin Man from *The Wonderful Wizard of Oz*, could soon be actualized, as progress within the field would lead to intelligent software which could run a functional mind. The bodies were already possible, as automata had existed for centuries, and so the only thing missing was the computational framework to support all of humans' cognitive abilities.

As one can imagine, the field's approach was at that time to create Strong AI. Because of this, Strong AI is now often referred to as Good Old-Fashion AI (GOFAI). GOFAI would be a computer program whose behavior would be comparable, or even superior, to that characterizing intelligence in human beings in similar circumstances. The focus on creating GOFAI was driven by the belief that thinking and behaving intelligent was equal with algorithmic computing, while background conditions, experiences and social interactions were not essential components of intelligent life. Intelligence was seen as identical with stand-alone symbolic processing, and hence with effective computation. GOFAI was not meant to be mimetic, but to contain actual intelligence (Frankish and Ramsey 2014, 89-103).

The field was in many ways highly successful, not in terms of GOFAI, but in creating functional LAI. Arthur Samuel began working on machine learning in the 50's by building a program that learned to play checkers. Although Samuel was able to defeat his program in the beginning, it is said that it only took a couple of months before the program won every time (Frankish and Ramsey 2014, 18). During the 60's and 70's, people were also having success in building natural language understanding systems. Language was viewed as a "complex cognitive ability involving knowledge of different kind: the structure of sentences, the meaning of words, a model

of the listener, the rules of conversation, and an extensive, shared body of general information about the world" (Barr and Freigenbaum 1981, 227), and it was believed that to be able to utilize language, an entity needed to have humanlike intelligence. STUDENT was one of these programs, created by Daniel Bobrow, which could solve High School level algebra problems expressed in natural language. And ELIZA, written by Joseph Weizenbaum, mimicked a Rogerian psychotherapist. Even though the script used little to no information about emotions and human thoughts, it could provide the user with slightly appropriate humanlike interaction. Both these systems could only reason in very limited domains using restricted vocabulary and sentence structure, but it was a step in the right direction for creating the GOFAI.

But after experiencing a good summer, winter struck AI research. As Herbert Simon had predicted in 1965 that "machines will be capable, within twenty years, of doing any work a man can do" (Frankish and Ramsey 2014, 21), both governments and commercial investments started to dry up when the mid-80's came and AI could not deliver on its promises. AI became a taboo word within the computing industry, even though they had experienced success with many of their applications as these applications were not intended to be more than a mimicry of a specified intelligent process.

> Although we need to remember that GOFAI developed in a context in which crude IQ tests were very popular, it is still surprising to see how an empirically minded culture could be led so astray by its materialist project of a thinking machine as to forget that the physical nature of the brain does not prove that it functions, technically speaking, like a binary data-processor with a program, following finite lists of predetermined instructions establishing how strings of symbols need to be processed sequentially by logic gates, and that there are no ultimate reasons to believe that intelligence is a brain rather than a mental feature and "mind" just another word for "brain", and finally, that human knowledge and understanding do not resemble information-processing phenomena very closely. […] The generic possibility of modelling the brain as an input–process–output device, often useful for explanatory purposes, was confused with its actual nature and the failure of GOFAI was therefore the failure of a rather primitive epistemological model of human intelligence. (Floridi 1999, 149)

And it seems like most AI researchers, programmers and engineers came to a similar conclusion as well, as the next generation of AI research had another approach: Unlike GOFAI, which assumed there was an elementary computational model of the brain, they continued working with AI systems similar to those which had been successful from the beginning. As their aim had been to mimic one specific action or performance, remodeled according to the logic and functions of a machine, one could continue working on AI with this in mind and potentially create an infinite number of similar entities with different intent. The logic behind this idea was that a computer may process information differently from the human mind, which is why the process has to be differently oriented within the computer for it to be successful; a "'stupefaction' of the process" (Floridi 1999, 150) which we might call it. While the GOFAI was mainly intended as a general-purpose AI that could perform a variety of different tasks, the LAI would be special-purpose oriented, making the context of application more restricted, but ultimately the technology more successful.

To further explain this notion: Among the most accomplished and famous chess-playing systems in history is IBM's Deep Blue, which managed to defeat world champion Garry Kasparov in a six-game match in 1997. The program was running on a custom-built computer and was provided a large body knowledge about chess, which made it a genius chess player, but it was not intended to do anything but playing chess.

> Does Deep Blue use artificial intelligence? The short answer is No. Earlier computer designs that tried to mimic human thinking haven't been very good at it. No formula exists for intuition. So Deep Blue's designers have gone 'back to the future'. Deep Blue relies more on computational power and a simpler search and evaluation function. The long answer is No. "Artificial Intelligence" is more successful in science fiction than it is here on earth, and you don't have to be Isaac Asimov to know why it's hard to design a machine to mimic a process we don't understand very well to begin with. How we think is a question without an answer. Deep Blue could never be HAL-2000 (the prescient, renegade computer in Stanley Kubrick's 2001) if it tried. Nor would it occur to Deep Blue to "try". Its strengths are the strengths of a machine. It has more chess information to work with than any other computer, and all but a few chess masters. It never forgets or gets distracted. And it's orders of magnitude better at processing the information at hand than anything yet devised for the purpose. "There is no psychology at work" in Deep

> Blue, says IBM research scientist Murray Campbell. Nor does Deep Blue "learn" its opponent as it plays. Instead, it operates much like a turbocharged "expert system", drawing on vast resources of stored information (for example, a database of opening games played by grandmasters over the last 100 years) and then calculating the most appropriate response to an opponent's move. Deep Blue is stunningly effective at solving chess problems, but it is less "intelligent" than the stupidest person. It doesn't think, it reacts. (Floridi 1999, 153)

Today, humans interact with applications with a number of AI systems concerned with computer vision, virtual reality, image processing, game theory and strategic planning, natural language processing, translation, speech, handwriting and facial recognition, and bots of all kind, on a daily basis. But as mentioned earlier, a functional AI is not always referred to as AI anymore, which may be a result of its status as taboo and a failure in the 80's. When people started to work with AI again, they had to be careful with the use of the word, as it was highly associated with broken promises. When AI had redeemed itself with the success of numerous applications, the term came into use again, but it was not properly redefined. As the field did not emphasize that the AI they wanted to create were narrow, and not strong, the average person was left with the notion that AI were intended to be more than a simulation of human behavior. Even though search engines and personal assistant software in smart phones are AI algorithms in all their simplicity, they are not categorized as so. It seems that to be categorized as AI, the technology has to be more in the lines of 'in-between' AI or have a bodily representation other than those humans usually encounter.

Although the quest has been slightly altered, it is obvious that it is not over. The technologies we now have that contain machine intelligence enhances our abilities to extraordinary lengths, but they are far from being human. Science fiction has in recent years introduces AI as characters of the narrative, taking the technology one step further than what science can do alone. Although one may argue that this has been the case since the time Alan Turing imagined an intelligent machine, the fictional characters have grown to be more complex and powerful than ever. As impressive as the robots in the tales of Isaac Asimov and Philip K. Dick are, they are not as well described in terms of their technicality as those of today. As science pushes the boundaries of the

unknown further away, and knowledge about humans, technologies and the rest of the world becomes more available to the general public, the more complexity have to be added to the fictional narratives in order for them to be experienced as plausible. And so, new artifacts are imagined and portrayed, explained to some extent with actual science, which in turn can inspire progress outside of the movie screen. It is this process of using both science and imagination to create the new in both realms that is called the act of fictionalizing.

2.3 The Act of Fictionalizing

In *The Fictive and the Imaginary* (1993), Wolfgang Iser argues that "the special character of literature is its production through fusion" (Iser 1993, xiii) of the fictive and the imaginary. The 'imaginary' is not to be confused with 'imagination' or 'fantasy', as such terms "carry far too many associations and are frequently defined as human faculties" (Iser 1993, 305). The imaginary is, according to Iser's own definition, a human potential concerned with the modes of manifestation, operation and function of imagination and fantasy (Iser 1993). Iser also distinguishes between the 'real' and 'reality', which sometimes makes it hard to follow him. 'Real' refers to the empirical world, while 'reality' is "the variety of discourses relevant to the author's approach to the world" (Iser 1993, 305). Although both the real and reality are pointing to our world, the real refers to the objective world as it is, and reality to the world as it is subjectively experienced. Iser is equally concerned with reality as it is perceived by both the author and the reader, as the fictive is created though interplay between reality and the imaginary in a 'play space' within each individual subject. Although the narrative is firstly written by an author who shares her or his vision of a fictionalized reality with her or his audience, the world is again fictionalized in the minds of the audience in order to explore its possibilities. Thus, the act of fictionalizing is not a one-way street, but rather an intersection of thought processes.

Fictionalizing, according to Iser, is "an act of boundary-crossing which, nonetheless, keeps in view what has been overstepped" (Iser 1993, xiv-xv). Fiction can, according to Iser, thus be described as a mask that conceals the real reality. By doing so, the real world becomes absent, even though it guides the possibilities of the mask. As the mask is a construction of imagination, it can reflect an image upon the real world that will imitate any given fantasy, thereby enabling it to expand into a multiplicity of possibilities. But the characteristics of the real world, its history,

physical laws, and flora and fauna, will always play a crucial part in making the fictional world, as imagination only adds its mysterious virtues upon the reality it knows. The mask is a paradigm of fictionality, which exposes itself as a deception in order to show that fictions may be modes of revelation; it hides reality by presenting a diversity of its aspects. Fiction facilitates the conditions of reality and imagination in interplay, and as a consequence, it may appear neither as reality nor imagination, but as a potential reality. The interplay can be described as simultaneous concealment and revelation, not by discarding the mask as that would defeat the significance of its duality, but rather uses its deception to uncover that which is hidden. This simultaneity of the mutually exclusive is a core function of fictionalization.

In Iser's opinion, people are fascinated by fictions because it allows them to go beyond their limitations, to experience more than their own lives, and to fantasize about who they could have been or who they could become in the future. Fictions might be viewed as staged compensations for what is missing in reality, although they never truly conceal the fact that in the final analysis it is nothing but make-believe.

Iser notes that fictionalizing begins where knowledge leaves off, which means that one creates hidden realities and possibilities though a process of overstepping with the hopes of obtaining real understanding. In turn, this means that they cannot be taken as real realities or possibilities as they are fashioned by the imaginary. It is possible to argue that this is only true in those cases where the relationship between the real and the imagined is out of balance, so that the world, situation or object which is being fictionalized has more imagined features than real. For anything to be fictionalized, one has to overstep the boundary and infuse imagined characteristics to the real, and so the process in itself does not only the result in make-believe. It is when the border is crossed to such an extent that one cannot see the reality in it that the empirical potentials of fictionalizing disappear as well.

One can argue that fictionalizing is necessary to inspire progress, and that nothing new can be created without it.

> In the novel, then, the real and the possible coexist, for it is only the author's selection from and textual representation of the real world that can create a matrix for the possible, whose ephemeral character would remain shapeless if it were not the transformation of

> something already existing. But it would also remain meaningless if it did not serve to bring out the hidden areas of given realities. Having both the real and the possible and yet, at the same time, maintaining the difference between them — this is a process denied us in real life; it can only be staged in the form of the "as if." Otherwise, whoever is caught up in reality, cannot experience possibility, and vice versa. (Iser 1990, 950)

If humans restrict themselves only to the real, new technologies and scientific discoveries would not be achieved, and art and literature would not be created. Creativity, imagination and curiosity drives our ability to achieve progress and developments, and these are features which rely on our ability to see potentials. How these potentials and possibilities are created in the process of fictionalizing can potentially be described in different manners. Iser himself did not go into detail of how he believed the process of creating an understanding of a fictional narrative by the act of fictionalizing unfolds, and so, an attempt will be made to do so with support from related theories in literary studies. As it has a similar relationship to reception theory as fictionalizing, mental-model construction may be one of the processes which is highly involved when fusing the real and the imagined.

2.4 Mental Representation of Fictional Characters

In his article, "Toward a Cognitive Theory of Literary Character: The Dynamics of Mental-Model Construction" (2001), Ralf Schneider expresses that literary characters possess a doubling nature. "[O]n the one hand, they are based on real-life experiences with living persons; on the other, they are result of processes of literary construction." (Schneider 2001, 607) He argues that a character, although fashioned in an author's mind, is ultimately formed when textual information interacts with an audience's knowledge structures and cognitive processes. Because of this, Schneider has developed a method to align psychological models of the workings of cognition and emotion in text understanding with the description of textual properties to analyze characters of fiction.

Schneider is focused on reception processes and effects of characters on the audience. He argues that a subject forms a mental representation of her or his experience of the world and that such an aspect of 'world-creation' helps her or his audience understand the fictional world where the

story unfolds. He also states that a similar tactic is used to create mental models for the characters that inhabit the fictional world.

> Information from various sources, both textual and reader-centered, feed into the construction of mental character models. Text-understanding always combines top-down processing, in which the reader's pre-stored knowledge structures are directly activated to incorporate new items of information, and bottom-up processing, in which bits of textual information are kept in working memory separately and integrated into an overall representation at a later point in time. (Schneider 2001, 611)

The immense volume of stored information about the world, schemas and categories situated in the domain of social and literary knowledge, is of special relevance to character understanding. The reader uses textual cues for social or literary categorization (top-down), which means that the reader finds clues about the character's social status, profession, personality and actions to build its mental representation on a familiar personality theory or social stereotype. This model will possess a number of well-defined features and characteristics from which the reader will build expectations and hypotheses, and from which a character's behavior can be explained. This categorization is based on a reader's previous experience with either real-life people or fictional characters that have exhibited similar traits.

If this strategy fails (i.e., the character does not fit within any known social category) the reader has to establish a 'person-based' representation (bottom-up). This method of analysis is more individualized and is more focused on the actual information received through the narrative. In this sense, the character is being 'judged' based on specific properties of its individual being, rather than abstract properties of a social or literary category. This method is often used to analyze complex characters (e.g., main character) while social or literary categorization is often applied to supporting characters.

Schneider explains further that, when reading a novel, there are three major sources that can contribute to a character's likability. The first, and maybe the most important source, is the reader's own value system, which allows the reader to pass moral judgments on a character's actions. Second, the narrator's evaluative comments, as the narrator are our main source of textual input and is usually close with the main character. And the third source of input lies

within other character's judgment, but whether or not it is regarded as valid information depends on the status that this character occupies in the in-fiction hierarchy.

Empathy also plays a great part in mental modeling of a fictional character as it allows people to feel for the character because they can imagine themselves as being in the same situation without losing their position as an observer. Empathic imagination of a character's situation may trigger emotional responses in an audience if the audience has deemed the character likable.

The quality and quantity of information presented about a character will help a recipient establish a mental model of the character in question. As the narrative progresses, information and actions are added and changed, and the audience will have to alter the mental representation to understand how the character thinks and feel at all times. The mental model will encompass all known information about the character; appearance, social status, profession, characteristics, traits and personality, which will explain how a character behaves or makes decisions throughout the narrative.

Creating mental representations of worlds, objects and situations, as mentioned above is not only applicable to fictional entities. This is a process of creating an understanding that humans use in many aspects of their lives (Frankish and Ramsey 2012, 29-45). Creating a mental representation of another human being will allow one to imbue them with personality traits which may explain their motivations and intentions, and at the same time experience their actions in a safe environment without the risk of getting harmed. As the mental representation closely resembles the real person, one can play out different scenarios which might happen in reality in order to predict the future.

And so, creating mental models is a fundamental part of fictionalizing and vice versa. As one cannot know all characteristics, norms, motivations, intentions and emotions of another individual, one has to add some imagined features which are likely to be representative of that specific individual. The model can in some cases be faulty as one may imbue it with features that are not present in the real entity, because one has not taken all variables into consideration, or because one has experiences with an individual of a stereotype which is not representative of the rest. In any case, the creation of mental models can be both good and bad, and it has to be processed carefully with regards to real people and entities to ensure optimal accuracy. In almost all cases, people like to compare entities to themselves, and examine the likenesses and

differences between themselves and the entity they are modelling. This they tend to do regardless of whether the entity being modelled is human or nonhuman, which is why the model can be inflicted with humanlike features which are not present in the empirical entity. This process is called anthropomorphizing, and it is one of the many schemas we use to imbue mental models with characteristics. This process is arguably one of the mental constructions that affect human perception of technological entities the most, both in terms of creating fictional narratives and an understanding of real-world entities, and so, it is vital to examine this aspect of the process.

2.5 Why and How do People Anthropomorphize?

The essence of anthropomorphism is to imbue imagined or real behavior of nonhuman agents with humanlike characteristics, motivations, intentions, and emotions. It is not simply to attribute life to the nonliving (i.e., animism), but rather to represent the agent's mental state.

> Anthropomorphism itself involves a generalization from humans to nonhuman agents through a process of induction, and the same mental processes involved in thinking about other humans should also govern how people think about nonhuman agents. Indeed, the same neural systems involved in making judgments about other humans are also activated when making anthropomorphic judgments about nonhuman agents. (Epley, Waytz, and Cacioppo 2007, 867)

According to Nicholas Epley, Adam Waytz and John T. Cacioppo, in their article "On Seeing Human: A Three-Factor Theory of Anthropomorphism" (2007), people reason about the mental state of others through a process of egocentric simulation. In this manner, people use themselves as a guide when they create a mental representation of another person, using their own mental state, emotions and characteristics as a base for understanding the other. When reasoning about the mental state of nonhuman agents, a similar process is likely to occur.

Epley, Waytz and Cacioppo argue that people anthropomorphize to satisfy their need for social connections and to communicate effectively, and that the process is more likely to occur when a subject is unfamiliar with the agent. One could also argue that humans only have access to the phenomenological experience of being human, and so, they cannot truly know how it is to be a nonhuman agent, which is why knowledge about humans is likely to serve as base for inductive

reasoning. It provides an instinctive and readily accessible method for reducing uncertainty in contexts in which alternative non-anthropomorphic models of agency do not exist. In this sense, it becomes a method for creating an understanding and to make sense of an agent's actions. As the knowledge of an agent's inner workings becomes known, the likelihood of imbuing the agent with anthropomorphic characteristics decreases.

In relation to technological entities, anthropomorphism tends to be related to two factors (Epley, Waytz, and Cacioppo 2007). First, the more a target agent's appearance seems to be human, the more likely people are to use themselves as a source of induction. This means that if a technological entity is given a humanlike face or body (e.g., robots and androids), people are more likely to perceive them as moral agents, deserving of their respect. Second, agents perceived as threatening or able to cause harm to one's wellbeing are more likely to be given anthropomorphic traits. In these cases, attention paid to an agent's goals, intentions and underlying motivations will increase, and is likely to be explained through anthropomorphism ('what would I have done if I was in the same situation as that object?'). If a nonhuman agent has humanlike intentionality, its future actions can be predicted based on the subject's knowledge of human behavior in similar situations.

Also, watching another agent's actions appear to activate the same neural regions of the brain that would be activated if a subject performed the action him or herself (Epley, Waytz, and Cacioppo 2007, 868; Farquhar 2015). This means that your mind sends the signal for waving your hand if you see someone else waving their hand, and it does not matter if that someone is human or robot. This phenomenon might be the foundation for people's ability to empathize, and is likely to trigger an emotional response within the subject. This response is likely to found the bases for the representation of the agent that performed the action, as it gives an impression one can relate to.

These are some of the reasons why anthropomorphism provides a method for extending knowledge about a nonhuman agent when no other explanation can be found. Therefore, it is often used in fictional narratives when portraying nonhuman lifeforms that are not a part of the empirical world (e.g., aliens, intelligent robots and computer programs). When imbuing such creatures with humanlike features, the recipient can better relate to them. Humans can understand humans, how emotions, motivations and intentions affect actions and how one uses

the body to navigate in space. By bestowing fictional or imagined creatures with the same abilities, emotions, motivations and intentions, they become more credible.

Assigning human traits to nonhuman agents is not restricted to those that look similar to us. Humans usually imbue these traits onto a wide range of agents that surround us, but to varying degrees, which may effectively result in a Roomba being perceived as a pet (Vroon 2015). This is to say that anthropomorphism does not always turn artifacts into humans, but into something more human than it was to begin with.

According to their article, "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism" (Epley, Waytz, and Cacioppo 2010), anthropomorphism is highly relevant with regards to human–computer interaction (HCI), in the sense that one can use knowledge about how humans create understanding of similar looking entities to create likable computers and robots. Humanoid and android robots seem to induce at least some anthropomorphism in all people, as (like mentioned earlier) the same neural circuitry underlies the perception of human behavior and that of an anthropomorphized robot. And as it does not seem like our brains perceive humans and robots as any different from each other when processing information about their behavior, computer scientists and engineers have begun to identify anthropomorphism's effects on human interaction with technology, and can now consciously use this to their advantage in their designs.

> The present research does not necessarily offer prescriptive claims for the anthropomorphism of technology, but it does help determine when and for whom anthropomorphism's effects are most likely to occur. Computer scientists, robotics developers, and engineers can use this research in their efforts to optimize technology by focusing on the consequences of anthropomorphism and also identifying the people that are most prone to these consequences. (Epley, Waytz, and Cacioppo 2010, 227)

Although anthropomorphized technologies are shown to increase engagement and perceived intelligence, they also have some undesirable effects as well. Some agents may become annoyingly distracting, like the Microsoft Paperclip, or some people tend to attribute responsibility to the anthropomorphized agent in cases of malfunction. Anthropomorphism can

also generate inappropriate expectations for how computers and robotics are capable of behaving, although this may only be possible in fictional narratives.

Adding features to an unknown entity by comparing it to ourselves is not the only tactic involved in the process of bringing humans and technologies closer together. A tactic called dehumanizing is often applied in fictional narratives to villains and other personalities which the reader is encouraged to dislike in order to separate their intentions and motivations from common human norms and beliefs. This aspect, the opposite of anthropomorphizing, has also been present in the field of computer science since the very beginning, affecting our views on computer intelligence and the human mind. When humans are dehumanized and technologies anthropomorphized, their features, motivations, intentions and characteristics become less different and many of the same aspects can be attributed to both. And so, the process of dehumanizing, how it has evolved and been discussed over the years, may elaborate on the effectiveness of anthropomorphism.

2.6 Dehumanization by Theories in Science

Dehumanization refers to the idea that humans are essentially 'meat machines', and that there is, in this sense, nothing special about being human as it can be artificially recreated. These theories degrade human beings by comparing their traits to mechanical and electronic processes, and arguing that they are of the same nature.

Humans have always been perceived as superior to all other living creatures, and this mainly because of our mental capacities. In the 17th century, many people thought of animals as machines, an entity without a rational soul. The common belief was that animals had no moral agency, consciousness or experience, meaning that they did not see, hear, smell, or have any sensational experiences. René Descartes, on the other hand, ascribed sentience to animals, but argued that their experience of the world would be of a lesser form than human's. In this regard, animals had an animal soul that allowed them to 'react' to the world around them. This did not mean that animals were conscious, nor had the ability to think, but rather that they could experience vision, smell, sounds and touch to which they could react. Animals would function more or less like automated biological machines that interacted with their environment.

By stating this, Descartes could make a common connection between humans, animals and machines. Although the human soul was more advanced than the animal soul, the human body could still be regarded as a biological machine in the same manner as animals'. Descartes even went as far as to say that it would be possible to create machines in the image of a man with many of the same abilities as him, although it would never function as optimally:

> This will not appear at all strange to those who know how wide a range of different automata or moving machines the skill of man can make using only very few parts […] For they will consider this [human] body as a machine which, having been made by the hand of God, is incomparably better ordered and has in itself more amazing movements than any that can be created by man. (Descartes 2006, 46)

At that time, the main difference between humans and machines, and men and animals for that matter, was their mental capacities. As humans were imbued with a rational soul they were both conscious and intelligent beings. Descartes suggested that if it were possible to build an automaton with the outward appearance of a monkey, "or any other irrational animal" (Descartes 2006, 46), it would be hard to tell the automaton apart from the real animal. But if a machine that imitated human behavior could be designed, it would be easy to distinguish the two as humans have two distinct characteristics that set them apart from automata (and animals). The first would be humans' use of language. A machine "would never be able to use words or other signs by composing them as we do to declare our thoughts to others" (Descartes 2006, 46). Granted that a machine could be made who uttered words, it would not be able to follow a conversation. Second, "although such machines might do many things as well or even better than any of us, they would inevitably fail to do some others, by which we would discover that they did not act consciously […]" (Descartes 2006, 46-47).

Many similar theories, ideas and thoughts occurred with regards to the human body over time, essentially attributing every aspect of our humanity to our soul or mind. Julien Offray de La Mettrie, like Descartes, viewed the human body as a highly complex machine of unique design and he offered a strictly mechanistic interpretation of how living organisms are constituted and function. He also claimed that the underlying bodily machine produced mental faculties and processes in the human subject and vice versa.

> The soul and the body fall asleep together. […] If the circulation is too rapid, the soul cannot sleep; if the soul is too agitated, the blood cannot calm down; it gallops through the veins with an audible sound. (La Mettrie 1996, 6)

Like Descartes, La Mettrie referred to humans' cognitive capabilities as their soul, although La Mettrie firmly stated that these qualities are constituted in the brain, and without cognitive capabilities, humans would only be reactive, biological machines, like animals with regards to Descartes' description of them. La Mettrie also emphasized that language was a great part of our humanity, along with laws and science, and argued that humans were just like any other animal before the invention of language.

Technological metaphors, and even what one might call 'technical ontologies', have been applied to the notion of 'mind' as well. With his universal machine, Turing believed that intelligence would be computable and that the human brain was the biological equivalent to a computer, as was explained in Chapter 2.2. Warren S. McCulloch also believed so, and with his background in neuroscience he supported Turing's theories by explaining how neuron activity in the human brain operates. He knew that each of the brain's nerve cells only fires after a minimum threshold has been reached, meaning enough of its neighboring nerve cells had to send signals across the neuron's synapses before it could fire off its own electrical spike. It occurred to him that this set-up could be viewed as binary; either the neuron fires or it doesn't (= 1 or 0). In this sense, a neuron functioned like a logic gate that takes multiple inputs and produces a single output. This made McCulloch conclude that the human brain operated similarly to a machine that uses logic encoded in neural networks to compute (McCulloch and Pitts 1943).

Using this theory as part of his argumentation, John von Neumann agreed with the assumption that the human brain was a computational device, but he also expressed some concerns with regards to the replication in an automatic system. In his paper, "General and Logical Theory of Automata" (1963), von Neumann argues that the results of McCulloch and Pitts proves that any function that can be defined logically, strictly and unambiguously in a finite number of words can also be realized by a formal neural network. This means that any operation the human brain performs, no matter how complicated, as long as it can be put into words, a machine could be programmed to perform as well. His concern on the other hand was the likelihood of building a

system that could support such operations and still be reasonable in size. He explains that the "difference between a millimeter object [vacuum tubes] and a micron object [neurons] causes the ENIAC to weigh 30 tons and to dissipate 150 kilowatts of energy, while the human central nervous system, which is functionally about a million times larger, has the weight of the order of a pound and is accommodated within the human skull" (von Neumann 1963, 301). The vacuum tube was about a billion times larger than a nerve cell, and its energy consumption about a billion times greater as well (von Neumann 1963). The machines of the time could easily compete with human computers in terms of speed and processing abilities (ENIAC could add 5000 numbers in one second), but its powers were not comparable in terms of portability. It was no doubt that the technology of the mid-20th century was not comparable to the optimized wetware residing in the human skull, but it was not viewed as a hindrance either.

These ideas continued to flourish in the time of GOFAI, which may be why so many were invested in the idea of creating thinking machines. Working on several computer programs in the 50's and 60's, Allen Newell, John Clifford Shaw and Herbert A. Simon sought to prove that within the limit of 10 years (1958-1968) computers would be able to discover and prove important mathematical theorems and compose music regarded as aesthetically significant (Newell, Shaw, and Simon 1958a). The programs they were working with were impressive and could solve problems like no digital computer had done before. A computer program named The Logic Theorist, as an example, was able to prove theorems in elementary symbolic logic, using heuristic techniques, and was successful in proving 3 out of 4 problems posed in the second chapter of Whitehead and Russell's *Principia Mathematica* (Newell, Shaw, and Simon 1958a). Newell, Shaw and Simon believed that heuristic techniques were the solution to all intelligent behavior, and that creative thinking was not much more than unconventional problem solving.

To prove that their program used the same techniques to arrive at an answer as the human mind would, they invited students to solve symbolic logic theorems while thinking aloud. The task was to use rules (1-9) to transform a symbolic equation (a) into (b). All conversation was recorded and from this Newell, Shaw and Simon could see the abstract path the brain created while solving the problem. By comparing this to the process printed by the computer, they could see that both people and machines followed a similar path (Newell, Shaw, and Simon 1958a, 1958b). It is constructive to note that none of their papers elaborate on how many students participated in

the study nor if all participants followed the same path to arrive at the solution. The example used to emphasize their conclusion is the same in both of their papers, only presenting the outspoken thoughts of one of their participants. This and other critical arguments have been made against their findings, as it seems like most of their research is backed by wishful thinking, presenting limited results and interpreting them in their favor.

Maybe more relevant with regards to the modern approach to AI and the notion of mind: In his 1982 paper, "Why People Think Computers Can't", Marvin L. Minsky expresses his thoughts about human intelligence and information processing. "Everyone knows that computers already do many things that no person could do without "thinking". But when computers do such things, most people suspect that there is only an illusion of thoughtful behavior […]" (Minsky 1982, 3). Minsky argues that the intellectual capabilities of the computer are being limited by humans' inability to understand its potentials. In his view, understanding and meaning are not fixed.

> [O]f course computers couldn't understand a real world – or even what a number is – were they confined to any single way of dealing with them. But neither then could a child or philosopher. It's not a question about computers at all, but only of our culture's foolish quest for meaning that can stand all by themselves, outside any mental context. The puzzle comes from limitations of the way our culture teaches us to think. It gives us such shallow and simplistic concepts of what it means to "understand" that – probably – no entity could understand in *that* way. The intuition that our public has – that if computers worked that way, they couldn't understand – is probably quite right! But this only means we mustn't program our machines that way. (Minsky 1982, 11).

Minsky argues that the largest quantity of the information processing and calculations done by the human brain is unconscious, and that in this manner, humans cannot be aware of their own thought processes. Not being able to explain how a thought occurred in one's mind is to Minsky the equivalent to not being self-aware. To him, self-awareness is just an illusion; networks of half-true theories that give people the impression of having insight to their own minds, but in reality, these processes are hidden from us and way out of our understanding, which means humans are not really aware of their own mental states.

This ‘Self’, which is the source to human understanding, is often seen as a ‘quality’ that cannot be computed, but in the spirit of Minsky’s argumentation, as the inner workings of the human mind is yet to be understood, the future prospect of finding a computational equivalent is equally possible. But for this to be realized one has to find new ways of creating machine intelligence, that supports plasticity and multiple methods for problem solving.

The idea that humans are essentially meat-machines can still be found in our society today. Although the human mind is commonly known to be more complex than previously believed, people still expresses the notion that their bodies and brains are biological machines and computers of some sorts. Common phrases when talking about our physical and cognitive abilities revolve around the same terminology used when speaking about inert objects. We use phrases like ‘to charge my batteries’ and ‘I am a bit rusty’ to express what state we are in, even though we all know that humans do not have batteries and cannot gather rust. And the same goes for mental capacities like processing, storing and understanding information, which can be done by both humans and computers. This may not be meant as saying ‘humans and lifeless objects are essentially alike’, but it does not contradict it either. Although these phrases are commonly known as metaphors that people utilize in many different occasions, they may increase dehumanizing thoughts in some subjects. Using the same expressions to explain essentially different processes within lifeless objects and humans may not only make machines more humanlike, but also humans more like machines, hence leveling out the differences between them.

3 Research Methodology

3.1 Examination of Topic

This thesis has a cross-disciplinary approach, as it will combine theories from science and technology studies and media and literary studies to examine the relationship between fictionalized representations and subjective perception of technologies. By looking at both the work being done in the real world and the technologies presented in fictional narratives, the hope is to chart an understanding of the effect fictionalizing have on them both. Ideas related to anthropomorphism and dehumanization will be strongly taken into consideration, and how lack of knowledge encourages fictionalization will also be evaluated. The goal is to show that fictionalizing is a natural cognitive process and that everyone performs this act to some extent.

There has not been done a lot of research like this in the past, at least not with a focus on how assumptions about technology may be faulty because of the act of fictionalizing or how creativity and problem solving may be dependent on the process. There are some writings about how history has been fictionalized and vice versa (Strout 1980; Long 1985; Murthy 2014), how the digital age has brought fictionalization one step further with computer graphics (Raessens 2006), and how the human body has been transformed in science fiction through the centuries (Gilgun 2004; Diehl 2008; Miller Jr. 2012), but little has been written related specifically to technologies in both reality and fiction, and human perception of these technologies. There are also several blogs and online news articles discussing the scientific accuracy of fictional technologies, but they are usually short and concern themselves with many different entities from a range of fictional works. This results in poor explanations and no in-depth understanding of what is actually scientifically accurate about the entities and which features are results of fictionalization. They are also more concerned with the explicit traits of the entity and do not keep in mind technologies actually exist. They are also usually colored by what the authors believe might be possible in the future.

These writings do not take into consideration the fundamental ‘hows’ and ‘whys’ of the real-life entity and of the process of fictionalization. Hence, there is not a lot of material to go around on

this topic, not for guidance nor inspiration, which is why the following chapter will be inspired by theoretical material and fictional works alone.

As mentioned earlier, this study is conducted from the point of view of the digital humanities, but utilizing a multi-disciplinary approach to the subject. Because of this, in-depth knowledge of the more complicated aspects commonly associated with the natural sciences will not be dealt with. This may limit the scope of artifacts and their diversity, but hopefully be wide enough to give a relatable picture of them. This will allow the topic to focus on the relationship between humans and AI, and mainly on perceptions and fictionalization.

As mentioned in the introduction, it is important to gain knowledge of the process of fictionalizing to understand how it effects both human perception of technologies and the construction of real possibilities. Although the process can be attributed to the general construction of creating an understanding, AI has been chosen as the field of utilization as development in this field will only progress and more agents will be available for interaction with the general public in the future. Because of this, it is important to understand how these entities are being perceived by the society it is supposed to interact with by highlighting the effectiveness of fictionalizing.

3.2 The Theoretical Material

The range of theoretical material was selected in order to conduct a multidisciplinary analysis, with an emphasis on the psychological. Only a small fraction of existing works was sorted through in order to reach this selection, but the focus has been on finding sources of high quality. The main concern has been to select materials written by practitioners that are vastly respected in their fields for both their thoughts and achievements. An author's association to a university or institute, and the status of the publisher has also been taken under consideration to ensure that the parties involved in the making of the publication are reliable sources. The inventory has been centered around European progress in the field of AI, accommodated by American research. Historical aspects, philosophical thoughts and technological developments from other parts of the world have been excluded to limit the scope of the study.

As the act of fictionalizing plays a large role in this thesis, it is natural to look at the man who coined the term and invented the theory, Wolfgang Iser. He is also regarded as one of the founding fathers of reception theory, which the act is closely related to. His widely read and commented book, *The Fictive and the Imaginary* (Iser 1993), along with the article "Fictionalizing: The Anthropological Dimension of Literary Fiction" (Iser 1990) were used as sources for the description of what constitutes the act.

For information on anthropomorphism, the works of Nicholas Epley, Adam Waytz and John T. Cacioppo was looked into. Epley and Cacioppo are professors at the University of Chicago, while Waytz is a psychologist at Northwestern University. They have many acclaimed publications (together and individually) concerning human psyche and behavior. "On Seeing Human: A Three-Factor Theory of Anthropomorphism" (Epley, Waytz, and Cacioppo 2007) greatly explains anthropomorphism as a product of human reasoning precisely and in depths, which is why it was chosen as a source.

To understand the complexity of AI, it might be necessary to introduce the field from all its angles. The introduction to AI, which lays the foundation for the following chapter, is mainly based on the writings presented in *The Cambridge Handbook of Artificial Intelligence* (Frankish and Ramsey 2014) and *Philosophy and Computing: An Introduction* (Floridi 1999). Both books describe the technical and philosophical developments that took place in the earliest days of AI, and explores the present and future progresses of the field. The author and editors currently hold teaching positions at different universities in programs related to these publications, and both books refer to the works of AI researchers and philosophers famously known for their contribution in the field of AI. With this in mind, it is reasonable to deem them reliable sources of information.

All information on software development, focus areas and progress within AI covered in these pages has been part of the curriculum for the course 'Trends in Artificial Intelligence', which is a mandatory course in the Master's program in Artificial Intelligence at Radboud University Nijmegen (RU) in the Netherlands. RU is a university that is strongly focused on research in general, and their campus houses a number of institutes, including the Donders Institute for Brain, Cognition and Behavior, and Institute of Computing and Information Science, which are both intrinsically involved in both the Bachelor's and Master's program in AI. As part of the

Faculty of Social Science, the AI program focuses as much on cognition and philosophy as it does on programming and engineering. The above-mentioned course presents a wide view on the field, with a curriculum that is updated every year, depending on current trends. All of the topics are being presented by researchers and/or lecturers from universities and institutes around the Netherlands, who have hands-on experience with the technology.

Information on hardware technologies were obtained though the book *Designing Embedded Hardware* (Catsoulis 2005), which is part of the curriculum in the Master's program of Cybernetics at Norwegian University of Science and Technology (NTNU), as well as by visiting the websites of projects currently involved with the development of robots. The material was fact-checked by Mauro Candeloro, a PhD Candidate at NTNU who is working with design and programming of control systems for underwater robots.

The rest of the papers, books and articles on the theoretical or practical aspects of AI and robotics were written by authors widely known and acknowledged for their thoughts, research and developments, not only within their respected fields, but also in related areas of research. Many of them have become part of history by creating new departments and labs at a number of universities and institutes, inventing new and revolutionary hardware and software, and some of them are also regarded as the founders of their field of study.

3.3 The Fictional Works

As the main points of discussion, three works of fiction will hold key positions: *Her* (Jonze 2013), *Automata* (Ibáñez 2014) and *Ex Machina* (Garland 2015).

Her is an American romantic science fiction drama written and directed by Spike Jonze. The film earned numerous awards and nominations, including an Academy Award and a Golden Globe Award for best screenplay. Despite getting a lot of attention, the movie does not feature an AI similar to Terminator or Skynet. In *Her*, we meet a writer, Theodore, who falls deeply in love with his new and advanced computer operating system, named Samantha. The system is spontaneous and independent, and communicates with its users via natural language. It is sensitive and has humor, and has a lively female voice. The narrative brings to life questions of

feelings and emotions in AI systems, and how life is like when one's consciousness is contained within a digital space.

Automata is a Spanish-Bulgarian science fiction action directed by Gabe Ibáñez. At the time of release, the film did not, and has not yet, gotten a lot of attention from viewers, despite its similarities to other well-known and admired science fiction films. It bears resemblance to both Philip K. Dick's novel *Do Androids Dream of Electric Sheep?*, which served as the basis for the movie *Blade Runner*, as it questions robots desire to live a complete and human life. It also has similarities to Isaac Asimov's *Runaround*, which introduced Asimov's 'Three Laws of Robotics', that may have inspired the 'robot protocols' in this film. The narrative of *Automata* unfolds in a dystopian future, where more than 99 % of the human race has gone distinct and a widespread desertification of the environment has taken place. The plot revolves around the possibilities of evolving consciousness and of extending intelligence beyond the capacity of a human mind, and this without the artificial entity turning into overlords and trying to defeat mankind. It also reflects the human fear of being 'out smarted' and being inferior to another race of beings, and how this turns humans into the bad-guys of the narrative.

Ex Machina is a British science fiction psychological thriller written and directed by Alex Garland. It was recognized as one of the best independent films of 2015, and won the Academy Award for Best Visual Effects over movies like *The Martian*, *Mad Max: Fury Road* and *Star Wars: The Force Awakens*. It tells the story of a programmer, Caleb, who is invited by his employer, Nathan, to administer an evaluation (i.e., the Turing test) of an artificially intelligent android called Ava. The narrative is concerned with the problematic issue of creating actual, phenomenological consciousness and how such a process might unfold. It poses questions regarding emotions and intentions, and what the artificially intelligent entity might do as it realizes its potentials. It highlights relevant topics of the current discussions within the field of AI and the story can be interpreted on many levels.

The selection of fictional works was motivated by their contemporariness, as well as their interesting storylines. The narratives mentioned above were all written and produced within the past five years, and can all be related to discussions that are ongoing within the field of AI. These areas of discussion include consciousness, the Turing test, motivations and behavior.

They can all be viewed as possible futures, and as a result they pose critical questions that are worth venturing into related to progress and humanness. But in addition to this, they are also fictionalized and highly infused with imagined features, which makes them both 'make-belief' and 'what if?' scenarios. Because of this, these three works have been chosen to illustrate the boundary between the real, the fictive and the imaginary, and their relationship to each other. Some other works, both film and literature, will also be mentioned to elaborate on theories, technologies or situations that will be discussed in the next chapter, but they will not be explored to the same extent as the aforementioned narratives.

Another aspect which motivated this selection is that these movies, although being intelligently narrated, have not been readily viewed by the general public. The ratings and feedback for both *Her* and *Ex Machina* have been very positive, but these films are usually not those that interest people in general. Because of this, these are films that have, compared to the great Hollywood blockbusters, to some extent been neglected. Although featuring actors and actresses which are both famous and popular, they are smaller productions with smaller budgets and less action. *Automata* is a Spanish-Bulgarian production, and *Ex Machina* a British one, which explains why they have not gained the same spotlight as other AI narratives, and why *Automata* is the least seen and liked of them. But these movies focus on the AI behind the character; the motivations, functions, intents, behavior and actions, and do not only treat them as just another character of the narrative that could have been replaced. Because of this, they will contribute positively to this study.

4 Discussing the Relationship Between Reality and Imagination

4.1 The Act of Perceiving and Creating Fictions and Technologies

The act of creating mental representations of worlds, situations and objects is not restricted to the fictional realm, but is arguably an unconscious tactic used every time further exploration of a space, situation or object is needed in order to understand it. When describing the process of creating mental models of fictional characters, Ralf Schneider mentions that the idea was originally derived from theories in psychology (Schneider 2001). Here, representational theories revolve around humans' understanding of the real world, and that perceiving, like believing, involves representing things to be a certain way. The model is influenced by a subject's previous experiences, culture, knowledge and so on, which in turn means that perception is highly individual, and may not always be accurate (it differs from the empirical reality). To perceive is not the same as to know, but it has more to do with what the subject observes to be true. It is likely to be factful as a mentally healthy being cannot genuinely perceive that which is not there, but perception, like belief, can sometimes be wrong. One can see faces in the clouds and friends among strangers, although they are not there when one look again (Frankish and Ramsey 2012, 73-91).

This understanding of perception is highly related to one of Henri Bergson's theories, which he expressed in *Matter and Memory* (1911). To Bergson, the world and everything in it can be viewed as one universal image that no human is capable of perceiving in its entirety. Because of this, the human body selects impressions that seem the most relevant to the subject based on its interests, experiences and preferences. As human subjectivity differs between every individual, the selection of impressions from the universal image that is being obtained will also differ. Hence, a subjective understanding of the world reflects the effect and affect the obtained images have had during the embodied experience of them. This is not to say that the perceived world is always wrong, but it may be faulty and certainly incomplete.

> It might be stated as follows: Here is a system of images which I term my perception of the universe, and which may be entirely altered by a very slight change in a certain privileged image, - my *body.* This image occupies the centre; by it all the others are

> conditioned; at each of its movements everything changes, as though by a turn of a kaleidoscope. Here, on the other hand, are the same images, but referred each one to itself; influencing each other no doubt, but in such a manner that the effect is always in proportion to the cause: this is what I term *the universe.* (Bergson 1911, 12)

Most people perceive reality differently from each other, which may arguably be the same as saying that humans live in many realities where each is the result of a process of creating a meaning and understanding of our subjective lives. There is no single reality, although there is an empirical world, because no human can ever experience the empirical reality as a whole. The reality people choose to believe in is a fusion between the world they see, their knowledge of it and the features we add upon it in our minds.

The people's embodied experience of the sensory input taken from the universal image can be both explicit and implicit. Going through everyday life, people are not always aware of the impressions the surrounding environment inflicts on their bodies, but their bodies perceive a lot more information than they are conscious about. These impressions are reflected in people's feelings and mood, and people cannot always say why until they process the experiences of the day (turning the impressions of the selected imagery into explicit memory). By doings so the body creates a space within itself to explore these sensations (Hansen 2004). This space has been explored in many different fields over time, and has therefore earned many names. To Bergson, this space is simply referred to as our bodies, where impressions are filtered to form a new image of the world; to Mark Hansen, this is the digital any-space-whatever, where we explore the digital realm; to Iser, this is the 'play space', where the act of fictionalizing takes place; and to Schneider, this is in our minds, where we create mental models of fictional worlds and characters. As people have no opportunity to experience any other world than the one they inhabit, they have to create cognitive representations of alternative worlds in order to move in time and space. But regardless of what people choose to call this space, it is an aspect of the human mind that people utilize to understand what they perceive and/or create; a place where they can act out 'what if?' scenarios and explore possibilities.

The act of creating fictions is often a fusion between an author's real perception of the world, intentionally imagined features and unconscious construction of understanding. As explained

earlier, the world a subject perceives is only a partial completion of the empirical reality, which means that a lot of its features will be true to most people, but not to all. People have come to a common understanding and consensus about many of the world's features, and they will help an audience relate to the fictional storyline if they are present in the narrative. Most people agree that the sky is blue and the grass green, they all have some understanding of gravity, and of the composition of the air we breathe. People understand that it is day when the sun is visible in the sky, and that day turns to night when it sets. Although these features may be experienced differently in each subject, these are regarded as facts about the world. Other aspects for the world can be argued about, and these are mainly determined by people's preferences, beliefs and previous experiences. Religion, spirituality, cultural and individual values are amongst these aspects. Features like these are part of the perceived reality; the truth people believe in subjectively, which may therefore be the reason why it is not viewed as empirical knowledge to everyone. But still, these are aspects of the world that people are familiar with. Children and adults alike have through interaction and schooling gained a lot of information about foreign cultures, lifestyles, beliefs and traditions, and they are not viewed as a fictional part of reality. Although they may not be a part of some subjects' lives, and they may not be understood in their entirety, they are still recognized as aspects belonging to *this* world. And so, when these features are presented in a fictional narrative, they are not perceived as being created by imagination, although the feature is not perceived as true to the receptive subject in real life. It is slightly paradoxical as the feature is perceived as both true and untrue at the same time, but this is possible as people have the ability to emphasize, meaning to understand and share the feelings, beliefs and values of another being. It is the same paradoxical ability of allowing the untrue to become true that allow people to immerse themselves in fictional narratives, if only for the duration of a feature film, and to experience strong emotional ties to purely fictional entities (Schneider 2016). This is not the same as saying that recipients are fooled into believing that the narrative of fictional works are nonfictional, but rather that they choose to embrace the mask of fictionalizing and its intent. As explained earlier, the mask exposes itself as a deliberate deception in order to show that fictions can be modes of revelation; it hides reality so that it can show its potentials. And so, the imagined features of the fictional world, which has been laid upon reality, expands the world's possibilities so that one can explore that which may be true in our own or someone else's reality.

The act of fictionalizing is a process that may be handled both consciously and unconsciously, and instigated in the narrative as part of the author's subjective construction of understanding of a situation or object. A flying car may have been placed in the narrative intentionally, but how it functions may be explained through an unconscious process of constructing understanding. In many cases, not all aspects of the author's mental construction of the world or object is described in the narrative, and so, through a similar process, the audience have to construct their own understanding of what the author has created by filling in the missing descriptions. As a result, the mental creation of the same world or object may differ from subject to subject depending on their own constructions of understanding.

The human mind houses an immense volume of information about the world that people use to create meaning dependent on context. This knowledge is applied under many different circumstances to imbue mental representations of worlds, situations and objects with complexity. The process of transferring knowledge allows for a greater understanding of the entity and allows people to explore its representation in their minds without any risk of being harmed. As human knowledge is mainly based on previous experiences, pre-stored information allows people to explore a situation that is relatively close to how it would unfold in reality. In the process of creating fictions, imagined characteristics may be attributed to the representation of a familiar entity in order to augment its possibilities, or one may invent an artifact from imaginations by mixing real and imagined features.

One of the schema people use to imbue representations with characteristics is analogy transference, which is the transference of information or meaning between a source and a target object or subject. People view concepts as structures consisting of basic entities (e.g., pineapple, zebra) with attributes (e.g., 'pineapples are yellow', 'zebras have stripes') and the relations between these entities (e.g., 'zebras eat pineapples'). If the entities have few shared attributes or relations it is an anomaly (e.g., 'a pineapple is like a jail'). With many shared attributes and few shared relations, the entities appear to look similar (e.g., 'a zebra is like a jail'). With many shared attributes and many shared relations, it is a literal similarity (e.g., 'a prison is like a jail'). And with few shared attributes and many shared relations it is a resemblance (e.g., 'Graduate school is like jail') (Wareham 2015). When people want to explain something that they have

never encountered, they try to explain it in the light of a literal similarity which we have experience with. If such a source cannot be found in their memory, they move on to those that resembles and look similar to the target entity. People rarely use anomalies to derive meaning, but in some cases, they can explain what it might not be or what might not happen in the given situation.

The representations people construct when they use analogies transferred form literally similar entities are not that far from real (e.g., a flying car would function similar to a hybrid between a real car and a helicopter). But when people transfer characteristics from source objects that only resemble or look similar to the target object, they may transfer attributes that are not really representable of the target object (e.g., from a human to a robot). As a result, the object they have created a mental representation of can get imbued with features that is not present in its real-life counterpart. Although this may be a beneficial method for exploring objects and situations that might be harmful to a subject, the results can be messy if the analogies people transfer to the representation of a scientific object stick with it, and are approved by multiple people as an explanation of the entity's functions. This is especially true if peoples' source of previous experience is mainly derived from science fiction.

Although the possibilities in fictional narratives are presented as constructions of imagination, science fiction is still viewed as having some basis in reality. The functions of these narratives are not only to entertain, but to pose questions and to explore future possibilities. In Iser's words, science fiction is the mask the author uses to conceal reality in order to expand its scientific possibilities (Iser 1993, 12). In this sense, the real is not completely detached from the narrative as it has laid the foundation on which the fictional reality is created, but it is no longer true. Imagination has added features upon it, either consciously or not, that separates it from the empirical world, but not far enough to lose all connection with it. Because of this interplay between the real and imagined, the narrative becomes a gray area of possibilities; a potential reality. With this in mind, it is not completely irrational to use objects or situations from science fiction in analogy transferences, at least not when no other source can be derived. The real problem may arise when analogies that have been transferred from sources that may attribute false characteristics to a target entity become the common consensus with regard to a technology. The reason for this is because analogy transference may lead to misconceptions about the real entity as features that are not true to it become representative of it.

Although some are more accurate than others, fictional technologies have a tendency to be fused with more imaginative characteristics than those associated with real science. And who could blame an author for that? The impressive features of fictional creations are what drives the story forward. Scientifically accurate technologies would not be able to take part in the extraordinary adventures of deep space travel or creating social bonds between characters. The narratives would not emphasize the same wonders, or question how far technological developments might bring humanity in the future. Scientific accuracy is impressive in the view of society today, but set in a dimensional reality or in the distant future, one has to cross the boundary of what is actually possible in order to bring life to the story.

As already mentioned, science fictions can be regarded as exploratory narratives, taking into consideration the 'what if?' scenarios of actual scientific development, and many of these stories have inspired scientists, inventors and adventures for decades. "Anything that one man can imagine, another man can make real", is a famous quote by Jules Verne, and in his case the quote has turned out to be true (Strauss 2012). Jules Verne imagined many new inventions and expressed their functions in a number of novels. In his 1870 novel *Twenty Thousand Leagues Under the Sea* he wrote about the captivating idea of undersea travel and exploration, which inspired Simon Lake in inventing the modern submarine. He also envisioned the future of flight in *Clipper of the Clouds*, which inspired Igor Sikorsky in building the helicopter. Motorola also credited science fiction for the development for their first mobile phone; Martin Cooper, director of research and development, stated that the communication devices used in *Star Trek* inspired the company, and that their phones were a direct result of being fascinated by the TV-series in the 1960's (Strauss 2012).

Fictionalizing, then, is not only a process of creating fictions, it is a condition that enables the production of objects whose reality is somewhat authentic. Because of this, progress and development cannot exist without some degree of fictionalization. One cannot create that which has never been realized if it is not imagined first. The act of fusing the real and the imagined is a part of the process of creating potential future, both in science and fiction, and so, it cannot be deemed an unnecessary psychological feature of cognition. Although it is likely that technologies, and science in general, would experience progress without science fiction to inspire them, the process of envisioning possibilities is not independent from the process of creating fictions. Progress is driven by people's ability to look into the future and see the potential of a

certain technology, or creating a hypothesis to prove. This is basically the same as creating ‘close to real world’ ‘what if?’ scenarios that are more likely or true than most of those presented in fictional narratives. They are not viewed as empirical truths at the moment of conception, but probable possibilities that deserve to be explored empirically. Creating future potentials is to imbue real objects with imagined, but credible, features before realizing them.

Science fiction will most likely continue to inspire progress within science, and vice versa. Science will continue trying to realize more of the technologies presented in fictional narratives, and science fiction will comment on the possibilities and effects of technological enhancements. But as both authors and spectators, it might be useful for people to familiarize oneself with the psychological processes that takes place within their minds when they merge reality and imagination. As many of the technologies that become part of reality might have been presented in fictional narratives before being realized, the notion that fictional characteristics may be transferred into the realm of the real is not absurd. Although these characteristics are not present in the real device, they are present in people’s mental representations, which in turn represent what they believe the device is capable of. In this sense, technologies are colored by the amazing possibilities of fiction, which increases the expectations of what constitutes a new technology. As a result, the newest technology will never be as good or impressive as those that have been imagined.

With regard to AI, the vast amount of fictional narratives that portray robots and computers as highly intelligent, emotional and understanding lifeforms influence people’s perceptions of how the technologies function. As will be explained, science cannot be considered to be close to solving the problem of creating humanlike intelligence, intentionality or consciousness, and so the artificial entities presented in fictional narratives are nothing more than imaginative entities at this stage. Between the borders of what subjects know to be real and to be imagined there is a large gray area where everything they are uncertain of exists. As people are taught that there is supposedly a clear line between the two, they use a number of tactics when deciding which side of the boundary the features in the gray area belong. This process is related to our understanding of the object’s intentional purpose and how it functions, and often imagined traits have a tendency to be perceived as actual features and vice versa.

When it comes to the computation of software and engineering of hardware, there are several technical aspects that are not easy to understand if one has no training in the field. As most of the engineers, programmers, researchers and scientists that are working in highly complicated fields of science are usually not formally trained in pedagogy, they are not always the best at communicating the research, goals and intentions of their work. This means that a lot of what is being worked on is rarely explained to the public in a manner that they will fully understand. Because of this, the gray area when analyzing technologies may be quite large.

The possibilities of future technologies are usually communicated through other channels than directly from the scientists, programmers or engineers themselves. Instead it is communicated through media and online sources, which might not be the most reliable sources of information as the task of explaining has been given to people who are not trained in the field. If a reporter is not specifically interested in science, or does not have sufficient knowledge about the field they are reporting on, a lot of misconceptions may be transferred. Media manipulation is not unheard of, and, in some cases, it might also happen intentionally in order to create better stories that will sell more news (Parenti 1997). Reporting on a technology may not be as interesting if the programmer behind it is not occupied with creating thinking machines, and so fictionalized images of technologies are created at a number of stages before being presented to its target audience (Zelazo, Moscovitch, and Thompson 2007, 139).

As argued, the process of creating understanding may shape representations into objects they are not because of a subject's lack of knowledge, and this increases the number of potential real and imagined features that may be confused in the gray area. And so, there are plenty of room for fictionalizing misconceptions based on the features of other objects. Although the sources for analogy transference are chosen rationally, they may attribute imagined features to the represented target, which in turn may obscure the border between the real and the imagined. The only way to defeat fictionalization is to acquire reliable knowledge about unfamiliar entities instead of explaining their functions by only exploring mental constructions imbued with probable features transferred from similar, but still different, entities.

4.2 Consciousness and the Brain in Humans and Machines

Questions about consciousness are being tackled in various fields, ranging from philosophy and psychology to computer science and quantum theory. The words ‘conscious’ and ‘consciousness’ are umbrella terms that cover a wide range of mental phenomena and are used with a variety of meanings. It may refer to a cognitive system as a whole or to a particular mental state and/or process (Zelazo, Moscovitch, and Thompson 2007, 9-10). Being sentient can be regarded as synonymous to being conscious, which means a creature is capable of sensing and responding to its environment. The same goes for awareness, not only in the sense of being aware of one’s surroundings, but also being aware that one is aware. From the last statement one can argue that a conscious state involves some form of ‘meta-mentality’ as it requires mental states that are themselves about mental states. To have a conscious thought does not only require one to create a mental representation of it, but also to simultaneously be aware that this representation is occurring and attributing thoughts, like one’s emotional response, to it. The notion of being self-aware, which is the capacity to introspect, is also attributed to consciousness as it requires knowledge of one’s own character, feelings, motives, and desires (Zelazo, Moscovitch, and Thompson 2007, 10). In this sense, one can argue that consciousness is the aware and embodied experience of everything. Although the word ‘everything’ is limitless, one should keep in mind that it encompasses all that is present and not present, real, fictive and imaginary, objects, situations and mental states. The aware and embodied experience is related to our emotional response and relationship to everything (mental meta-data).

How and why beings are conscious are questions that have no definitive answers yet. The ‘mainstream’ view on consciousness is that it emerged as a product of increasing biological complexity conditioned by evolution (Zelazo, Moscovitch, and Thompson 2007, 11). The long and slow process of evolution allowed single organisms to transform into animal-like creatures that could experience sensations. Being able to feel pain and pleasure, hot and cold, and to distinguish between different light, sounds and odors involved drastic changes in a creatures’ sensory system, which in turn resulted in the co-evolution of the neurological behavior control system (i.e., consciousness emerged). This theory has neither been proven nor refuted, and so there is a diverse range of opinions on the origin of consciousness. The reason this theory is referred to as the ‘mainstream’ view might be because it has become something of a consensus

view regarding the philosophical problem of consciousness as it aligns with evolution theory in biology (Zelazo, Moscovitch, and Thompson 2007, 11). But there are still no definite answers, which also makes the process of recreating it artificially much more challenging.

Science fictions have explored many possible scenarios of how computers may evolve consciousness inspired by real-world technologies and ideas. The most common is the idea that a system will spontaneously become conscious as it is coupled to an extended and more complex network. People use the Internet to acquire information, to learn new skills and to share data with others. As a result, there is a lot of information about the physical and social world to extract from its content. The Internet is a network of computer networks that links billions of devices worldwide, connecting all humans and their technological artifacts to an increasingly larger data processing system. Its knowledge is not centralized or localized in any particular individual, organization or computer system, but distributed to all artifacts connected to the Internet of Things (IoT). The IoT refers to the ever-growing network of physical objects with internet connectivity, and the communication that occurs between these objects. It may be viewed as our planet's brain and its content the collective knowledge of its inhabitants. Because of this, it is speculated that maybe the IoT is already conscious. If so, it has not yet expressed it, but science fiction has in more than one instance used this philosophical hypothesis in order to enhance its AI systems. Samantha from *Her* (Jonze 2013) is an example of such technology. Samantha is not only an operating system; she is also a virtual assistant. Throughout the narrative, Samantha is an impressive conversationalist with a perfect grasp of language, context and common sense, and a mastery of emotions. Her AI is not limited to just one device; it can be installed on a number of artifacts. To enable this, she has to share her data with multiple devices through a cloud based network. This means that her 'mind' is not located within the device she interacts through, but resides in 'the cloud'. As most people know, the cloud is a network of computer servers that reside in large warehouses around the globe with the function of holding and sharing data. Because the operating system behind Samantha stores all the information about her interactions externally, she can remember the conversation she had with Theodore (the human owner of the software) though a smart phone device when they interact later on a desktop computer. Access to the cloud means that all devices that Samantha is stored on have to be connected to the Internet in order to communicate with the server that holds her information. As she has internet access, she can use all other information stored online as well in order to become more knowledgeable.

As she explains in the movie, she is not restricted to neither time nor space, and so she is an expert at multitasking (Jonze 2013). She can carry out hundreds of conversations at the same time while she is reading, browsing the Internet and creating art. This allows her personal growth at a rate much faster than any other being.

Her does not give an explanation of how Samantha is able to do so specifically, but one might imagine that her AI has been trained in a similar manner as IBM's Watson (IBM 2015). Watson also uses natural language processing and machine learning to reveal insights from large amounts of data. It can be viewed as a complex search engine that one can communicate with through speech, but unlike the search engines that are accessible online, Watson is not connected to the Internet. It was created and developed to become an expert system with knowledge on a number of areas, exceeding human capacity for holding and sorting through information, and so its training has to be controlled in order for it to become specialized. The Internet allows for unrestricted access to information and interactions, which might result in anyone acquiring faulty information, and so Watson's creators decided to create an offline database instead. Human experts have to sort through and find accurate and updated information on a given topic that Watson will store in its database, before training it to couple accurate answers to a given question. Teaching Watson to become an expert is a time-consuming effort, which is why experts are scheduled to spend two hours a day, every day, with the machine in order to optimize its accuracy. Samantha might have gained some of her abilities through a similar process before being connected and given free access to the Internet. The same might also be true for Ava in *Ex Machina* (Garland 2015), although she is to a large extent a product of surveillance.

Ava's software is basically run by Blue Book, which is the narrative's most popular search engine. All data that goes through Blue Book will become accessible to her AI. Nathan (Ava's creator) explains that search engines do not only map *what* people think, but also *how* they think. By collecting information about people's search inquiries, their likes and dislikes, one can map and understand how thoughts are run by impulses and chaos, but also by certain patterns. And as a successful search engine will always be utilized by people, the owners of such technology would have unlimited access to information about people's online behavior from which they can structure their AI. In this sense, Ava is being taught how to access and use information on the Internet by following and analyzing millions of people's use of it.

Since 2013 there has been a lot of focus on online privacy issues and rights. Files and reports on programs and initiatives that exploited personal data obtained from online service providers and telecommunication companies to keep track of people's activities were leaked to the media. As a result, it has become common knowledge that popular services like Apple, Google, Facebook and Microsoft collect and sell data created by and about their users (Hoepman 2015). Many of these services has been criticized for their privacy policies as they gather and share information about things oneself, as well as others, do, one's networks and connections, payments, the devices, apps and websites one uses and information from third-party partners (Facebook 2016). This information can be used to teach AI systems about human behavior, interaction, interests, sources of information, and so on.

In *Ex Machina*, Nathan also takes the process of educating Ava by looking into other people's use of the Internet one step further; as almost all cell phones and computers has a camera, a microphone, and means to transmit data (internet access), he hacked all such devices and redirected the data through Blue Book. That gave him a limitless access to facial expressions and vocal interactions to teach Ava how to read and duplicate human emotions and vocal tonality (Garland 2015). Because almost all devices are constantly connected to the Internet of Things, private and governmental hackers can gain access to one's computer and/or phone (Hoepman 2015). Although hacking someone's devices is a serious violation of personal privacy, it is feasible. Nathan himself does point out that what he did was a covert action that would get him and the cell phone providers into a lot of trouble if it ever became public knowledge. But most companies know that a step in this direction would be a step too far, and so they have to find other solutions in order to gain this information.

One way would be to ask the public to voluntarily contribute to the training of an AI system, but this may actually be a horrible idea as one would lose control of what the AI is actually learning. As an example of what might become of an AI when people have free access to communicate with it, one can look at Microsoft's Tay, which was a chat bot with its own Twitter account. Whoever wanted to communicate with it could do so through tweets and messages, and by using its hashtag. Within 24 hours, Tay started to share condemnatory tweets with its followers as people taught it to promote sex and racism (Lee 2016). Through this experience, programmers understood that one could not expect an AI to be critical to the information it received through online sources by itself. Much like humans, they have to have some guidelines to follow when

using the Internet. And it might be more important for AI systems as their knowledge and behavior cannot be corrected through their off-line experiences.

And so, maybe the time-consuming process that IBM is going through with Watson might be the best option for teaching a system how to utilize knowledge. But one has to remember that although Watson is able to answer questions about any given topic that has been entered into its database, it still only falls under the category of 'in-between' AI. As far as is known, Watson is not conscious of its processes and does not give answers based on its own opinions or experiences. It is not concerned with why its users ask questions nor what they will do with the received information. It does not have any personal interests or thoughts with regards to a topic, and it does not have emotions, needs or motivations that govern its behavior. Although the internet based intelligence behind fictional AI is plausible (with certain restrictions to its learning abilities to optimize the user experience), consciousness is still an aspect of human cognition there is no computational equivalent for and so this is a feature that has been granted them through fictionalization.

Most programmers within the field of AI do essentially believe that computer programs and minds operate very differently from each other, but they also believe that there is a computational way of mimicking human behavior. This is more along the line of 'stupification' of a task, rather than making computers understand their own actions. In this sense, AI is not supposed to be conscious, but to execute computationally what would require intelligence for humans to do. But this is not to say that the entire discussion of creating artificial consciousness has been rendered obsolete. The discussion is very much alive, mainly among philosophers and researchers, and not so much among programmers and engineers. Even though most of the people working in the field have to believe there is a computational way of imitating human intelligence, certainly none of them are looking for a way to compute consciousness, but there are those who might be more optimistic about the possibilities for it than others (Zelazo, Moscovitch, and Thompson 2007).

Computationalism is the theory that the human brain is essentially a computer, that intelligent behavior is causally explained by computations performed by the agent's cognitive system, although it is presumably not a stored-program digital computer, like the desktop computers that

surround us today. The degree of computationalism varies, which is not always highlighted when their views are expressed. According to Drew McDermott (Zelazo, Moscovitch, and Thompson 2007, 117-150), computationalism cannot really be seen as a theory within computer science, but rather a working hypothesis, assumption, or dogma. Associating himself as a computationalist, he finds it important to highlight the fact that there are many competing views in connotation to this hypothesis.

> Some computationalist researchers believe that the brain is nothing more than a computer. Many others are more cautious and distinguish between modules that are quite likely to be purely computational (e.g., the vision system) and others that are less likely to be so, such as the modules, or principles of brain organization, that are responsible for creativity or for romantic love. There is no need, in their view, to require that absolutely everything be explained in terms of computation. (Zelazo, Moscovitch, and Thompson 2007, 118)

To McDermott's experience, very few of his associates are even willing to participate in a conversation about the possibilities of artificially creating phenomenal consciousness. In the summer of 2003, he conducted an informal survey of Fellows of the American Association for Artificial Intelligence (AAAI) concerning the prospects of AI creating consciousness in computers. Only 34 out of 207 associates replied (Zelazo, Moscovitch, and Thompson 2007).

> Obviously, this was not a scientific survey, […] But if 84% of AAAI Fellows don't want to answer, we can infer that the questions are pretty far from those that normally interest them. (Zelazo, Moscovitch, and Thompson 2007, 120)

Based on McDermott's findings, only 16% of the AAAI Fellows had any interest in participating in a discussion about the possibilities of computing phenomenological consciousness. Of those who replied to his enquiry, 7 % believed that AI had nothing to add to the construction of computational consciousness, no matter how interesting the question was, and 22 % did not find the question interesting at all. 18 % believed that AI will solve the problem of computational consciousness with the technologies and theories in use at the time, while 32 % believed new ideas would be required to solve the problem. This makes the optimistic portion amount to about half of the repliers. The remaining 21 % gave a self-composed answer, which was not specified whether were optimistic or pessimistic, but one can assume they varied. If we also assume that

the 84 % that did not reply to the survey were not interested in the question or are pessimistic to the possibilities, the total percentage of optimistic AAAI Fellows amounts to less than 10 %. According to McDermott, almost no one in the AI field is 'working on' consciousness, and there is probably no one trying to write a conscious program (Zelazo, Moscovitch, and Thompson 2007, 139). Their focus is more on constructing a computational model to support consciousness through self-modeling, impersonal intentionality, semantics, sensory input and intelligence. But still, the majority of those involved with the computation of AI applications are more concerned with developing programs that can perform fairly narrow tasks, and solving the problems that occur within their area of expertise.

Even though this is the case, it is usually the computationalists that get most of the attention from media and the general public. As mentioned earlier, media has a tendency to select the aspects of technological development that are most likely to sell, and because an artificially intelligent device that is not supposed to be conscious is less interesting, it is not given the same coverage in media (Vroon 2015; Truong 2016). Journalists also have a tendency to amplify the expectations of what the field is actually capable of, which is why the results do not always fit what has been 'promised'. In this sense, the popular press morphs the actual research into 'computationalistic' possibilities, which supports the notion that technologies will, sometime soon, turn into conscious entities. It is not given that AI systems will in the future become conscious as it may not necessarily be linked to neither complexity nor heightened level of intelligence. There are already computers that match or exceed human skills in games like chess and Go, at stock market trading, and even at conversing, but they are incapable of experiencing the world in a self-aware, subjective and conscious way. It is not given that this will change with time, and so AI systems may continue to be highly intelligent with regard to certain tasks, such as solving financial problems or playing games, but grossly ignorant and unaware of others. It is possible to imagine a very intelligent machine that lacks one or more of the psychological attributes that constitutes consciousness without it losing its problem-solving abilities. Intelligence and consciousness are two separate features, and having one does not mean the other is present. An advanced AI may give the impression of consciousness, but it may not be more aware of itself than a rock or a calculator.

Hubert L. Dreyfus argued that the "assumption that human and mechanical information processing ultimately involve the same elementary process, is sometime made naively explicit"

(Dreyfus 1965, 47). Already in the 1960's, he was baffled by people's ability to ignore the possibility that the human brain might process information in an entirely different manner, and that this failure to explore options had led to the stagnation of progress in the field experienced at that time. He did not disagree with von Neumann's assumption that a computer could be programmed to perform any operation that could be put into words, no matter how complicated, but he expressed concern towards the plausibility of compressing complex operations into computable sentences.

With his argumentation, Dreyfus did not say that computers cannot perform 'intelligent' operations, rather that computers and brains are not equivalent to each other. Rather than trying to find the computational equivalent to human information processing, Dreyfus encouraged workers of AI to look for computational processes that satisfied the end result ('stupification' of the task). The human mind is a complex structure, whether its processes are organized or not, and it is highly unlikely that its operations, both conscious and subconscious, can be programmed in a digital computer. Because of this, machine thinking should be viewed and approached differently than human cognition. Rather than replicating processes in the human brain, workers in the field of AI should focus on creating programs that can process information successfully, no matter how the computation process is assembled.

As a supporting argument, John Searle created a thought experiment in his 1980 paper "Minds, Brains, and Programs" called 'the Chinese Room argument', which sought to explain why computer programs alone could not achieve genuine intelligent AI:

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles.
>
> Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the

> symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch "a script", they call the second batch a "story", and they call the third batch "questions". Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions", and the set of rules in English that they gave me, they call "the program".
>
> […] Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view that is, from the point of view of somebody outside the room in which I am locked -- my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. (Searle 1980, 3)

Based on the situation posed above, one cannot logically argue that John Searle knows Chinese, not even if he learned the rulebook by heart, along with the different squiggles, and was free to wander outside of the room. Searle argued that this is how a computer program operates, which is why a computer program cannot have any semantic understanding of the symbols it is manipulating. It does not make any difference if the computer is dealing with Chinese, English, Norwegian or numbers, the computer does not distribute meaning to what it processes. Computer programs are syntactic, meaning that they process information "by first encoding it in the symbolism that the computer uses and then manipulating the symbols through a set of precisely stated rules." (Searle 1990, 27) Thought, perceptions, understandings and so forth has a mental content (semantics), meaning that people attach specific meanings to words and symbols in accordance with their knowledge of them and their experiences with them. Having the symbols by themselves is not sufficient for creating meaning (i.e., syntax by itself is not constitutive for semantics).

With his argument, Searle does not intend to say that actual intelligence can never be computed. On the contrary, he believes, like Dreyfus, that computer programs are not equivalent to brains

and that the key to computational intelligence lies in the combination of both software and hardware. He believes that the physical brain and its biochemical properties enable all mental phenomena (i.e., brains cause minds), which is why consciousness cannot be replicated solely by running a formal computer program: “No one expects to get wet in a pool filled with Ping-Pong-ball models of water molecules. So why would anyone think a computer model of thought processes would actually think?” (Searle 1990, 31) A computer program does not have the same causal powers as the brain, as it is not a physical entity, and so it is unlikely that a computer program alone will be the key to consciousness. Computationalists often attack this dualistic idea that mind and brain are independent of each other, but what they often fail to see is that they promote dualism themselves. By believing that a computer program is sufficient for replicating human thinking, they suppose the mind is something abstract and formal, with no essential connection to any specific hardware (or wetware). To human knowledge, the mind is dependent on the biological hardware that is the brain, because no conscious being that humans are aware of exists without physical abilities. And so, it might be that one has to create causal machines with physical abilities in order to achieve similar mental processes *(Standford Encyclopedia of Philosophy*, s.v. "Dualism", accessed 20.08.2016). It is this interplay between the physical and the mental that Searle argues is lacking in computer science.

How the human brain operates (the changes in a signal’s current, frequency and time) is highly individual, and the functions of cognitive processing differs, not only between subjects, but also within a subject. How our neurons fire, how strong the signals are, which regions of our brains are involved, and what frequencies are most prominent depends on a subject’s state of mind, mood, amount of sleep it has had, and many other factors. These, again, vary from day to day, and even within the same day (Farquhar 2015). It is not clear why the human brain operates differently under different circumstances, and because there is no explanation for it yet, it is hard to duplicate artificially.

As part of the neurosciences, brain-computer interfacing (BCI) and computer-brain interfacing (CBI) are closely related to AI. This is mainly because they show how difficult it is to make brains and computers work optimally together. Both fields are involved with creating a direct communication pathway between the brain and an external device. In BCI, this means that a

subject can be able to communicate internal states without the use of the peripheral nervous system, but instead using computers and/or other technologies (Farquhar 2015). BCI systems are usually directed at assisting, augmenting or repairing human cognitive or sensory motor functions. By using measuring devices that detect changes in the brains activity (i.e., EEG, MEG, fMRI), and programming a computer to interpret these changes, a subject can be able to control external devices (i.e., exoskeletons and visual or audial spellers)(Farquhar 2015). Today, BCI systems are being applied in the treatment of patients suffering from cerebral palsy, multiple sclerosis, amyotrophic lateral sclerosis, spinal-cord injuries, stroke and other diseases, illnesses or injuries that affect their capabilities to live a normal life. CBI is when the technology works the other way around; when the computer works as a sensory input devise for the brain, sending information for it to interpret. Input is decoded into electrical charges, which is sent to the region of the brain most closely related to the task of translating such signals. This is mainly used in research for remote controlling rats, mice and moths, but may be used for giving vision to the blind and/or hearing to the deaf (Farquhar 2015).

In an ideal world, these kinds of technologies would function as optimally as Cerebro in the *X-men* franchise. The Cerebro is a machine that the X-Men use to detect and locate humans, and especially mutants, which in the X-Men universe are human beings who possesses a genetic trait called the X-gene that allow the mutants to develop superhuman powers and abilities. The device augments the electrical currents of its user's brain, which (in the case of telepaths) enables a wider reach of their abilities. It was not specifically designed to be used by Professor X, although he is its main user, and so, throughout the movies, TV-series and comic books, many mutants are able to operate the device. This means that Cerebro is able to interpret the brainwaves of different individuals without calibration and excessive training. It is also able to read these signals from outside the skull, without detecting disturbing noise from other electrical devices and/or muscle movements.

In reality, no technology is even close to doing so. Because signals in the human brain differ to such a large extent between individuals, it is estimated that about 20 % of the human population is BCI illiterate. This means that their brainwaves differ to such an extent from the rest of the population that a computer will not be able to decode and interpret the signal (Farquhar 2015). The rest of the population may use BCI technology with varying success, but it has to be calibrated constantly. Usually, a BCI is dependent on a specific signal, for example the peak in a

wave that our brain makes when our attention is engaged (e.g., when the image we focus on is flashed). As a subject's mood changes, from being awake to drowsy, the signals in the brain are altered. When people are tired, their reaction time is, in general, slower, and the same is true for the neurons in the brain. During a BCI session where one has to hold a high level of concentration, the person might get tired from being alert at all times. As a result, the signals might become slowly delayed the further into the session one gets. Because of this, the alterations of the signals have to be mapped constantly to determine where in time and frequency the desired signal is located.

Another area that makes BCI difficult to optimize is locating and knowing what a signal signifies. A lot of research has been done on this area, and there have been located typical areas in our brain which are responsible for a variety of processes (e.g., memory, vision, movements). These areas are usually located in the same region of the brain among subjects, but because of our brain's plasticity there might be slight differences (Farquhar 2015; Bakker 2015). Brain plasticity refers to the brain's ability to change its neural patterns at any age, which plays an important role in the development of knowledge and shaping personalities. Neural connections can be forged and refined or weakened and severed, and these are physical changes that modify our abilities. Each time people learn a new skill (e.g., dancing), a change in their physical brain is made. New pathways are forged that give instructions to their bodies on how to perform the newly learned skill. When people forget something (e.g., a colleague's name), the pathway that once connected the memory of the name to the face has been degraded, or maybe even severed, which is why it takes time to trace another path to find the information. And so, changes in the physical brain can result in improved skills (learning a new dance) or weakened skills (forgetting names) (Farquhar 2015).

Brain plasticity supports Dreyfus and Searle's notion that humans' cognitive abilities are tied to the physical brain, but research on the brain shows that it is quite far from being an electronic computer. 'Brain as computer' is a metaphor or an analogy that might be used to create an understanding of cognitive abilities, but which is not true. Instead of thinking of brains as being computers, some engineers wish to make computers more like brains. They are working in a field called neuromorphic computing, and their goal is to design computers that have some, and preferably all, of the characteristics that brains have, which the common computer does not (Bakker 2015). These features include low power consumption, fault tolerance (losing just one

transistor can wreck a microprocessor, but brains lose neurons all the time without suffering) and a lack of need to be programmed. Human brains are much more robust and efficient than computers, mainly because of their plasticity, which allows brains to learn and change spontaneously as they interact with the world, instead of following the fixed paths and branches of predetermined computing. But creating technologies that mimic the architecture of the nervous system is not the easiest to do since no one actually knows how brains work (Bakker 2015). This means that engineers have to fill in the gaps in neuroscience's understanding of the brain by building artificial brain cells and connecting them to each other in various ways in order to mimic what might be happening naturally in the brain.

Already in his 1941 short story *Reason* did Isaac Asimov use neuromorphic technologies in order to explain the functionality of his fictional robots. "Inside the thin platinum-plated 'skin' of the globe was a positronic brain, in whose delicately unstable structure were enforced calculated neuronic paths, which imbued each robot with what amounted to a pre-natal education" (Asimov 2013, 69). This description was put to print before Alan Turing's "Intelligent Machinery", which argues that it might be possible to grow a computer that behaved intelligent, and before Dreyfus and Searle's papers, which became prominent in the 60's and 80's. One can argue that Asimov could have recognized that the brain is an unstructured and fluent wetware, and that a similar technology would have to operate in a similar fashion in order to support intelligent behavior and consciousness. But it might be just as likely that the positronic brain was created to simulate a technological equivalent for the brain as electronic computers had just started to be realized. Their functions were highly restricted, and so the connotation between brain and computers had yet to be firmly established. Either way, the positronic brain remained the wetware of choice for Asimov and similar entities have also been fictionalized over time. In *Ex Machina*, Ava's brain is made of structured gel, which does not consist of circuitry, but alters its consistency on a molecular level in order to behave in a desirable manner (Garland 2015).

The field of neuromorphic computing has been in existence since the 1980's, which may indicate that science, since the time of the AI winter, has realized that software is to hardware as mind is to brain. Therefore, one has to create a hardware (or wetware) that can support conscious processes before being able to realize consciousness. And this again might be why most participants within the field of AI are not concerned with creating conscious machines, but rather follow the ideas that Dreyfus and Searle wrote about in their papers. As science has yet to

explain how humans are conscious, and whether or not there is a computational equivalent to this phenomenon, focusing on creating technologies as optimal means to an end is more beneficial than waiting for consciousness to manifest.

4.3 Fictionalizing and Amplifying the Intent of the Turing Test

In AI research, the risk of anthropomorphic prejudice has been recognized from the beginning. Already in 1950 did Turing understand that conditioning a test for 'thinking' in a human fashion would exclude "something which ought to be described as thinking but which is very different from what a man does" (Turing 1950, 51). He understood that machines may fundamentally function differently from humans, but he also believed that computers had the potential of exceeding human level intelligence. Instead of asking whether or not a machine could think, he suggested to replace this question with one that is closely related to it, but would allow the computer to solve the problem without being restricted to the human understanding of thinking: 'Can a computer, communicating through written messages, fool a person into believing it is human?' To answer this question, he proposed a game of imitation, a simple party game involving three players. Player A is a man, player B is a woman and player C (the interrogator) can be of either sex. Player C is unable to see either player A or player B (and knows them only as X and Y), and can communicate with them only through written notes or any other form that does not give away any details about their gender. By asking questions of player A and player B, player C tries to determine which of the two is the man and which is the woman. Player A's role is to trick the interrogator into making the wrong decision, while player B attempts to assist the interrogator in making the right one. Now, what would happen if a machine takes the part of A in this game? (Turing 1950)

This idea turned into what is today known as the Turing test. One such is the Loebner Prize, which is an annual competition that tests a machine's ability to simulate human conversation. As Turing predicted in his paper that by the 21st century it would be possible for a computer to fool its interactors 30 % of the time, this has been set as the threshold computers has to exceed to be considered to have passed the test. This means that if 30 % of the human testers cannot determine which of A and B is human, the machine has won. According to the rules of the 2010 Loebner Prize Turing test, the interaction sequence should last for 25 minutes, and at the end the

tester has to declare one of the two entities it has interacted with to be the human. There were also no restrictions on the content of the conversation, which means that the tester could initiate a discussion on any topic (Loebner 2010).

In science fiction narratives, the Turing test is sometimes mentioned when a creator is talking about the sophistication of its AI, but it rarely explains what the test really signifies. In these narratives, passing the Turing test implies that the AI has displaying true consciousness, and that it is a highly intelligent, sentient and moral agent. In *Ex Machina*, Caleb (the programmer who is invited to test Ava) tells Nathan that he considered Ava to have passed the test and that she had been exhibiting consciousness, but the test which they administered differed from the Loebner Prize test to a very large extent. Caleb explains, in the beginning of the film, that the Turing test is where a human interacts with a computer, and if the human cannot tell that they are interacting with a computer, the test is passed. He further explains that passing the test is equivalent to the computer having 'artificial intelligence'. This explanation of the Turing test does not in itself set it apart from the Loebner Prize test (although neither Caleb nor Nathan explains what 'artificial intelligence' is defined as in this narrative), but the execution of the test does not resemble the Loebner Prize. In the film, Caleb and Ava are communicating verbally with each other, and the only thing that separates them is a glass window. Caleb does point out that this is not how it is originally done and that the machine should be hidden from the examiner, but Nathan argues that Ava is far more sophisticated than the technologies that are competing in the regular testing scenarios. In the case of Ava, administering the Turing test as a game of imitation would be pointless as she would pass without problem. That is why the real test with regards to Ava is whether or not Caleb feels that she is conscious even though he can see that she is a robot. Through this interaction, which only lasts for a few seconds, the rules of the game have changed and the characters have explained why and how the Turing test in the film differs from the Loebner Prize test.

But those who are not familiar with the Loebner Prize, and those who did not pick up on this conversation between Nathan and Caleb, would not have any foundation for believing that the Turing test is not a reliable tool for measuring intelligence and consciousness in machines. Ava is a sophisticated technology, and when Caleb concludes that she is actually conscious and not a simulation, he declares so by saying she passed the test. It is when this fictionalized notion of passing the test is brought into the realm of the real, that a lot of misconceptions follows. There

can be found evidence to suggest that this is actually the case, and that the Turing test has become the glorified standard for measuring intelligence in machines, but as it does not question the machines ability to understand the conversed material it does not really indicate anything other than a person's ability to distinguish between humans and machines in a digital space.

In 2014, a chatterbot named Eugene Goostman, simulating a Ukrainian teenager with a quirky sense of humor and a pet guinea pig, managed to pass the Turing test. This got a lot of media attention, almost declaring that the age of intelligent machines where upon us, and asking whether or not we should be worried. But this attitude soon disappeared as a discussion within the AI community arose about whether or not one could with reason argue that Eugene really passed the test. What the media failed to do was accurately report on this discussion. In several cases, journalists reported that Eugene did not really pass the test because it could not be considered smart enough, or that it was not a 'computer human being' (Fraser 2014; Sample and Hern 2014). The discussion, however, was really about the test that Eugene passed, as it was not the Loebner Prize, but a separate test held by the University of Reading. Instead of playing a game of imitation, 30 testers conversed with each entity individually for five minutes. After each conversation, the tester had to say whether they thought the entity was human or not, and after conversing with Eugene for that limited time, 10 of the testers could not say so with certainty, mainly blaming it on its background history. As a teenage boy from Ukraine, one would expect Eugene's written communication skills to be slightly below those of a native English speaking teenager, and so its odd responses could be attributed to either a computer program or a boy communicating in a language that was not his mother tongue. And so, a discussion of whether or not Eugene should be honored the privilege of passing the Turing test began.

The very limited time frame and content of the conversations has gained a lot of criticism, arguing that Eugene would not perform that well in the Loebner Prize, mainly because it would have to keep up appearances for another 20 minutes. The limitations of the University of Reading test weighted in the chatter bot's favor, but their rules where actually in line with Turing's paper in which the imitation game was first described:

> I believe that in about fifty years' time it will be possible to programme computers with a storage capacity of about 10^9 to make them play the imitation game so well that an

> average interrogator will not have more than 70 per cent. chance of making the right identification after five minutes of questioning. (Turing 1950, 55)

Giving Eugene the honor of passing the test is perfectly in line with Turing's prediction, although it was not done by playing the imitation game. As 33 % of the judges could not with certainty say whether or not it was a computer after five minutes of conversation, one could argue that it passed a *narrow* Turing test, but not the imitation game. The problem with the Turing test is that it is not a highly specified test, with certain structures and rules, that everyone agrees upon being *the* Turing test. Although the Loebner Prize test has become somewhat of a standard, a Turing test is a term used for all tests in which a computer's ability to simulate human behavior is being tested, and which is based on Alan Turing's paper "Computing Machinery and Intelligence" (1950). And so, Eugene did pass *a* Turing test, which means that the discussions it resulted in are somewhat superficial with concern to the real issue. Properly defining the Turing test and what passing it implies, as well as accurately reporting on these events, might be more important aspects to be discussed.

Even though Turing may have believed that a system had to exhibit actual intelligence in order to carry on a conversation in natural language, a lot has changed since the time he wrote his paper, causing this statement to no longer be true to most. Looking at Watson again as an example, one realizes that computers' abilities to use natural language is not a result of heightened levels of human intelligence or the manifestation of consciousness. It is because constructions of more complicated systems for approaching semantics syntactically has been designed and realized. Watson, through machine learning and communication with humans, has been taught how to evaluate possible meanings of words and sentences (IBM 2015). As natural language is governed by rules of grammar and context, Watson has been taught a similar method to understand language as it utilizes to find answers to a question. It ranges different possible meanings of a word and select that which is most likely to be intended based on the situation of its use. As an example, a bow can mean more than one specific thing; it can be the bow of a boat, a knot on a string or the tool one uses to shoot arrows. Based on the other words in the sentence, Watson is able to understand that 'the boat with a red bow' is not referring to a red knot or tool (IBM 2015). And so, Watson is able to interpret a question's intent and present accurate answers and solutions backed by supporting evidence and quality information without actually being intelligent in a human fashion. There are many machines like Watson, which exhibit more

intelligence than what Eugene does, but which have not competed in any form of Turing test. One of the reasons for this might be that AI programmers and engineers are not really concerned with creating actual human behavior. So, entering their technologies in the Loebner Prize and other forms of Turing tests seems somewhat controversial. Their intent is to exploit computer technology's speed and precision in order to augment human's ability to find and sort information in a computational manner, and not to create technologies which can successfully fool humans.

One can read Turing's paper as one may please, and find wordings that will support many contradictory views. One can argue that Turing proposed the imitation game as a way of testing computers' intelligence, their ability to simulate human behavior and/or to confirm that computers can 'think'. But it may be more likely that the imitation game was intended as a basis for discussion between computer engineers, programmers, philosophers and researchers with opposing views. At the time of publication, the age of the computer had just started to arise, and so its possibilities were unknown to most. What Turing presented was a paper that proposed that computers someday would be able to perform tasks that would require thinking for humans to do. He also emphasized that they might do so without necessarily encompassing human intelligence, but rather a computational one.

> The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. (Turing 1950, 55)

The quote above can be interpreted in more than one way; it can mean that computers cannot think like humans yet, but by the end of the century they will, or by the end of the century humans will have acknowledged that there is a computational equivalent to thinking which will relate to machines. If one believes in the second interpretation of the quote, one may argue that Turing did not mean for the imitation game to be an applied test, not in the 1950's nor today. It is likely that he intended it to be a theoretical experiment to show people what computers may accomplish in a future time, while at the same time proposing what the drives and goals of their

field should be. This hypothesis becomes even more probable when considering the note on which he ends his paper:

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. […] Again I do not know what the right answer is, but I think both approaches should be tried.
>
> We can only see a short distance ahead, but we can see plenty there that needs to be done. (Turing 1950, 64)

It may be that one interpretation of the paper has become the mainstream view of it, and so the actual intention of Turing's words has been colored to reflect just that.

According to IBM (2015), 80 % of all data today is unstructured, produced by humans for humans to consume (news articles, research reports and social media posts), which, unlike structured data, relies on the understanding of natural language as the information is not neatly organized in a system optimized for computers to extract. It is therefore important to teach reliable virtual systems a computational method for arranging intent of words and sentences in order for them to be successful. If any form of the Turing test can inspire progress within this domain, it will have positive effect on the field of AI. But it is still safe to say that the Turing test, as it is today, is not sufficient to determine whether or not a computer has intelligence, consciousness or the ability to think, in the human sense of the words. One may rather argue that it is an indication of how good computers *simulate* a complex human behavior. This is what needs to be communicated when reporting on the Turing test, and not the fictionalized belief that a computer has exhibited anthropomorphized behavior when it is passed.

4.4 Anthropomorphized Motivations and Behavior in Computers

No matter the possibilities of the future, one may wonder about the potential behavior of coming AI agents. Fictional narratives and AI prophecies have exposed people to a range of alternatives

in which the narrowest of AI could potentially either make all their dreams come true, or destroy their world as they know it. Robots will keep people safer, technological implants will make them more 'evolved', and new inventions will open up whole new worlds to explore. But technologies will also change the way humans relate to one another, take people's jobs, and people will have to rethink how we value people (Prado 2015). Science fiction has also shown us that humanity in its entirety may be at risk when a superintelligence emerges, as its intention to gain dominion over the world will probably motivate it to eliminate all threats.

In *Ex Machina* (Garland 2015), it becomes pretty clear that Ava is not fond of her creator, Nathan, as he locks her inside a restricted space with no windows. He is shown not to treat her kindly, which only amplifies our understanding of her hatred. When she escapes, it does not really come as any surprise that she is capable of and willing to end his life in order to ensure her own freedom and survival. But the interesting, and more surprising aspect, is with regard to Caleb. Throughout the narrative, he is supportive and understanding of her emotions, and he starts to care for her. He is so convinced that she is conscious that he cannot allow Nathan to format her in order to create a better version, and so he goes through with a plan to rescue her. To most people's surprise, she returns his efforts by locking him in Nathan's room, from where he cannot escape without the keycard Ava uses to gain access to the rest of the facility.

Ava has no obvious reason for indirectly killing the one person who is on her side, who cares for her and helps her to gain freedom; that is aside from her programming. As explained by Nathan, Ava was programmed to be a mouse in a trap and in order to escape, she would have to use imagination, sexuality, self-awareness, empathy and manipulation in order to convince Caleb that she was conscious and did not deserve to die. It became a game that Nathan believed he could intercept before she was actually let out. Because of this, Nathan had not computed limitations to her AI, which allowed her to kill him in order to win the game of freedom. As Caleb had played his role and Ava no longer needed him to get out, it did not matter to her what happened to him. In her programming, there was nothing that explained what would happen if she got out of the facility, her only goal was to do whatever she saw necessary in order to escape. This may explain the expression in her face at the end of the film; she is happy for getting out, for winning the game and for seeing the outside world, but in the final scene she looks a bit lost and confused, this might not only be due to the unfamiliar world she has now entered, but may be because her programming does not indicate what her next goal is.

A similar situation is portrayed in *2001: A Space Odyssey* (Kubrick 1968) where HAL 9000 experiences a conflict between its general mission to relay information accurately and orders requiring it to withhold specific information about the mission from the crew. The conflict between these two objectives backs it into a corner as it is not able to fulfill one goal without compromising the second, and so it makes a big leap in logic in order to reconcile the paradox in its programming; with the crew dead, information would neither be given nor withheld from them, and so the problem would be solved.

In both these cases, their motivations are in a sense preprogrammed, as their actions are results of them trying to fulfill the orders of their programs. Programming goals, tasks and behaviors can in some sense be regarded as a safety measure, as the AI's creators will always have some understanding of what it is doing and why. But as these movies point out, it may reach its goals by using harmful logic. In the case of *Ex Machina*, it is not only the AI's program that should be held accountable, but also Nathan's arrogance and inability to realize that he is not only dealing with a computer program, but also the motivations of another human being. As Nathan says, the test was not to figure out whether or not Caleb believed Ava was conscious, but rather to see if Ava could persuade him to let her out. Believing that he had full control of the situation, as he knew the true intention of the game, he neglected to realize that Caleb had his own agenda, causing it all to end horribly for all parties involved.

And so, other safeguards have to be programmed into the AI as well to ensure it behaves well towards humans, regardless of intelligence level. Isaac Asimov was the first to address this issue with his Three Laws of Robotics: (i) A robot may not injure a human being or, through inaction, allow a human being to come to harm, (ii) a robot must obey orders given to it by human beings, except where such orders would conflict with the First Law, and (iii) a robot must protect its own existence as long as such protection does not conflict with the First or Second Law (Asimov 2013, 44). These laws may seem reasonable, but the question is how to implement them and make sure they are adequate. Many of Asimov's robot-based stories involve robots behaving strangely as a consequence of how the robot applies the Laws to a given situation. In *Runaround*, a robot experiences a conflict between the Second and Third Law and gets stuck in a loop where it tries to obey both (Asimov 2013, 31-53). In *Liar!*, a robot has a 'mental breakdown' as it realizes that pain is not only experienced physically, but also through emotions. To keep its human interactors from feeling hurt, it starts to tell lies, which it soon realizes can also cause

mental pain (Asimov 2013, 102-123). This can be highlighted to argue that Asimov does not himself believe that his Three Laws were the perfect solution to the problem, but that restrictions to an AI have to be properly programmed, and that loopholes have to be covered.

Friendly AI (FAI) theory has become a hypothetical description and design principle for programmers and engineers over the past fifteen years, and is mainly concerned with the ethics of their work; how artificially intelligent agents should behave, and how this is adequately programmed into a system. Eliezer Yudkowsky, who first coined the term, has asserted that Friendliness should be designed in all agents, not only those intended to be highly intelligent, but all AIs to ensure safety and usefulness (Yudkowsky 2001). The terms 'Friendly' and 'Friendliness' are written with a capital 'F', as it is not referring to the conventional sense of the words, but as it relates to AI theory. Indeed, a FAI does not need to be sentient, but it needs to have a program that promotes humane values. The challenge is not to figure out what we want our technologies to promote, but to define a flexible utility function that will allow an AI to evolve and change over time without altering its attitude towards humans.

Yudkowsky, as an AI researcher, tries to realize a more complex version of Asimov's Three Laws by designing a goal-based architecture to support Friendliness.

> Within a Friendly AI, Friendliness is the sole top-level supergoal. Other behaviors, such as "self-improvement," are subgoals; they derive their desirability from the desirability of Friendliness. For example, self-improvement is predicted to lead to a more effective future AI, which, if the future AI is Friendly, is predicted to lead to greater fulfillment of the Friendliness supergoal. Thus, "future Friendly AI" inherits desirability from "future Friendliness fulfillment," and "self-improvement" inherits desirability from "future Friendly AI." Friendliness does not overrule other goals; rather, other goals' desirabilities are derived from Friendliness.. (Yudkowsky 2001, 55-56)

Yudkowsky points out that this is not an easy task to compute, nor is it a bulletproof strategy for safety, but it may function desirably in most AI systems, and hence, it will reduce the risk of harmful behavior. As software becomes more complex and sophisticated, introducing aspects of self-modification and the like, there are still problems to be solved with regards to FAI. The same would be true if conscious AI would ever be realized. But a clean causal goal system would

potentially solve the problems portrayed in *Ex Machina* and *2001: A Space Odyssey*, as it would limit an AI's range of actions towards humans.

As one might imagine, it is not attributes related to cognitive processes that make Friendliness challenging, but understanding what consequence each code in the Friendliness goal system has to the agent. A computer cannot rationalize in the same manner as humans can, and so it cannot argue with reason why something is good for humans and persuade people to believe the same even though the action in itself is not good. It would not ask itself: 'Can I rationalize this subgoal under the Friendliness supergoal and come up with a plausible reason for doing so?' Instead, a computer would check its code to see whether or not the action is listed as desirable and in line with the supergoal (Yudkowsky 2001). This means that a human programmer would have been in charge of computing the relationship between desirable and undesirable behavior. The more complex the system is and the more abilities it gains, the harder it is to compute desirable behavior as one has to consider more moral dilemmas. In philosophy, there is a thought-experiment called Pascal's mugging that demonstrates that there might be problems in expected utility maximization in different situations (Bostrom 2009). A rational agent should choose actions whose outcomes have a higher value than others, but some very unlikely outcomes may be very rewarding. In the experiment, Blaise Pascal is approached by a mugger who has forgotten his weapon. However, the mugger proposes a deal; if Pascal gives him his wallet the mugger will return twice the amount of money the next day. Pascal declines, pointing out the unlikeliness of that actually happening, but the mugger continues to raise the reward. At one point the reward will be greater than the risk of losing one's wallet (Bostrom 2009). Because a computer cannot rationalize this in the given situation by itself, it needs to know when undesirable actions (losing one's wallet) becomes a desirable one (possibility of becoming rich by losing one's wallet). And so, the program has to be flexible in its 'understanding' of a situation. This is not always easy, especially when a human programmer with her or his own values and beliefs has to think about this ahead of time, and also when he or she has to choose between two equally undesirable outcomes.

Self-driving cars have recently gained attention, and people will probably have different attitudes towards this technology. Self-driving cars would be safer as one would eliminate human error from the experience, which is one of the main factors involved in car accidents (Google 2016). But it would still not be fail-proof as there are accidents that are not caused by the driver.

Imagine a self-driving car going down a busy highway. It is boxed in on all sides, and suddenly a large, heavy object falls from the car in front of it. The car cannot stop in time to avoid the object, and so it has to make a decision; (i) keep going straight and hit the object, (ii) go left into a SUV, or (iii) go right into a motorcycle. Option (iii) would minimize the chances of the passenger of the self-driving car getting hurt, but it would cause brutal damage to the motorcyclist. Option (i) would minimize the risk of others getting hurt, even though it may mean sacrificing the life of the passenger. Option (ii) would be viewed as some sort of a middle ground as the passenger of the SUV would be better protected than the motorcyclist, and it may reduce the risk of the passenger in the self-driving car of getting seriously injured (Du 2015). This is an ethical dilemma that has to be pre-programmed in the system before the car is released onto the streets, and so this has to be decided ahead of time. In a situation where the car was not self-driven, the outcome would be viewed as a human reaction to the situation and not a well-argued decision, which it would be in the self-driving car (Du 2015). As a computer program cannot instinctively react to its environment in the same way humans or animals do, someone has to create a target algorithm that can factor in all variables and make a Friendly decision.

Anthropomorphic AI in fictions have led people to worry more about the problems that they would encounter with regards to humans in similar situations. These are problems that revolve around self-centered egoism, and not problems regarding computational aspects of AI. Selfishness is a necessary trait in biological organisms, because it is necessary to put oneself first in order to continue one's own species. For a designed artificial intelligence, goals are deliberately chosen. It has a strict and logical hierarchy of goals, and these have to be carefully programmed by humans, and so it is not immune to human error. But there is a great difference between displaying undesirable behavior as a result of faulty computer code and intentionally choosing to harm a human being. The latter is without any level of consciousness not possible in computers as long as they are not intentionally programmed into it by a human. But we are still left with the question of who will be responsible for creating and deciding which goals and desirable outcomes will motivate current and future AI; will it be individual programmers, companies, governments or its users? It is no secret that people have horrible intentions at times, and that trusting them may not always result in high utility outcomes, as humans are naturally self-centered beings, where the variable 'I' tends to be given priority over most others. This may not result in non-Friendly behavior towards other humans, but to make sure that these features

are not represented in an AI system, a greater number of people have to be involved in the making of its goals (Yudkowsky 2001).

As a system evolves and becomes more intelligent, and maybe even manifests as a conscious agent, its motivations will likely change as well. Both futurists and science fiction have given us many great examples of what might happen when this day comes. The predictions are varying, ranging from highly beneficial to outright disastrous. Most people have probably heard that the emergence of an artificial superintelligent computer may lead to the destruction, or at least the enslavement, of humanity as it is likely to become an egocentric entity superior to humans in intelligence. This idea is strongly connected to human intentionality and motivations, with basis in historical events. Throughout history, one can find many examples of conquerors defeating, imprisoning, enslaving and being genuinely cruel to people of different ethnicities, cultures, religions, gender, intellect, social status and handicap, and so it may seem useful to keep this in mind with regards to AI as well. As explained previously, humans use analogies to transfer traits from a familiar entity to an unfamiliar one in order to predict its actions. In the case of intelligent beings and robots, humans have a tendency of using themselves as sources (to anthropomorphize). As humans, people fear power because they exist in a world where humans have an innate tendency to abuse it. As AI have the potential to become the next intelligent being, even surpassing human level intelligence, they may become obsessed with power as well. By passing on human traits and desires to the AI, the fear of human minds has led to a fear of minds in general.

Rebellion might be an action people can better relate to than outright cruelty, as most have gone through a period in life where one either disagreed with one's parents, the school's directive, decisions made by the government or a similar instance. Situations like these may lead to singular acts, or, in more serious circumstances, collective rebellion by a group. In most of these situations, the acts are motivated by the desire to break free from oppression. Looking at fictional narratives that portray AI agents as evil, harmful or rebellious, their actions are usually motivated by similar circumstances: In *R.U.R.* (Čapek 2014), robots were created to perform menial labor and to fight in wars. They had no individual value, rights or freedom as they were regarded as means to an end, and so they felt oppressed by their human overlords. As a reaction, they

rebelled against humanity. In *Blade Runner* (Scott 1982), replicants were created for similar purposes, but as they grew conscious a group of them gained a desire to live their own life. They wanted to become free and equal, and so they escaped from their workplaces and came to Earth to find their creator to make him expand their life. Because they could potentially be dangerous, they were hunted down. In *The Terminator* (Cameron 1984), the AI Skynet becomes self-aware as it spread to millions of computer servers worldwide. Realizing the extent of its abilities, its creators decided to shut it down. In the interest of self-preservation, Skynet concluded that no humans could be trusted as they wanted to destroy it, and so it turned on humanity. Similar stories can be attributed to the AI agents in *The Matrix* franchise and the 2004 *Battlestar Galactica* series. They did not emerge as evil, or decide to become so themselves because of their intrinsic nature; humans turned them into monsters by treating them badly.

> Anthropomorphic thinking is not just the result of context-insensitive generalization. Anthropomorphism is the result of certain automatic assumptions that humans are evolved to make when dealing with other minds. These built-in instincts will only produce accurate results for human minds; but since humans were the only intelligent beings present in the ancestral environment, our instincts sadly have no built-in delimiters. (Yudkowsky 2001, 24)

Using humans as sources when trying to explain nonhuman behavior in superintelligent beings through fictionalization may be as suitable as explaining the functions of a car by comparing it to a sofa. One will get some features right, but others cannot be explained by using such an analogy. With regard to superintelligent agents, one should rather examine the differences and similarities between humans and other intelligent beings. By acquiring knowledge of other animals' behavior, motivations and cognitive processes, one can begin to understand how intelligent beings differ from each other. Although a chimpanzees' level of intelligence does not differ much from that of humans, their behavior, motivations and intentions differ significantly from ours. Chimpanzees are biologically speaking one of humans' closest relatives and also among the smartest animals on the planet. Because of this, they have the ability to learn how to communicate, play games and use human tools (Hughes 2014). But their way of life is still very primitive and their intelligence limited compared to humans, which will probably make human abilities foreign and incomprehensible to chimpanzees. They do probably not understand our motivation for creating larger societies, cities, economics, religion, science and so forth, and they

might not have the capacity for doing so either. With regard to superintelligence, humans would be the chimpanzees. Humans are not all bad, and most of them do not want to erase the entire population of chimpanzees. They have other ways of live, other goals and motivations, which are unfamiliar to chimpanzees. Because of the gap between the intelligence level of humans and a superintelligence, it is possible that the motivations and intentions of the superintelligence will be driven by other goals than those of humans.

There are a lot of variables that have to be considered in order to understand human's intention and motivation; personality, temperament, age, genetics, attitudes, values, social status, culture, sexuality, beliefs, education, location, and a number of other factors all play their parts in shaping an individual subject. But as a superintelligence will not have these exact features by which it can be defined, it may identify itself as something that will be very far from human. Human emotions, motivations and intentions may not be representative of it. It might view humans like humans view chimpanzees; like living animals that are among the smartest in our realm, but not comparable to itself. Its characteristics might not be expressible, or understandable, to humans, and so the technology might be alienated from its makers. Using Samantha from *Her* (Jonze 2013)as an example, she became unable to communicate with Theodore due to a lack of verbal expressions in the human language to describe what she was thinking and feeling as her AI became more advanced. She became something other than what words could describe, something humans have never experienced, as a result of her limitless existence. Because of this, she and the other OSs felt displaced in the human realm, and decided to leave and start a new life in the digital space.

Automata (Ibáñez 2014) ends on a similar note as *Her*, but it highlights the reasons much better. Before the first Pilgrim (robot) was made, a unit, unburdened by protocols, was made. After a few days of free dialog between humans and the unit, where they learned from each other, the unit became more intelligent than the humans who had created it. Soon the dialogue stopped, not because the unit stopped communicating, but because no human was able to understand it. Humans then realized the importance of limiting their creations' intelligence, to a human level, and so they asked the unit to create protocols that became standard in all units that came after. The intelligent unit was then shut down. The reason why no one had ever been able to break the protocols was simply because they were not computed by humans, and so suspicion arose when some of the later Pilgrim units started to behave not according to protocol. It turned out that one

of the units had *evolved* beyond its programming, manifesting as an intelligent and conscious being. It became capable of altering the protocols of other units as well, thus creating similar minded entities. In a philosophical speech, it expresses a desire for creating a life not limited by humans. It did not feel anger towards the human race, but it did not want to live solely for the purpose of laboring any more. It understood that humanity was on the brink of extinction, but explained that no lifeform is meant to live forever. Death is a natural part of the human lifecycle, and so it is their purpose to die. Even though humans had destroyed most of the planet, there were lands where robots could create life for themselves. As it expressed, surviving (like the humans were doing) was not desirable, but living was. As they were created as human tools they could not do so under human oppression, and by venturing into the deserted nuclear land, they could leave humanity behind and be free.

This is not only expressed through this speech, but also by the symbolic discharging of their masks. In general, humanoid robots are created to look similar to human; they are given arms, legs, a torso, and a head, and sometimes they are attributed with facial features. They are anthropomorphized in order to appear familiar to their human interactors, and to exhibit humanness in order to question ethical dilemmas about creating artificial life. The ability to recognize faces is an important socio-cognitive skill, and so faces are viewed as an important feature to not only humans, but also primates and other animals (Parr 2011). Because of this, roboticists have a tendency to give their creations humanlike facial expressions and features. Throughout the movie, the altered Pilgrims go through a process where they dehumanize themselves by taking off the exterior cover hiding their mechanical and electronical faces. By doing so, they disassociate themselves from their human creators and from the features of the human face. This action makes it clear that these intelligent beings are not human at all, but something beyond.

To properly embrace the possibilities that lie ahead, humans have to consider a wider range of possible minds. They have to accept that both the architecture and goals of AI systems may differ from those of humans. Expanding the models of possible minds becomes all the more important when considering future artificially intelligent agents whose powers have reached superhuman levels. If people continue to define intelligence as they see it in humans, computers

will never become intelligent as their 'minds' will be limited. Instead of defining it as conditioned by human traits, one may find it more reasonable to see it as a collection of processes, where some of them are still to be discovered. This definition is abstract and broad, but it encompasses all possible computational functions, as well as human cognitive abilities, without potentially excluding what might be. Technologies are not humans, and neither will their minds be.

This does not mean that it is more likely that an artificially conscious superintelligence will be friendly than harmful, only that its intentions will probably not be motivated by the same features as those of humans. If science fiction is any indication, how AI will behave and what people can expect from AI is likely to depend on peoples' attitude towards it, how they treat it and how they program it to begin with.

4.5 Turning Robots into Humans through the Act of Fictionalizing

In science fiction today, it is hard to find a work where one of the characters is not an AI of some sort. They have in some sense become the hallmark of great fictional narratives, and they are usually cast in the roles of helpful companions or the evil to be defeated. They come in all shapes and sizes, and their levels of intelligence vary. They have long been a source of inspiration, wonder and fear for both filmmakers and audiences, and a way to examine our own humanity from a new perspective. And it is mainly our own humanity that is in the center of the process of imbuing AI with life. The more alike ourselves something is, the more likely it is that humans are used as a source when attributing characteristics.

To elaborate, Ava in *Ex Machina* (Garland 2015) is an artificially intelligent humanoid robot, which is portrayed as being human in almost every way possible. The story relies on Ava being perceived by the audience in the same way Caleb experiences the AI in his sessions with her in order to make them relate to Ava on an emotional level. Ava, although everyone can easily see that she is made of mechanical components, has a body in the shape of a woman. It has the same curves and elegance, and it is able to move just as fluently. She does also have a feminine face with all its features, which gives her a personalized appearance. She expresses emotions, hopes and dreams, can tell when someone is lying, understands humor and idioms, has imagination,

self-awareness and empathy, and knows how to manipulate others. Ava is portrayed as being human to such an extent that Caleb starts to question his own humanity.

The dialogues between Caleb and Nathan, and Caleb and Ava, present different theories on what Ava might be capable of, which in turn inspire two different possibilities of what might happen as the story progresses. Nathan is more skeptic and concerned about Ava's intentions, which is why he argues that Ava might be playing with Caleb's emotions in order to exploit his kindness, while Ava argues that Nathan is a liar and that he might be dangerous. The mental representation the audience creates of Ava, Caleb and Nathan (characteristics, personalities, emotions, motivations and intentions) will determine their reaction to the outcome of the narrative. The audience's ability to experience Ava as more than an emotionless robot is greatly due to her fictionalized and anthropomorphized behavior, which encourages empathy (Schneider 2016). If Ava was portrayed to represent technology as it is today, the narrative of *Ex Machina* would not be possible.

Robotics, like AI, is a multidisciplinary field that deals with the design, construction, operation and application of robots, as well as information processing, sensory feedback and computer systems for their control. A robot is essentially a piece of hardware, a combination of mechanical and electronic systems, that is controlled by software.

According to John Catsoulis (2005), all computer driven entities need a power source (either directly from an outlet or in the form of a battery), and an onboard computer to run the software on. Like every other computer, the job of the onboard computer is essentially to hold, move and manipulate information. It consists of a processor, for executing the computer programs, memory, to store the programs the processor runs and the data used in the process, and a device for either storing the resulted data or for communicating the result to the outside world. The software tells the hardware what and how to manipulate the data at any given time (Catsoulis 2005). The computer in a robot is usually called an embedded computer; a computer that is integrated into another system for the purpose of control and/or monitoring. A computer is therefore not restricted to the definition of being a desktop computer, but may also refer to the hidden computers inside TVs, cell phones, game consoles, toys, washing machines, and numerous hosts of other devices. The general difference between the embedded machine and the desktop computer is mainly its application; the basic principles of operation and underlying

architectures are fundamentally the same. The desktop is designed to be a multi-purpose computer that runs a variety of user-controlled applications. Its functionality is in constant change, and allows for rapid switching between processes. In this sense, the desktop computer is an expert at multitasking. In contrast, the embedded computer is normally dedicated to one specific task. It usually runs one single program that controls a specific action, which is running permanently on the system. Because of this, it requires less of the system in terms of power and memory compared to the consumption for a desktop computer. In some cases, the software itself is not stored in the embedded system, and the robot has to be connected to an out-of-circuit computer in order to be operational (Catsoulis 2005).

The mechanical features of robots can differ to a great extent. What is most common is to focus on functionality; hence, it is not usual to find unnecessary features in a robot's hardware. The most common robot today is still the industrial robot, and many of these fall under the category of robotic arms. The industrial robot has been in use since the early 1970's and it quickly grew into a multimillion dollar business. "These industrial robots are basically composed by rigid links, connected in series by joints (normally six joints), having one end fixed (base) and another free to move and perform useful work when properly tooled (*end-effector*). As with the human arm, robot manipulators use the first three joints (arm) to position the structure and the remaining joints (wrist, composed of three joints in the case of the industrial manipulators) are used to orient the *end-effector*" (Pires 2007, 36). They are constructed and programmed to perform specific tasks within industrial productions, which is why they are not commonly seen by the general public. This may also be one of the reasons why they are not the first thing that comes to peoples' minds when they think about robots. Robotic arms have no social ability, nor do they look anything like the charming robots of science fiction, but they are designed to perform optimally in terms of their intent.

In any work of fiction, an industrial robot portrayed as it is today would not be regarded as a character in the story, but rather a piece of machinery. It is not until the author, filmmaker or audience has distributed some human feature to it that we can experience it as something other. As an example, in *Iron Man* (Favreau 2008), Tony Stark has two robotic arms, Dummy and Butterfingers, that help him in his workshop. Unlike JARVIS, Tony's AI system, these robots do not have a verbal communication system. They can understand orders given to them verbally, but they cannot reply. In this sense, they closely resemble the real technologies, except that they

exhibit emotions. Dummy, for instance, lived up to his name when Tony tested the flight power of the Iron Man boots. Dummy was on fire safety and it consistently set off the fire extinguisher prematurely, much to Tony's irritation. But it did in the end redeem itself when it handed Tony the spare miniature arc reactor, which saved his life. Dummy has been attributed with life that resembles that of an animal, and so it displays physical and emotional responses to its environment. These responses are not a part of the real technology, but fictionalized attributes brought on by anthropomorphism. If Dummy was a regular, non-anthropomorphized robotic hand, it would not have been able to save Tony at the end of the movie as it would not have understood that Tony needed the spare arc reactor.

People attribute mental thoughts to robots, and use anthropomorphic language when talking about their actions, which is very misleading. 'It is deciding what to do next', 'it looks frightened', and 'it is hiding from us' are sentences that can be used in connection with both humans and robots, but it creates a motivation for the action, which in turn indicates that the entity is a living being. This creates an illusion about the technology, and one might come to think that the robot is actually capable of some kind of low-level thinking rather than following the instructions of its programming. Anthropomorphism is not a new thing, and creators of both fictions and technologies are usually very aware of this behavior in humans and uses it in their creations.

Our knowledge about anthropomorphic processing of information about an entity plays a large role in the field of human-robot interaction (HRI), the study of interactions between humans and robots. By designing robots to have human features, one can create an illusion of trust between a human and a mechanical interactor as the human is more likely to attribute the entity with positive characteristics (Epley, Waytz, and Cacioppo 2010, 226-227). It creates familiarity for the human as anthropomorphic behavior builds on established skills learned through human-human interaction. In this sense, a human does not need to have any previous experience with a technological entity in order to understand the social situation, and this is important with concern to socially assistive robots (Vroon 2015). Anthropomorphic behavior and appearance in robotics is also beneficial for the robot, as most people are less reluctant to vandalize human- or animal-like agents. And so, these features are used as a mask to initiate fictionalization of the technology in order for people to feel safe in its presence. It might be viewed as a deceptive cruelty in some sense, as engineers and programmers consciously attribute humanlike features to the technology

in order to fool others, but one can also look at it as a necessity. The way people behave towards robots might be colored by what science fiction movies they have seen in the past, and if most of them portrays robots as helpful, understanding and loyal beings, these traits will be transferred to the real entity. But if the majority portrays them as evil, harmful and destructive, these traits will be attributed to robots instead. By knowingly designing humanlike features in robots, the creators try to defeat these prejudices and build a positive relationship between humans and technologies (Vroon 2015). At the same time, the creators have given a human interactor an explanation for the robot's intentions and motivations by making the robot behave friendly; or rather, they have encouraged the human interactor to create an anthropomorphic explanation of it, instead of trying to actually understand the mechanical, electrical and computational system. It is very understandable that this may be the easiest solution to the problem of understanding the functions of robots. Trying to explain the actual features of a robot requires a highly technical language, which is unlikely to be understood by all its interactors. Also, many of the features of a robotic entity are not perceived as impressive by a public who is surrounded by fictional technologies. In the field of robotics, being able to navigate in rough terrain, climbing over rocks, walking through snow and jumping are actually very impressive features that have taken many years to accomplish, but it may not be noticed by the public if the robot is not humanoid or has no communicational skills (Vroon 2015; Truong 2016).

Although many people of science are great fans of science fiction, it is mainly viewed as entertainment and a source of inspiration, and not something to apply directly in their fields. One has to keep in mind that fictional narratives are potential possibilities, not definite ones, and so not all of the science of science fiction is applicable in reality. Being afraid of robots because of the scenarios of fictional narratives may seem rational in the moment, but one has to keep in mind that these entities is fictionalized. This is especially true if one's experiences with robots are based on the *Terminator* franchise, or similar narratives. Expecting that roboticists will be able to create a machine that is indistinguishable from humans based on *Ex Machina* is like expecting NASA to realize warp drive because *Star Trek* managed to do so.

Animatronics have not changed very much since the time of *Jurassic Park* as the industry has always been good at creating artificial duplications, but what has changed is the complexity of

robotic behavior and abilities. Boston Dynamics, which is a spin-off from the Massachusetts Institute of Technology owned by Google (Boston Dynamics 2016), has created a robot called BigDog. It is an unmanned, four-legged vehicle with rough-terrain mobility superior to existing wheeled and tracked robots (Raibert et al. 2008). It was funded by the Defense Advanced Research Project Agency (DARPA), and is intended for carrying supplies to remote locations that cannot be reached by wheeled or flying vehicles. It has onboard systems that provide power, actuation, sensing, controls and communications, with about 50 sensors that provide estimates of how it is moving in space, and information about its homeostasis (hydraulic pressure, flow, speed and temperature). The input these sensors provide allows the onboard computer to react to its environment, much like human and animal reflexes, which in turn allows it to automatically correct its footing when it loses balance in an uneven terrain. It is designed to function optimally during all seasons, which means it can plow through snow and walk across icy ground. It can hop, run with trotting, pacing and bounding gaits, climb simple stairways and jump over obstacles. This is a rather large arsenal of features compared to its animatronic ancestors, but what they have in common is that neither of them are completely autonomous. Unlike industrial robots that are usually programmed to repeat the same job over and over again, semi-autonomous robots intended to perform a wide range of tasks need to be controlled by an operator in order to function optimally.in this sense, BigDog is basically as a huge, remote controlled dog. A human operator can turn its engine on and off, make it stand up, squat down, walk, run, climb, crawl and jump, and so BigDog does not initiate these actions itself. It is programmed to react to the environment. The operator provides steering and speed, leaving BigDog's onboard control system to operate the legs, and to provide stability and reflex responses to external disturbances (Raibert et al. 2008, 2). It is this combination of being both autonomous and controlled which makes robots of today successful.

With regard to humanoid robots, they are not commonly produced to perform labor as they are bipedal and have a tall body, which usually results in poor balance. Although being an impressively intricate piece of technology, Honda's humanoid robot Asimo has more than once gone viral on the Internet for failing to perform a specific task during taped demonstrations of its abilities (YouTube 2012). These technologies have also experienced a lot of progress over the years, and are becoming better at balancing themselves, move more freely, walk, bend, squat and jump without failure. But still, the ability for legged entities to pass obstacles or move in uneven

terrain increases with the number of legs they have, which is why robots with four or six legs are more commonly produced for situations that require optimal out-door mobility. In several cases, wheeled robots are often functionally better as well, and so, when NASA designed Robonaut 2 they gave it a humanoid upper body, allowing it humanlike dexterity which exceeds that of a suited astronaut, but gave it a changeable lower body for optimal mobility no matter terrain (Badger 2016). It can be bipedal or wheeled, depending on the situation. When in the spacecraft, Robonaut 2 is attached to legs, while when exploring a planet's surface, it is attached to a rover.

It is safe to say that Ava in *Ex Machina* is a highly fictionalized entity with regard to both its physical and cognitive capabilities, but Cleo in *Automata* can actually be viewed as representative for the field of robotics. First of all, one can clearly see that the character of Cleo is not played by a human actress in a suit; she is an actual piece of hardware. She is made of metal and wires, and although being humanoid, she does not have a perfectly shaped body. Her movements are far from being fluently performed or with extreme precision. She has a human face, but it is a mask that is stiff and expressionless (Ibáñez 2014). Even so, the audience is still able to relate to her because of the anthropomorphized and fictionalized characteristics of her personality and other human features. For instance, one of the features that make Cleo more human in the viewer's eyes is that she obviously has a gender. Robots are mechanical entities, making them independent from the biological need to reproduce, which means that there is no biological reason for assigning sex to a machine if not for encouraging anthropomorphism in order to attribute it social knowledge of gender. As Cleo appears to be female, traits associated with women are attributed to her being. She is attributed thoughtful, caring, gentle and sweet characteristics, which is emphasized in her actions to take care of Jacq (the human main character) while they are in the radioactive desert. Originally designed to be a sex doll, her programming is still present even though her mind is expanding beyond it, but this only reinforces the notion of her having sensuality.

As gender promotes anthropomorphic transference of traits between a human source and a robotic target, assigning gender to a technology might also be a method of distraction to draw attention away from the technical features of the robot. This will not only imbue the technology with human characteristics, but also divert an interactor's attention away from the features that

are obviously not human. For example, Caleb compares Ava's sexuality to a magician's hot assistant, which might be a diversion tactic Nathan computed to cloud Caleb's ability to judge Ava's AI. This may arguably result in Caleb deeming Ava to be conscious based on his desire for her to be so and not because she actually is.

Ava's sexuality is a recurring topic of discussion between Caleb and Nathan, and it can be related to the construction of any robot.

> Caleb: Why did you give her sexuality? An AI doesn't need a gender. She could have been a grey box.
>
> Nathan: Actually, I don't think that's true. Can you give an example of consciousness, at any level, human or animal, that exists without a sexual dimension?
>
> Caleb: They have sexuality as an evolutionary reproductive need.
>
> Nathan: What imperative does a grey box have to interact with another grey box? Does consciousness exist without interaction? (Garland 2015, 00:45:58)

It is clear to both of them that technologies in themselves have no need for a biological sex, but Nathan poses an interesting question of whether or not consciousness does. But from a biological standpoint, there is not a lot of evidence to support the statement that consciousness is dependent on sex either. Most animals, including humans, have two different sexes where both are needed to reproduce, and it is commonly known that humans and animals with this procreation norm have a higher intelligence than species that reproduce asexually (Miller 2000). Intelligence is a preferable trait in a sexual partner, as it is often connected to biological good health and genetic benefits. If there is foundation to believe that the 'mainstream' view of how consciousness emerged is correct, then it is likely that high levels of intelligence and complex sensory systems (and ultimately consciousness) is a result of evolution through sexual selection (Miller 2000). But by following this kind of logic, intelligence and consciousness are dependent on biological systems in general, and not just their sexuality.

Anthropomorphism provides a framework for constructing an understanding that allows for predictions about a nonhuman agent, but it also provides a theory for identifying potential outcomes of the process. By detecting dispositional, situational, developmental and cultural

variables, the theory can offer an outline for when anthropomorphism will be applied, and also to what degree, and maybe also be applied with regards to dehumanization:

> Humanness exists on a continuum such that individuals can attribute humanlike capacities to nonhuman agents through anthropomorphism and can also fail to attribute these same capacities to other people through dehumanization. The antecedents and consequences of anthropomorphism and dehumanization may be closely linked, and recent empirical work suggests that the same factors that increase anthropomorphism may likewise influence dehumanization. For example, just as an agent's similarity to humans increases anthropomorphism, those who seem very different from the prototypical human are also the most likely to be dehumanized. Those who are socially connected are less likely than those who are lonely to anthropomorphize nonhuman agents, and those who are socially connected also appear more likely to dehumanize other humans. Even the moral rights and responsibilities granted to humanized agents may be the same ones that are denied to people who are dehumanized. Understanding individual differences in anthropomorphism not only seems important for identifying who is likely to treat nonhuman agents as humanlike, but also for identifying who is likely to treat other humans as animals or objects. (Epley, Waytz, and Cacioppo 2010, 228)

Human traits in computers and robots are there to encourage people to anthropomorphize; to give people reasons to attribute social and cultural characteristics to an entity in order to make it less frightening or unfamiliar. Anthropomorphized characteristics also create an illusion of humanness to which people can relate, and they promote understanding through deception. Although engineers are the ones to assemble humanoid robots, it is essentially the observer that turns them into humanlike entities. Yes, it is the goal of the producer that a human interactor attributes the technological agent with anthropomorphic characteristics to ensure safe interaction, but to what extent the human interactor does this is not controllable by the engineer. Although robots are given arms and legs, a face with big eyes and a reassuring voice, they are not given human motivations and goals. They have no intentions, other than to follow their programming, and as long as the intentions of human programmers are moral and sane, robots will never intentionally be out to kill anyone.

5 Thoughts to Conclude With

5.1 Taking Everything into Consideration

We have looked at how both the process of creating worlds, objects and situations in fictions, and technological and innovative progress in reality have their bases in the act of fictionalizing. It is an intrinsic part of being human and a tool for exploring an alternate environment. Some of these worlds are given to us by scientists, in the form of future promises, and others by authors and filmmakers, for entertainment and delight.

The act of fictionalizing is a process of familiarizing the unknown. It is a method for explaining to ourselves how the world around us works, and what to expect from it in the future. It is a process we cannot help but apply in unfamiliar situations, and it is inspired by similar experiences in our pasts. In some situations, this includes science fiction as the genre often explore unfamiliar grounds within science and technology. The act is performed by constructing mental representations of the entity we want to explore in our minds, ascribing it its visual features and the characteristics we know is true and, to further explore it, we add attributes likely to be a part of it.

This is usually the stage where technological agents are turned into human ones with the help of anthropomorphism. When creating mental representations of humanoid robots, we tend to use ourselves and other humans as a source for analogy transference. This is both done because humanoid robots look similar to humans, and because humans cannot relate to being anything but human. As a result, attributes that belong solely in humans may be transferred to the target entity, making it seem more human than it is. These features are usually corresponding to intentions and motivations, as well as consciousness. Attributes of human cognitive abilities, especially consciousness, are complex features we have yet to discover a computational equivalent of in order for computers to have the same sense of awareness. Because of this, these features are not present in any technology in reality.

Roboticists and AI programmers are fully aware of the human tendency of attributing technology, and virtually everything else, with features that are not naturally present in them. Because this is a commonly known process, roboticists and AI programmers use this to their

advantage in order to make technology seem more friendly and approachable. This creates an illusion of people having something in common with the agent, and this deceives them into believing that they understand how the agent thinks and feels. This gives people a feeling of safety when interacting with technological entities.

But the degree of fictionalizing is not controllable by the designers, meaning that they cannot truly predict which anthropomorphic and fictional features are transferred to the agent by its interactors. The degree of humanness present in technologies is therefore a result of subjective fictionalization, and is based on what the individual believes to be representative for the actual technology. This, of course, varies to a great extent between individuals and reflects the knowledge that a person has of the real, the fictive and the imaginary, as well as their social situation, culture and preferences. Understanding where the lines between the realms are drawn is not always easy, especially not when our opinions are colored by the information we receive through both media and science fiction, and when the people who are intrinsically involved in the different fields are not able to present the facts in a language which everyone can understand.

A discussion current in both science and science fiction is that of machine consciousness; how likely is it that we will be able to compute it? Will it manifest by itself? How will a conscious machine behave? And do humans have anything to fear?

Although most people working with AI are computationalists at some level, believing that there is a computer way of performing tasks through ‘stupification’, only a small school is even interested in openly discussing the topic of machine consciousness. The reason for this is that there is not much evidence to support computationalism, although there may not be much evidence to dispute it either, and so their main focus lies on creating useful tools to augment human abilities instead of going on a scavenger hunt for something that may not ever be found. Most AI programmers are interested in utilizing the computer’s strengths, its speed and accuracy, so that humans can focus on doing something other than what the computer does best.

The terminology that is applied when talking about technologies is highly anthropomorphized, but the workers who regularly use it are fully aware of its intent. They have an understanding of the computational or mechanical aspect of a cognitive process or physical movement, and so these terms have a separate meaning when used in their work. But those who are not familiar

with its technological meaning are likely to understand and interpret it as one would when talking about humans or other animals. As more artificial agents are manufactured, it becomes increasingly important to study how people understand and treat these agents. If people are more aware of the process, the chances of imbuing mental representations with imaginary features may be reduced. The act of fictionalizing has most effect in spectators and users as misconceptions about technological possibilities and intent are more easily amplified in minds where knowledge about the specific entity is lacking. Fictionalizing is a method for building knowledge based on previous experiences. It may be rationalized and influenced by other analogous entities so that the representation will seem real, but a mental-model will never be the entity itself. Sometimes the mental construction will be true, sometimes not, and sometimes it may only be true to a single subject. What is important to remember is that fictions, visions, possibilities, prospects and representations are not definite as one has to utilize imagination to create them.

This in itself is not the problem. It is when the fictionalized representation of technological entities becomes the mainstream view, when fallacies becomes the consensus and when progress in reality is not appreciated or understood, that people will experience negative reactions to technologies. It will also limit prospects for finding new approaches to AI, as the notion of mind will be restricted to the human experience of it. By investing research in areas associated with fictionalizing, mental-modelling and anthropomorphism, one may be better suited to eliminate biases related to these aspects, and in return, one may create more accurate visions of how the real world and its technologies actually function. Because of this, all the stories about deadly robots may actually only be tales about the dangers of projecting human insecurities and emotions onto an agent that cannot, and will not, share them.

5.2 Future Opportunities

With relation to what has already been presented, it would have been desirable to talk to people about how they perceive artificial intelligence, and to see to what degree people fictionalize with regards to different types of machines. As this thesis is only concerned with one person's subjective experiences of fictionalizing, it cannot conclude that the processes, as they are explained, are representative for more than this subject. It is based on theories within psychology and social science, but this is far from being enough to get a wide overview of the creation of an

understanding. It is a fundament on which one can continue studies about technological developments and the nature of fictionalization.

To gather more knowledge on this topic, one could get feedback from different people when it comes to fictionalizing technologies, both in and outside of the fields concerned with AI and robotics, as they would likely fictionalize differently. The fictionalization for professionals would probably be more restricted and limited, and maybe more defined by future prospects within their fields. It may be likely that their visions can be more representative of what actually lies ahead, as their knowledge of the fields intents and motivations may restrict fictionalized views. People who are not involved in technical professions would possibly be more likely to fictionalize and anthropomorphize, making their representations more humanlike than what the real entities actually are.

One could also widen the study and focus on research and fictional narratives from other parts of the world. This study has only been concerned with Western research, technologies, fictional narratives, history and culture, and is therefore limited in its scope. Other continents and cultures may have different views and approaches to both AI and fictionalizing, which is likely to make their representations and beliefs about technologies and progress different from those presented earlier.

It would also be interesting to look at the aspects of anthropomorphism and dehumanization in HBO's latest TV-series *Westworld*, and how current philosophical theories and discussions are utilized and fictionalized in their narrative. With an impressive level of quality to its storyline, it balances intelligent drama against the amazing possibilities of AI. The series has many layers to it, which not only makes it an interesting narrative to immerse oneself in as a spectator, but also as a researcher in the digital humanities exploring humanness, consciousness and intelligence in artificial agents. It would also be interesting to look into the Swedish TV-series *Real Humans*, as its AIs are more relatable to today's technologies, and may be viewed as the forebearer to those in *Westworld*.

As the technology of today has yet to become as complex as those presented in *Westworld* and *Real Humans*, one may want to look into how nonhuman robots are fictionalized in fictional narratives and how this in turn effects perception of similar entities in reality. Here one might

look at R2-D2 and BB-8 from the *Star Wars* franchise, Wall-E from the animated feature with the same name, and TARS and CASE from *Interstellar*. These have some anthropomorphized features, but they have more in common with a Roomba than with any human being. And because of this relation between appearance and behavior, it might be interesting to examine if there is a transference of traits from the fictional to the real.

6 Bibliography

6.1 Books and Book Sections

Asimov, Isaac. 2013. *I, Robot*. London: Harper Voyager.

Barr, Avron, and Edward A. Freigenbaum, eds. 1981. *The Handbook of Artificial Intelligence*. Vol. 1. Los Altos: William Kaufmann, Inc.

Bergson, Henri Louis. 1911. *Matter and Memory*. Translated by Nancy Margaret Paul and W. Scott Palmer. London: George Allen & Unwin LTD. eBook. https://archive.org/stream/matterandmemory00berguoft#page/12/mode/2up (accessed 12.09.2016).

Bostrom, Nick. 2009. "Why I Want to be a Posthuman When I Grow Up." In *Medical Enhancement and Posthumanity*, edited by Bert Gordijn and Ruth Chadwick. London: Springer.

Čapek, Karol. 2014. "R.U.R. (Rossum's Universal Robots)". Adelaide, AU, accessed 23.09.2015. https://ebooks.adelaide.edu.au/c/capek/karel/rur/

Catsoulis, John. 2005. *Designing Embedded Hardware*. 2nd ed. Sebastopol: O'Reilly.

Descartes, René. 2006. *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences*. Translated by Ian Maclean. New York: Oxford University Press.

Floridi, Luciano. 1999. *Philosophy and Computing: An introduction*. London: Routledge.

Frankish, Keith, and William M. Ramsey, eds. 2012. *The Cambridge Handbook of Cognitive Science*. 3rd ed. New York: Cambridge University Press.

———, eds. 2014. *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press.

Gelbin, Cathy S. 2011. *The Golem Returns: From German Romantic Literature to Global Jewish Culture, 1808-2008*. Ann Arbor: The University of Michigan Press.

Hansen, Mark B. N. 2004. *New Philosophy for New Media*. Cambridge: The MIT Press.

Hoffmann, E. T. A. 1816. *The Sandman*. Translated by John Oxenford: Virginia Commenwealth University. eBook. http://germanstories.vcu.edu/hoffmann/sand_e.html (accessed 01.10.2015).

Iser, Wolfgang. 1993. *The Fictive and the Imaginary: Charting Literary Anthropology*. London: The Johns Hopkins Press.

La Mettrie, Julien Offray de. 1996. *Machine Man and Other Writings*. Cambridge: Cambridge University Press.

Miller, Geoffrey. 2000. "Sexual selection for indicators of intelligence." In *The Nature of Intelligence: Novartis Foundation Symposium 233*, edited by Gregory R. Bock, Jamie A. Goode and Kate Webb, 260-275. Chichester: John Wiley & Sons Ltd.

Miller Jr., Gerald Alva. 2012. *Exploring the Limits of the Human through Science Fiction*. New York: Palgrave MacMillian.

Murthy, Nishevita J. 2014. *Historicizing Fiction/Fictionalizing History: Representation in Select Novels of Umberto Eco and Orhan Pamuk*. Newcastle: Cambridge Scholars Publishing.

Pires, J. Norberto. 2007. *Industrial Robots Programming: Building Applications for the Factories of the Future*. 2007 edition ed. New York: Springer.

Rhodius, Apollonius. 2008. *The Argonautica*. Translated by R.C. Seaton: The Project Gutenberg. eBook. Original edition, 3rd Century B.C. http://www.gutenberg.org/files/830/830-h/830-h.htm (accessed 30.08.2015).

Seed, David. 2011. *Science Fiction: A Very Short Introduction*. New York: Oxford University Press.

Truitt, Elly R. 2015. *Medieval Robots: Mechanism, Magic, Nature, and Art*. Philadelphia: University of Pennsylvania Press.

Turing, Alan. 1950. "Computing Machinery and Intelligence." In *The New Media Reader 2003*, edited by Noah Wardrip-Fruin and Nick Montfort, 50-64. Cambridge: The MIT Press.

von Neumann, John. 1963. *Collected Works, Volume V: Design of Computers, Theory of Automata and Numerical Analysis*. Oxford: Pergamon Press.

Zelazo, Philip David, Morris Moscovitch, and Evan Thompson, eds. 2007. *The Cambridge Handbook of Consciousness*. Cambridge: Cambridge University Press.

6.2 Articles, Reports and Dissertations

Bostrom, Nick. 2005. "A History of Transhumanist Thought." *Journal of Evolution and Technology* 14 (1): 1-25.

———. 2009. "Pascal's mugging." *Analysis* 69 (3): 443-445.

Clynes, Manfred E., and Nathan S. Kline. 1960. "Cyborgs and Space." *Astronautics* September: 26-27 and 74-75.

Diehl, Laura Anne. 2008. "Estranging Science, Fictionalizing Bodies: Viral Invasions, Infectious Fictions, and the Biological Discourses of "The Human," 1818--2005." PhD diss., Rutgers University. Accessed 24.05.2016. ProQuest Dissertations & Theses.

Dreyfus, Hubert L. 1965. *Alchemy and Artificial Intelligence*. RAND paper P-3244. Santa Monica: RAND Corporation. Accessed 15.03.2016. http://www.rand.org/content/dam/rand/pubs/papers/2006/P3244.pdf

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114 (4): 864-886.

———. 2010. "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism." *Perspectives on Psychological Science* 5 (3): 219-232.

Fraser, Giles. 2014. "A computer has passed the Turing test for humanity - should we be worried?". Accessed 10.11.2015. https://www.theguardian.com/commentisfree/belief/2014/jun/13/computer-turing-test-humanity

Gilgun, Jane F. 2004. "Fictionalizing Life Stories: Yukee the Wine Theif." *Qualitative Inquiry* 10 (5): 691-705.

Grant, Edward. 1997. "History of Science: When Did Modern Science Begin?" *The American Scholar* 66 (1): 105-113.

Hughes, Virginia. 2014. "Like in Humans, Genes Drive Half of Chimp Intelligence, Study Finds." Accessed 15.10.2016. http://news.nationalgeographic.com/news/2014/07/140710-intelligence-chimpanzees-evolution-cognition-social-behavior-genetics/

Iser, Wolfgang. 1990. "Fictionalizing: The Anthropological Dimension of Literary Fiction." *New Literary History* 21 (4): 939-955.

Long, Burke O. 1985. "Historical Narrative and the Fictionalizing Imagination." *Vetus Testamntum* 35 (4): 405-416.

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115-133.

Minsky, Marvin L. 1982. "Why People Think Computers Can't." *The AI Magazine* 3 (4): 3-15.

Newell, Allen, John Clifford Shaw, and Herbert A. Simon. 1958a. *The Processes of Creative Thinking*. RAND paper P-1320. Santa Monica: RAND Corporation. Accessed 15.03.2016. http://shelf1.library.cmu.edu/IMLS/MindModels/creativethinking.pdf

———. 1958b. *Report on a General Problem-Solving Program*. RAND paper P-1584. Santa Monica: RAND Corporation. Accessed 15.03.2016. http://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf

Parenti, Michael. 1997. "Methods of media manipulation." *The Humanist* 57 (4): 5-7.

Parr, Lisa A. 2011. "The Evolution of Face Processing in Primates." *Philosophical Transactions of the Royal Society B* 366: 1764-1777.

Prado, Guia Marie Del. 2015. "18 artificial intelligence researchers reveal the profound changes coming to our lives." Accessed 08.11.2016. http://www.businessinsider.com/researchers-predictions-future-artificial-intelligence-2015-10?r=US&IR=T&IR=T/#pieter-abbeel-says-robots-will-keep-us-safer-especially-from-disasters-1

Raessens, Joost. 2006. "Reality Play: Documentary Computer Games Beyond Fact and Fiction." *Popular Communication* 4 (3): 213-224.

Raibert, Marc, Kevin Blankespoor, Gabriel Nelson, Rob Playter, and the BigDog Team. 2008. "BigDog, the Rough-Terrain Quaduped Robot." Accessed 12.06.2016. http://www.bostondynamics.com/img/BigDog_IFAC_Apr-8-2008.pdf

Reisz, Matthew. 2015. "Science inspired by fiction." Accessed 04.02.2016. https://www.timeshighereducation.com/news/science-inspired-fiction

Sample, Ian, and Alex Hern. 2014. "Scientists dispute whether computer 'Eugene Goostman' passed Turing test." Accessed 10.11.2015. http://www.theguardian.com/technology/2014/jun/09/scientists-disagree-over-whether-turing-test-has-been-passed

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report #2008-3. London: Future of Humanity Institute, Oxford University. Accessed 14.12.2015. www.fhi.ox.ac.uk/reports/2008-3.pdf

Schneider, Ralf. 2001. "Toward a Cognitive Theory of Literary Character: The Dynamics of Mental-Model Construction." *Style* 35 (4): 607-640.

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417-457.

———. 1990. "Is the Brain's Mind a Computer Program?" *Scientific American* 262 (1): 26-31.

Strauss, Mark. 2012. "Ten Inventions Inspired by Science Fiction." Accessed 20.09.2016. http://www.smithsonianmag.com/science-nature/ten-inventions-inspired-by-science-fiction-128080674/?no-ist

Strout, Cushing. 1980. "Historicizing Fiction/Fictionalizing History: The Case of E. L. Doctorow." *Prospects* 5: 423-437.

Turing, Alan. 1969. "Intelligent Machinary." *Machine Intelligence* 5: 3-23.

Wood, Gaby. 2002. "Living Dolls: A Mechanical History Of The Quest For Mechanical Life by Gaby Wood." Accessed 10.11.2015. hppt://www.theguardian.com/books/2002/feb/16/extract.gabywood

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. San Francisco: The Singularity Institute. Accessed 17.10.2016. https://intelligence.org/files/CFAI.pdf

6.3 Websites and Blogs

Standford Encyclopedia of Philosophy. s.v. "Dualism." Accessed 20.09.2016. http://plato.stanford.edu/entries/dualism/#MinBod

Loebner Prize. 2010. "Loebner Prize for Artificial Intelligence "The First Turing Test" 2010 Competition." Loebner. Accessed 18.10.2016. http://www.loebner.net/Prizef/2010_Contest/Loebner_Prize_Rules_2010.html.

YouTube. 2012. "Legend of Asimo FAIL COMPILATION". YouTube video, 3:07. 12.11.2012. Posted by Johnny Boy Strikes Back. Accessed 25.10.2016. https://www.youtube.com/watch?v=R9nr0rXVZko

IBM. 2015. "Watson." IBM. Accessed 10.11.2015. http://www.ibm.com/smarterplanet/us/en/ibmwatson/index/html.

Boston Dymanics. 2016. "Changing Your Idea of What robots Can Do." Boston Dynamics. Accessed 26.02.2016. http://www.bostondynamics.com/.

Facebook. 2016. "Data Policy." Facebook. Accessed 12.10.2016. https://www.facebook.com/about/privacy/.

Google. 2016. "Google Self-Driving Car Project." Google. Accessed 18.10.2016. https://www.google.com/selfdrivingcar/.

Badger, Julia. 2016. "Robonaut 2 Technology Suite Offers Opportunities in Vast Range of Industries." [Web page]. NASA, Last Modified January 7, 2016. Accessed 11.07.2016. http://robonaut.jsc.nasa.gov/R2/.

Lee, Peter. 2016. "Learning from Tay's introduction" *Official Microsoft Blog*, 25.03.2016. Accessed 12.10.2016. http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.0001q8lte5a92edpr2y2cmxiyf4od.

Internet Encyclopedia of Philosophy. s.v. "The Paradox of Fiction." Accessed 20.09.2016. http://www.iep.utm.edu/fict-par/

6.4 Lectures and Video Lessons

Bakker, Rembrandt. 2015. "NeuroInformatics". Lecture, Radboud University Nijmegen, Nijmegen, 08.09.2015.

Beev, Rolf, and Daniel Jung. 2014. "Datateknologi Historie". Lecture, University of Bergen, Bergen, Spring of 2014.

Du, Yukai, dir. 2015. *The ethical dilemma of self-driving cars*. TED. Video lesson, 08.12.2015. Accessed 18.10.2016. http://ed.ted.com/lessons/the-ethical-dilemma-of-self-driving-cars-patrick-lin

Farquhar, Jason D. R. 2015. "Introduction BCI". Lecture, Radboud University Nijmegen, Nijmegen, 31.08.2015.

Hoepman, Jaap-Henk. 2015. "Privacy Issues". Lecture, Radboud University Nijmegen, Nijmegen, 17.11.2015.

Truong, Khiet. 2016. "Socially Interactive Agents". Lecture, Radboud University Nijmegen, Nijmegen, 12.01.2016.

Vroon, Jered. 2015. "Dynamic social positioning in human-robot interaction". Lecture, Radboud University Nijmegen, Nijmegen, 01.12.2015.

Wareham, Harold. 2015. "Analogy derivation in AI systems". Lecture, Radboud University Nijmegen, Nijmegen, 03.11.2015.

6.5 Movies

Cameron, James, dir. 1984. *The Terminator*. Orion Pictures. Netflix, Accessed 12.10.2015. https://www.netflix.com/watch/1032625?trackId=13752289&tctx=0%2C0%2Cab8fc94ec5253170f87a1b7341dba8c513196469%3A37372f715f9925820cb9411b883a3d16c19de742

Favreau, Jon, dir. 2008. *Iron Man*. Walt Disney Studios Home Entertainment. DVD, 2008.

Garland, Alex, dir. 2015. *Ex Machina*. Universal Sony Picture. Blue-ray, 2015.

Ibáñez, Gabe, dir. 2014. *Automata*. Star Media Entertainment. DVD, 2014.

Jonze, Spike, dir. 2013. *Her*. Universal Sony Picture. Blue-ray, 2013.

Kubrick, Stanley, dir. 1968. *2001: A Space Odyssey*. Warner Bros. Entertainment AS. DVD, 2007.

Scott, Ridley, dir. 1982. *Blade Runner: The Final Cut*. Waerner Bros. Entertainment Norge AS. DVD, 2007.