UNIVERSITY OF COPENHAGEN

MASTER THESIS

Department of Information Studies

"A Comparative Analysis of Research Paradigms within Artificial Intelligence"

Submitted by
Emma Ohlin

For the degree of
Master of Science (MSc.) in Information Architecture and User Studies

Total number of characters incl. space: 179 687
Standard pages: 75
Date of submission: 31 July 2018
Supervisor: Niels Ole Finnemann

Abstract

This thesis examines the differences and similarities in the paradigmatic development of the central paradigms Traditional, Neural Networks, and the Evolutionary paradigm within Artificial Intelligence in the time period from the 50's to modern age. In order to examine the differences and similarities, a comparative analysis method was applied to highlight how the concept of Artificial Intelligence has been approached within the different research paradigms of traditional, neural networks and the evolutionary paradigm. Indeed, a comparative analysis methodology allows for a deep analysis of both differences and similarities but also calls for an extended perspective of related philosophical, cultural, social and information processing influences. It is fundamental to understand these influencers in order to outline a comprehensive answer to how the different paradigms utilizes different approaches to examine Artificial Intelligence. The analysis has been conducted involving a hermeneutics approach which has resulted in the suggestion that the different paradigms has connections to philosophical theories such as rationalism and empiricism.

Keywords: **Artificial Intelligence, Reasoning, Information Processing, Problem-Solving, Neural Networks, Consciousness, Natural Language Processing, Evolutionary AI, Rationalism, Empiricism, Robotics, Hermeneutics, Cognitive Science**

Table of Contents

1.0 Introduction

Artificial Intelligence has accelerated tremendously in popularity during the past years. The increased popularity is due to several reasons, including availability, cost and advanced technology. The technology development has generated in large amounts of accessible data which enables for multiple Artificial Intelligence solutions. AI technology has entered our everyday lives, including for example Google Translate, Siri, cloud services, vacuum cleaning robots among plenty of others. AI provides enormous opportunities to improve a wide range of areas as for example healthcare, transportation, finances, communication, manufacturing and many more.

Although AI technology brings endless of opportunities, recent debate has also revolved around the possible impact AI technology might have on human society. It is argued that Artificial Intelligence will fundamentally affect society as we know it today (Vinge, 1993, p. 366; Kurzweil, 2005, p. 204-205). The British theoretical physicist and cosmologist Stephen Hawking (bbc.com, 07.06.2018) claimed that the invention of AI technology will eventually lead to extinction of humankind. However, before we accept the alarming speculations about the possibly unfathomable results that the invention of AI might result in, it is relevant to define what the concept of AI actually means.

The following thesis is about Artificial Intelligence and in particular, the differences and similarities in the paradigmatic development of the central paradigms Traditional, Neural Networks, and the Evolutionary paradigm in the time period from the 50's to modern age. Artificial Intelligence is, according to Oxford Dictionaries, defined as follows;

> The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages (en.oxforddictionaries.com, 06.06.2018).

The research field of Artificial Intelligence is an interdisciplinary area spanning over a wide range of research disciplines which means that it is a topic that can be examined from multiple perspectives. The aim of this thesis is to examine the differences and similarities in the paradigmatic development of the central paradigms Traditional, Neural Networks, and the

Evolutionary paradigm within Artificial Intelligence in the time period from the 50's to modern age. While numerous researchers have interest in Artificial Intelligence with various approaches and aims, the common denominator is that AI is an ambiguous concept and despite the general definition provided above, there are as many other definitions of the concept as there are scientists. This thesis is divided into three paradigms, based on some of the most significant research influences at that time and as I will demonstrate, paradigms are dynamic and I have chosen to examine these three paradigms which I consider to be significant for the AI research and has contributed to valuable knowledge. A comparative analysis methodology was applied to best highlight the similarities and differences between the different research paradigms. In particular, a hermeneutics approach has been utilized which enabled for understanding and deep analysis of the texts and material.

The first part, the Traditional paradigm, concerns information processing, mathematics, logic, rational reasoning and problem-solving. The aim is to explore different pioneers who have influenced the research area of Artificial Intelligence and to look at their most distinctive arguments. This part includes arguments about language processing which is a central part of the whole thesis and for the concept of AI. The Traditional paradigm suggests that consciousness works rationally and the paradigm will be argued to have similarities with philosophical rationalism and cartesianism, where reasoning is about rational and logical decision making.

The second part of the thesis, the Neural Networks paradigm, revolves around the cognitive processes which enables for humans to perform all the complex tasks we usually do nonconsciously or without full attention. What enables all this are the neurons in our brain, consciousness and self-knowledge, and studying these might uncover solutions to how we can build machines with similar characteristics. This paradigm suggests that consciousness works associatively, thanks to the highly complex neural network in our brain which enables us to learn and constantly improve. I will argue that this paradigm, in contrast to the Traditional paradigm, has connections to philosophical empiricism.

The third part of the thesis, the Evolutionary paradigm, has both similarities and differs from the dichotomy between the first and the second chapter. This chapter might be regarded as a chapter with connections to science fiction due to its concern with the development of AI which includes robotics and a postbiological world. The invention and development of intelligent technology forces us to reflect on what it means to be a human and to be alive. Different views on the development of AI technology are considered and it is suggested that

need to be proactive, rather than learning from our mistakes when we develop AI. The thesis ends with a concluding discussion to highlight the most significant similarities and differences.

This thesis has a strong relevance for Information Architecture and User Studies as Artificial Intelligence is a present phenomenon that has emerged with the development of digital information processing technology in recent decades. In this thesis I will both use the term Artificial Intelligence as well as the abbreviation "AI" to refer to Artificial Intelligence.

1.1 Research Question

The research question for this thesis was carefully analyzed and utilized in order to determine how to best communicate the goal of the research question. This thesis as determined and produced one question that will guide my research and methodology;

- What are the differences and similarities in the paradigmatic development of the dominant paradigms Traditional, Neural Networks, and the Evolutionary paradigm within Artificial Intelligence in the time period from the 50's to modern age?

1.2 Hypothesis

The research hypothesis for this thesis were deliberately developed after the creation of the research question. The hypothesis question is based on the research question but it also project a prediction of estimated results that might occur through the implementation of the research. To prove the specified hypothesis based off of methodology and analysis is the objective of the hypothesis but it is also possible to disprove of the hypothesis. This research as determined and produced one hypothesis that match the above research question;

- In a broader sense, the main differences between traditional and neural network theories of Artificial Intelligence is that the first relates to a Cartesian/rationalistic understanding of the consciousness while the latter relates to empiricism. In this thesis I will study new paradigms, assuming that this dichotomy is valid or not applicable.

2.0 Methodology: Comparative Analysis

> "Comparison lies at heart of human reasoning and is always there in the observation of the world─'thinking without comparison is unthinkable'" (Swanson in Rihoux & Ragin, 2009, p. xvii).

Certainly, we engage in comparisons all the time. We know the difference between things in the world because we have compared them. Any sort of categorization or classification is based on comparisons. In an empirical scientific sense, comparisons are therefore crucial (Rihoux & Ragin, 2009, p. xvii). To any study, the choice of method is dependent on the purpose, the research questions and what the study is going to represent (Trost, 1997; 16). Since the purpose of this thesis is to examine differences and similarities in research approaches within different paradigms of Artificial Intelligence from the 50's to today, a comparative analysis method approach was utilized for optimum results. To understand what a qualitative comparative analysis is, it is important to understand both its relation to other methods and the historical context. Systematic comparative methods originate from natural sciences such as anatomy and botany. As Benoît Rihoux and Charles Ragin argue, we have learned the difference between sciences, objects and ideas because we have compared them. A qualitative comparative analysis method consists of the guiding principles "method of agreement" and "method of disagreement" which points out both similarities and contrasting of cases. It is important to emphasize that this kind of method does only produce new discoveries by itself if truly relevant factors have been included in the comparison (Berg-Schlosser, De Meur, Rihoux and Ragin, 2009, p. 2-3). However, by comparing different views and arguments within the field of AI, a comparative analysis method can help establish new viewpoints and arguments on the topic even if this thesis might not produce new discoveries per se.

Comparative analysis methods revolves around cases, which means that they focus on only a limited amount of cases. The reason for the limitation of cases is because a comparative method requires a holistic perspective during the analysis in order not to lose any of the complex combination of properties that each individual case consists of (Berg-Schlosser, De Meur, Rihoux and Ragin, 2009, p. 6). Therefore, a careful case selection is required. The selection can be done in several ways but as already mentioned, the purpose of the study will determine the final selection. The cases must be representative for the domain or the area that is being studied since carefully selected cases will represent a diversity of the field, which will then maximize

factors of interest (Rihoux & Yamasaki, 2008, p. 124-125). In addition to that, a diversity among the cases helps maintaining a holistic perspective of the field. Moreover, it is important that the cases selected for studying needs to be comparable within specified criteria and this is where the adage "compare apples and oranges" comes from (Berg-Schlosser & De Meur, 2008, p. 20-21). It is therefore crucial to have a set purpose of the study before conducting the research since the purpose will set the frames for the selection of cases or literature to study. The cases and literature in this thesis have been carefully selected according to the criteria above and the overall common criteria is the field of Artificial Intelligence.

Although, as I will demonstrate, it is important to highlight that AI is an interdisciplinary research area. Many scientists selected for the comparison have contributed to new findings and even paradigm shifts. Their research has had a great influence on the research of AI which also makes this a representative selection. In particular, I implemented a so called "snowball selection", where I located one significant source who then refers and locates another source and so on. All researchers selected for this comparison are regarded as pioneers within AI and they are therefore representative (Larsson, 2000, p. 58; Denscombe, 1998, p. 142).

More pioneers could have been added since many researchers have contributed to the AI research throughout the years. However, as Dirk Berg-Schlosser, Gisèle De Meur, Benoît Rihoux and Charles Ragin argues, a comparative analysis method requires a limited amount of cases in order for the analysis to keep a focus on the important parts. This argument is further supported by Larsåke Larsson (2000, p. 73) who argues that after a certain point of researching, some information will be repeated since the researchers are analyzing the same phenomena. I have chosen to focus on the source of origin. For example, Alan M. Turing was a computer scientist and mathematician who invented the Turing Test which became an important landmark within early AI. His work has been very influential on successors work within AI and Turing is therefore an important and influential pioneer to study. Many successors have referred to and analyzed his work throughout the years, which sometimes opens up for new perspectives on his work, however, Turing is in this case the primary source which makes him relevant for this study. Moreover, the limitation of cases is also about the amount of workload and in order to reach the time frame, the number of cases had to be selected carefully.

Understanding the difference between primary and secondary sources is of high importance. As Paul A. Frisch (2001, p. 991) explains, "Primary sources provide the documentation for history as it is being made". As already argued, Turing is an example of a primary source whose studies document the history as is was being made. Different traditions tend to define primary and secondary sources differently, however, it is always important to

understand the context of the creation of the source such as who created it and why. What characterizes a primary source is that the researcher has had a firsthand interaction with the data or created the data by observing or conducting research. Significant for secondary sources is secondhand interaction with the data, meaning that the research rest upon sources by other persons (Lombard, 2010, p. 251-253). Despite what sources that are being used, one needs to always be critical by identifying, locating and evaluating the material. To determine whether the primary or the secondary resource is better than the other depends on the researcher's expertise in relation to the topic. If the researcher is an orthopedic studying knee injuries, then interviews with patients are more valuable than secondary sources. On the other hand, an engineering university student might not have any theoretical or practical background, which might make secondary sources (articles written by e.g. engineers) superior to the primary sources (e.g. science lab experiments). However, Emmett Lombard (2010, p. 253) claims that consistent referral combined with understanding might help regarding information literacy and should always be applied.

2.1 Hermeneutics: The Science of Understanding and Interpreting

In order to carry out a comparative analysis, one needs to understand the texts and material to enable comparison. Therefore, I have also applied a hermeneutical approach to create an analysis. The history of hermeneutics has its roots in the ancient Greek and later became a common technique for interpreting the Bible. Hermeneutics is a methodology that explores understanding, interpretation, how we read and handle texts. Hermeneutics is a classical discipline which only concern is the art of understanding texts, to simply deal with possible problems that arises due to meaningful human actions. It is difficult to separate theoretical questions from practical problems that concern almost everyone in hermeneutics. For example, how is the meaning of a text constructed? Is meaning constructed by the readers as they read or is the meaning given by the author within the text? These are complex core questions of the hermeneutics theory (Thiselton, 2009, p. 1-2; Gadamer, 1998, p. 164, 210-211), but questions like these are central within communications theory as well. Communication concerns what is being transmitted by a text and also what is interpreted and understood by the readers or target audiences. The terms "sender" and "receiver" are commonly used within communication theory but also within linguistics to highlight the message transmitting process. Hermeneutics concerns the whole process, involving the text

author and the reader as an act of communication is what distinguishes exegesis from hermeneutics (Thiselton, 2009, p. 3-4). The philosopher Hans-Georg Gadamer (1998, p. 164), who spent most of his research on hermeneutics, argue similarly saying that hermeneutics must be considered as a whole in order for it to do justice to the text. Just focusing on one single part will take it out of its context and will therefore not do justice to a text.

According to Gadamer (1998, p. 210-211, 294), the perfect culmination of human self-awareness is to understand things. The better understanding one reaches, the more one is able to recognize the value of different phenomenons in the world. Gadamer argues that one can fully participate in life if one has an understanding of the environment and the objects and phenomenons in it. The specific content is what determines and guides the readers understanding and the readers unconsciously assumes there is a meaning inherent in the text. However, the meaning and understanding of a text is dependent on what we, as readers, bring to the text prior to reading it. Our understanding is shaped by past experiences, opinions and our knowledge of the subject. Similarly, Wolfgang Iser (in Thiselton, 2009, p. 30) argues that every reader of a text brings something of their own such as perceptions and experiences to the text. Readers also fill in potential gaps in texts where it is non explicit. On the other hand, such arguments can be criticized for being too objective. For example, the literary theorist Stanley Fish (1980, p. 349) argues that everything that exists are interpretations and there is therefore nothing within the actual text which Iser claims. Texts, authors and readers all gets shaped by their place in history. The readers also gets shaped by their own place in society and history as readers and interpreters. This can also be referred to as interest. Interest is a kind of pre-understanding with emphasis on the self-interest, the self-affirmation or the gratification of desire by the self. Two persons can therefore read the same text and interpret it entirely different. Interest is therefore something that rises from self-centered values (Thiselton, 2009, p. 32). The concept of the "self" is something that I will develop later on.

We tend to believe what is being said in a text without questioning it, because we believe that the person who has written it has better insight and knowledge about the subject than we have. Only when we fail at accepting the content as true we seek to understand the text further, in other words, we try to understand another's opinion. Understanding a text is primarily about understanding the content of the text and secondary about understanding another's meaning (Thiselton, 2009, p. 32).

Align with Gadamer, Anthony Thiselton (2009, p. 13-14) suggest that understanding is similar to putting together a jigsaw puzzle. Holding one puzzle piece, we might assume the color green represent a plant or grass and then we try different possibilities where it might fit

in. Piece by piece we are putting together a picture. Some initial guesses will be proven wrong during the process, while some others will be proven right. In order to put together the whole puzzle, we need to entertain some working assumption about what the final picture might represent.

It is not until the very end that we will find out what the picture signifies. This is what Thiselton refers to as the "hermeneutical circle". Although the term "circle" is misleading because what it actually designates is an upward and constructive process from a pre-understanding to a fuller understanding. After gaining more understanding, one will turn back to review and possibly edit or change the preliminary understanding. Analyzing and understanding how researcher's within AI define the concept of Artificial Intelligence can be compared to a jigsaw puzzle, like Gadamer and Thiselton suggests. Understanding one researchers perspective doesn't facilitate an understanding of the subject in a bigger context. I gained a better understanding for each paradigm after reading about the different paradigms. Therefore, I agree with both Thiselton and Gadamer in the sense that understanding requires a holistic picture. Furthermore, the stages of preliminary understanding and understanding will merge together into a process where the different puzzle pieces are examined in order to relate them to the whole picture. To understand the whole picture, we need to scrutinize the individual pieces and get a sense of the picture as a whole, which the individual pieces alone cannot provide (Thiselton, 2009, p. 13-14).

Applying a hermeneutical approach is a way of establishing bridges between contradictory arguments or viewpoints. This helps providing an understanding of diverse motivations and journeys that have led to each perspective or viewpoint. According to Thiselton (2009, p. 5-6), it has even been argued that a hermeneutics approach encourages such receptiveness and open-mindedness that it should be mandatory at all universities. Hermeneutics provides a good training in tolerance, reciprocal listening, mutual respect, patience and integrity. The text actively "speaks for itself" or what a person seeks to understand within hermeneutics. The human inquirer can therefore be regarded as the object of scrutiny (Thiselton, 2009, p. 5-6). Thiselton's argument goes align with Gadamer's argument that meaning and understanding of a text is dependent on what the reader brings to it. I, with an academic background in media and communication, will most likely interpret a text about AI different than someone with an academic background in for example biochemistry. It is also about what one is looking for in a text, two persons might have similar academic, or not academic, background but they might be looking for different arguments or information in the same text. Having said that, one perspective is not better than the other, they are just different.

Hermeneutics is a multifaceted concept and according to Gadamer (1998, p. 307-308), ever since early traditions of hermeneutics, it has been subdivided into three different fields; understanding, interpretation and application. Consisting between these three elements is the concept of understanding. Gadamer emphasizes that these three subdivisions are closely connected where interpretation in fact, is an accurate form of understanding. The third element, application, revolves around that people *apply* their own prejudices when interpreting and trying to understand a text. Hermeneutics can, in this sense, therefore be explained as a process including these three elements. Gadamer further argues that understanding and interpretation to a large degree is about languages. Language is so closely related to our thinking as well as part of cultures and about who we are that language naturally becomes an important part of interpreting. Language is also our main tool for communicating with each other, whether it is written or orally, and language therefore becomes a central part of hermeneutics (Gadamer, 1998, p. 378). In this thesis I will demonstrate that language and communication is a very central part of Artificial Intelligence and therefore, I consider it is necessary to explain the communication model here.

2.2 The Mother of Models: The Communication Model

As both Gadamer and Thiselton argues, communication and language are central to hermeneutics. Therefore, understanding the basics of communication is crucial when applying a hermeneutical approach. A famous and very influential communication model, which is also referred to as "the mother of models", is a model about encoding and decoding by Claude E. Shannon and Warren Weaver (1963, p; 34), see Fig. 1 below. The simple communication model consists of three basic elements that visualize how information transfers from sender A to receiver B. The first phase is the encoding phase. Sender A formulates its message so that receiver B can understand and retrieve the message. In the second phase, the sender selects one or more channels to transfer information to the recipient. The last phase is about how the message is received and interpreted by the receiver. This phase is called decoding. In general, we can consider our perceptual systems as channels for communication.

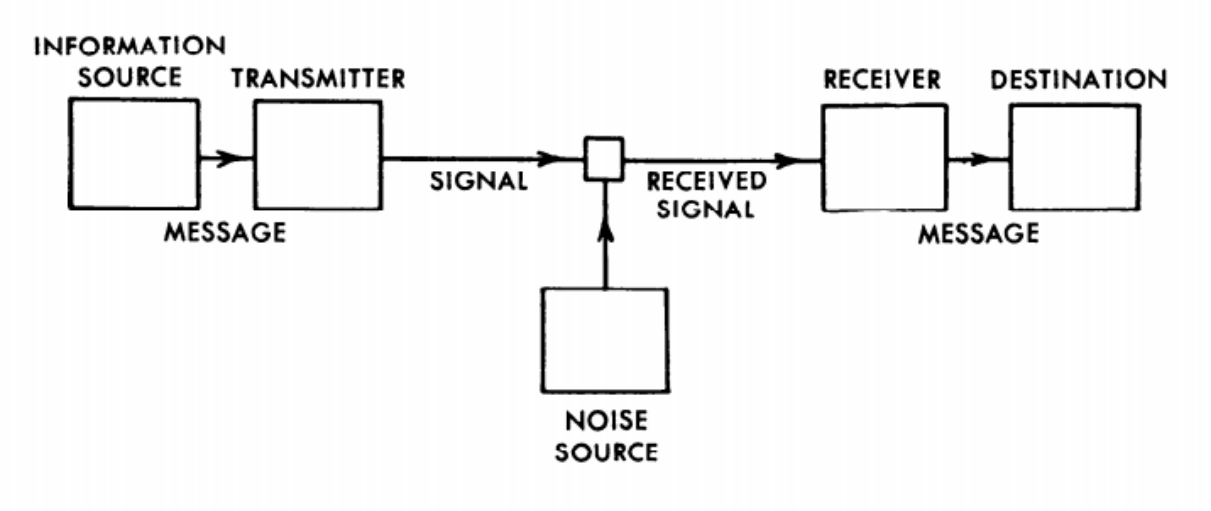

Fig. 1. Communication model by Shannon & Weaver (1963, p; 34).

This kind of model demonstrates a one-way communication process. The fundamental problem with one-way communication processes is the matter of feedback, or more so the lack of it. The sender cannot know how the message is going to be received and interpreted. The authors of the literature I have read and based this thesis on are in this case sender A, encoding information in a book or article as a communication channel and sending the message to me, who decodes their information. How I interpret the information might not be as it was intended by the author. This refers to Gadamers argument that what is being interpreted in a text is also dependent on what the reader brings to it. I might also be the sender A, sending information to receiver B through the thesis. Receiver B, anyone reading the thesis, will then receive the information and interpret it. There are simply no guarantees that whatever is being encoded in information will ever be decoded the way it was intended. Therefore, communicating is always a risk taking where everyone involved decode their own meaning (Durham Peters, 2000, p; 267). However, communicating on a daily basis is just a risk everyone has to take in order to be a part of social contexts such as a society.

The American linguist, cognitive scientists and philosopher Noam Chomsky is by many considered to be the "father of modern linguistics" because of his very influential work on this topic. According to Chomsky (1972, p. 4) we need to study and consider language as a cognitive system we are born with. This system develops as we grow older and is crucial to our ability to understand how languages are acquired. As this thesis will demonstrate, this is one of the greatest challenges within AI and might still be one of the greatest mysteries of the human mind.

Chomsky (in Bermúdez, 2014, p. 17; 1972, p. 100-105) wanted to understand why different languages are structured the way they are. In his book "Syntactic Structures" that was

released in late 50's and became a landmark within linguistics, he focused on the distinction between deep structured sentences or phrase structured sentences (which is how words in a sentence are structured and built according to constituents, the syntactical structures) and surface structures (the actual structure of words in a sentence). In order to specify the phrase structure of a sentence we only need a few words. This also applies on speech and every grammatical sentence. This is all part of what Chomsky refers to as transformational grammar which is a branch of generative grammar. The generative grammar aims to theoretically explain how children will learn their native language in their first years of life, without having encountered all the possible structures in the language. It is a phenomenon called generic grammar called "poverty of stimulus" and is the source of the idea that there is an internal grammar that every person is equipped with from birth. In other words, humans are born with a system that enables for them to learn and understand languages but they will only do so if they are exposed to language. The combination of inherent ability to develop an understanding of languages and the exposing of the language is what enables a human to understand language.

2.3 What It Means To Be Critical

Despite that communication in general might be a risk taking process, there are also other parts to consider. As a receiver of communication, it is crucial to be critical. Determining to which extent the author behind the text have provided adequate justification for their arguments is what critical reading is about. This assessment depends on what relevant knowledge any reader brings to the text, what past experiences and inference one might have. It is also dependent on the communication of the author, how they claim to position themselves in the field and what reason they have to draw the conclusions or arguments they make. On the one hand, as important it is to read critically, it is on the other hand also highly relevant to write self-critically. Convincing others to accept my arguments can be achieved by communicating reasons effectively and adequately and to provide evidence for these arguments (Wallace & Wray, 2011, p. 7). In this thesis, the aim is to examine how the concept of AI is approached within different research paradigms from the 50's to today and my role is to compare these different views critically.

Critical thinking and reading is the capacity to evaluate the text and be able to relate it to other information (Wallace & Wray, 2011, p. 9-11). This refers to both Gadamer and Thiselton who argued for the importance of holistic perspectives and information understanding

as a jigsaw puzzle. One therefore needs to look for potential weaknesses and strengths of the text. We could possibly simply accept what we read at face value if knowledge was just a set of fact. However, apart from the fact itself, knowledge also involves predictions about future facts with the use of past facts and interpretation. In addition to that, well done research means it should be possible beforehand to argue why one is going to read a certain book or an article. The reasons may vary, all from that one have been told to read a certain book or that it addresses a particular problem or topic. Despite the reasons for reading a text, one should always keep in mind how the author positions his/herself in the field and how the text will position itself in regards of other texts one have read. What to read, in what depth and to which extent is dependent on what questions one brings to the text and the reasons for reading that specific text (Wallace & Wray, 2011, p. 9-11).

It is also important to highlight that being critical means to also be critical towards oneself. People will always be affected by expectations, previous knowledge and prejudices. This is what shapes our understanding of the literature we read. The same thing applies to the authors. Both our own expectations and prejudiced and the authors underlying aims and agendas will influence how one come to understand the literature. One way to handle this is to get an understanding of who the author is and what agenda that author have. Taking this into account when reading and evaluating a text is a good condition to keep a distance to the text. "Reading between the lines" might therefore sometimes be valuable (Wallace & Wray, 2011, p. 29). Being critical also includes evaluation. Evaluating the arguments in texts helps to determine how reliable the literature is. According to Mike Wallace and Alison Wray (2011, p. 31) an argument consists of two parts, a conclusion and a warrant. The conclusion consists of one or more claims regarding how something is or should be. The warrant justifies why one should accept the claim or arguments. A warrant is often based on either evidence and findings from the author's research or other's professional evidences or experiences. A reliable and trustworthy conclusion is therefore adequately warranted by convenient evidence. Opinions, on the other hand, are unwarranted conclusions and should always be considered cautiously.

2.4 Revolutionary Science: Paradigms

In this thesis, I have divided and compared different paradigms within Artificial Intelligence. Gathering data, whether qualitative or quantitative, and then relating that data to relevant theories are scientific methods. Science on the other hand, consists of constellations of

facts, methods and theories. Scientific development is an ongoing process where new facts and findings are constantly added (Kuhn, 1962, p. 1-2). A paradigm shift means new and usually more stringent conditions for the research of the subject. "The Structure of Scientific Revolutions" by Thomas Kuhn (1962, p. 23) is by many regarded as a landmark event in the history and philosophy scientific knowledge when it was published. Kuhn argues that the reason paradigms have such status within science is because paradigms are contributing to more successful ways to solve problems than their predecessors. Whenever new facts and theories are added upon the stockpile of knowledge, it reflects on the past scientific work and contributes to a re-evaluation of prior facts. This is what Kuhn (1962, p. 7, 10-11) refers to as a revolutionary process.

Scientific research covers a wide range of subjects and each scientific domain contains its own community traditions, theories and models, in other words, paradigms. A fundamental change in these basic traditions and models within scientific disciplines is referred to as paradigm shifts. When scientists encounters new anomalies that cannot be explained according to current paradigms or theories within that specific domain, a scientific revolution occurs (Kuhn, 1962, p. 12-13). In order for a theory or method to be accepted as a paradigm, it needs to be better than its predecessors and competitors. It also contributes to a more definite definition within the discipline. Having said that, that doesn't mean the new paradigm has to answer to all questions it might confront. Not all agree with new paradigms and choose therefore to stay with older views (Kuhn, 1962, p. 17-19).

In order to understand how paradigms form and takes place, Kuhn argues that it is also important for us to understand that groups of people produces scientific knowledge and change occurs in social environments. Therefore, understanding social changes and social contexts is crucial if we want to understand science (Tredinnick, 2006, p. 188). Moreover, Kuhn (in Wray, 2011, p. 171, 175) emphasizes that part of the success of science is the social contribution, since scientific knowledge is produced and revealed in social contexts. Social changes influences and encourages significant scientific changes, simply because research communities are dynamic.

2.5 A World of Classifications

All new scientific discoveries does not contribute to paradigm shifts despite being better than their predecessors and despite contributing with fundamental change within the specific

domain. According to Kuhn, states of crisis occurs when many significant anomalies have gathered within a current paradigm. It is during this crisis period that new ideas and findings are tried which also means old ones are discarded. Slowly, a new paradigm shift takes place when the new ideas has gained new followers (Kuhn, 1962, p. 17-19). However, as Kuhn stated, not all agree with new paradigms and to some degree, division between paradigms is subjective and varies between domains. It is also important to emphasize that paradigms are not static and how one defines a paradigm shift can also be argued to be dependent or influenced on how we tend to classify the world.

The philosopher René Descartes famous statement "cogito ergo sum" (I am thinking, therefore I am) revolves around the notion that one should doubt all other knowledge except the own conscious processes and reflections. This provides a philosophical thinking starting point in how we can gain knowledge about things in the world, a concept which is also referred to as epistemology (Thiselton, 2009, p. 12). So, how do we know things? How do we know what different concepts means? These might seem as rather simple questions, however, they are highly complex questions without clear answers. Most people would be able to answer such questions, referring to common knowledge. For example, one can say that a dog is a mammal and if challenged, we can simply just look it up. However, in order for all people to live together in societies we have created rules and common agreements. We have therefore named and classified all things in the world. Classifications on the other hand, are never purely neutral. We constantly classify things based on our perceptions on the world, logic, empirical studies, heritage and the like. Some classifications are also based on central theories and knowledge, in other words, according to paradigms. Different scholars also tend to classify differently and there are therefore no correct ways of classifying.

All classifications are not scientific, such as kinds of clothes, pets, colors and so on. However, the one who classifies must have some sort of background knowledge or qualification in order to classify (Hjørland, 2013, p. 169-173). When I classify this thesis I do it based on my prior knowledge, experiences and foremost based on pioneers who have contributed to the paradigm shifts. Wallace and Wray's argument is applicable here, my expectations, aim and prejudices influences how I choose to divide the chapters in this thesis. As Luke Tredinnick (2006, p. 188) describes it, abstracting information always involves interpretative decisions and can therefore never be a neutral activity. This is very significant for any analysis or comparisons, as it always involves interpretative decisions both by the author and the target audience.

3.0 The Traditional Paradigm

For thousands of years, humans have been trying to understand how the human mind and brain works, without coming to any certain conclusions. What Artificial Intelligence aims at, is to build intelligent entities and in order to do that, it is crucial to understand how human functions, what intelligence is and how AI can be achieved (Russell & Norvig, 2010, p. 1). The Traditional paradigm concerns information processing, mathematics, logic, rational reasoning and problem-solving. This paradigm suggests that consciousness works rationally and the paradigm will be argued to have similarities with philosophical rationalism and cartesianism, where reasoning is about rational and logical decision making. This chapter is structured according to theme and scientist.

3.1 Can Machines Think?

In the year of 1950, Alan M. Turing published an article that came to introduce a new research paradigm. In the article "Computing Machinery and Intelligence", Turing proposed the question "Can machines think?". The paper is by many considered as one of the most important papers written on the topic of Artificial Intelligence and Turing's work has inspired many successors throughout the years which in this sense makes him an important primary source well worth of studying. Turing wanted to investigate whether or not machines have the ability to create thoughts and to think. However, Turing soon discovered that both the concepts of a "machine" and to "think" are ambiguous and difficult to define in a clear way. He therefore decided to replace the question by another, less ambiguous one. Instead of asking if machines can think, Turing asked if machines can do what humans (thinking entities) can do (Turing, 1950, p. 433-435). The new question therefore shifted focus from a very complex philosophical problem regarding what 'thinking' really is, to instead focus on different performance capacities of machines and whether systems may or may not generate them.

In order to make an attempt answering his new question, Turing (1950, p. 433-434) invented a game he called "The Imitation Game". The game involves three players where A is a man, B is a woman and C is an interrogator, who may be of either sex. Player A and B are separated from player C, who stays in a different room. They can only communicate through

written notes. The mission of the interrogator (player C) is to determine which of player A and B is the man and which is the woman by asking questions. While player A has a mission to trick the interrogator into making a faulty decision, player B is supposed to help the interrogator into making the right decision. The interrogator only knows of player A and B as labels such as A is 1 and B is 2. To ask a question, the interrogator must direct the question referring to the label and player A and B can answer as they like with their object of the game in mind. This later became famous as and referred to as The Turing Test. At the time, Turing's test was just what was needed for scientists who searched for a way to measure and evaluate their work of intelligent machines (Dreyfus, 1997, p. 73). This is one of the reasons Turing became an important pioneer within computer science. His article is regarded as a turning point for what later became to be referred to as the Traditional paradigm within AI, or sometimes "Classical" paradigm because of its technical and formal approach.

Turing (1950, p. 434) then questions what might happen if player A is replaced by a machine. With the roles switched, the players might act different than they did at first and Turing then argues that the new set up of the game might provide an answer the original question of whether machines can think or not. The way the game is intended to be played is in conditions that prevents the interrogator from both seeing and hearing the other players. In that way, the players are limited when it comes to use of their senses other than reading their notes. The interrogator might ask questions regarding the other players' haircut or a mathematical problem for them to solve. Turing further highlights that the way the game it set up separates physical and intellectual capacities of humans and there is no need in trying to make a thinking machine look more human, since the players senses are limited. Turing points out that the game can be criticized for making it too difficult for the machine (Turing, 1950, p. 435). After testing the game with different setups, Turing came to the conclusion that machines may carry out something that can be considered as "behavior" which can be described as similar to human thinking.

Despite coming to a conclusion regarding his rephrased question, Turing still had questions to answer. First, he had to define the concept of a 'machine'. He emphasizes that the concept of a 'machine' can undoubtedly include any kind of engineering technique. Needless to say, that would obviously be too broad for a game like the imitation game. Turing therefore decided to limit the concept of 'machines' to digital computers, machines that operates by simple rules to rewrite binary digits of 1 and 0 by manipulating them into memory. Turing argued that a digital computer is "universal", meaning that in theory, when given the right time and memory, they can simulate behavior of any other digital machine. He also drew the

conclusion that all digital computers can be considered to be equivalent and that every powerful digital machine have the possibility to act like it is thinking. Based on this argument, a digital machine is therefore a symbol manipulating machine (Turing, 1950, p. 436-438).

Digital computers can be programmed according to the principles outlined above, rewriting and manipulation of binary digits. Turing argued that machines can therefore be programmed to mimic the human actions very closely. Creating any kind of computer involves programming the mechanics behind it. Appropriate instruction tables must be put into the computer in order for it to carry out any desired task. The majority of the digital computers at the time (1950s) had a finite store, meaning they were somehow limited in their possibilities. However, Turing emphasizes that the idea of a computer with unlimited store is not theoretically difficult. Creating and programming such computers that can perform difficult tasks is of great interest, according to Turing. He continues that in only 50 years' time it will be possible to program computers with an enormously large storage capacity that when playing the imitation game the interrogator will make the right identification just after five minutes (Turing, 1950, p. 438-439, 442).

By considering nine different major arguments against AI up to 1950 when the paper was published, Turing (1950, p. 442-453) made an attempt trying to answer his original question whether machines can think or not. His arguments included different viewpoints such as a *Mathematical Objection* arguing that a computer system based on logic can only answer a limited kind of questions or *Lady Lovelace's Objection* based on the English mathematician and writer Ada Lovelace's argument that machines are not able to learn independently, implying that machines can only do what humans order it to do by programming it. As a response, Turing argued that machines can perform other tasks than what we program them to do, they can therefore, in contrary to Lovelace's argument, surprise humans. The arguments I will focus on are the *Theological (Religious) Objection* and *The Consciousness Argument*. I am arguing that these objections are of extra significance since a central theme regarding AI in general revolves around the human consciousness and the human mind.

The Theological (Religious) Objection. This objection states that machines cannot think because thinking is simply a function of a man's immortal soul. According to this argument, no animals or machines can think, only humans have the ability to think. Turing (1950, p. 443) clarifies that he does not agree with this objection, arguing that he is not impressed by past theological arguments and that he might would have been more convinced if animals were classed with humans. Either way, I believe this objection is relevant because in order to know whether machines can think or not, one needs to define what is meant by "thinking" and whether

animals can think or not is a question that still remains unanswered today, which I will later return to. According to the Oxford Dictionary (07.06.18), 'thinking' is defined as; "The process of considering or reasoning about something". Reasoning, on the other hand, is defined as "The action of thinking about something in a logical, sensible way" (en.oxforddictionaries.com, 07.06.2018). Based on these definitions, a machine that can perform logic decisions given the circumstances, can be regarded as a thinking machine. However, the question whether the machine actually thinks or not remains unsolved since it is difficult to prove that the machine make decisions based on *sensible* thoughts.

The Argument from Consciousness. Suggested by professor Geoffrey Jefferson, this argument states that "not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain /.../" (Jefferson in Turing, 1950, p. 445). Turing's response to that is that we should accept the imitation game solely because we simply do not know if individuals, other than ourselves, can experience emotions. To be a machine and describe one's own thinking is the only way to make sure a machine can think. This view is somewhat extreme and solipsist, and Turing emphasizes that he does not think Jefferson would agree with this extreme view. Whether he would agree or not, Turing raises an important point with this argument. Human consciousness is a mystery and as this paper will demonstrate, consciousness plays a significant part in Artificial Intelligence.

3.2 Machine Learning

To build an intelligent machine that can imitate an adult human mind, Turing (1950, p. 455) argues that it is important to consider the development of the adult human brain. He summarizes how the human brain develops into its present state with three steps;

1. The initial state of the mind, which is at birth.
2. The education to which it has been subjected.
3. Other experiences, not to be described as education, to which it has been subjected.

Turing then asks whether it is possible to program a computer that simulates a child's mind rather than an adult mind. He compares a child's brain with a blank notebook, hence, it can easily be programmed and filled with information (Turing, 1950, p. 456). He further divides

the problem into two different parts, one that consists of the programming of a child's mind and another that consists of the educational process. The learning process that involves both rewards and punishments will create desired patterns in a human mind. Turing argues that this learning process to a high degree is similar to the natural selection of evolution. He summarizes it as follows;

Structure of the child machine = Hereditary material
Changes = Mutations
Natural selection = Judgement of the experimenter

Turing was a computer engineer and mathematician which might influenced his research, like Wallace and Wray argues. The author's intentions and prejudices will influence the text. The fact that Turing was a computer scientist might explain why he in his paper provided a definition of what a machine is but not how he defines either thinking or intelligence. As I will show in the Neural Networks paradigm, Turing's arguments circle an important issue regarding cognitive abilities and what can and will be learned by exposure. However, Turing (1950, p. 456-457) does emphasize that a child and a machine does not require the same teaching process, since a child and a machine have different conditions. One of Turing's conclusions is regarding the nature of inherent complexity. He argues that the child machine can either be a complex system with inherent logical inference that has been programmed into it, or the machine could be just as simple as possible with no consistency with general principles.

Another important conclusion of Turing's paper is the ignorance of the experimenter. A machine's internal state during the process of learning is being ignored by the teacher. The machine will perform tasks we sometimes cannot make sense of or what we would just consider to be random. According to Turing, this is characteristic for what we usually consider to be intelligence. As long as the definition of intelligent behavior is consisting of the deviation of the complete determinism and conventional computation and not behavior we consider to be random, it can be regarded as intelligent (Turing, 1950, p. 459). However, by the end of the paper Turing (1950, p. 459) provides a different side of his own argument by emphasizing the importance of random behavior. He argues that some elements of randomness in a learning machine is of value to any system. The reason is that there sometimes might be different ways of doing things or if several unsatisfactory solutions to a problem needs to be investigated in

order to come up with the optimal solution. Random mutations are the process of evolution and that is how beneficial solutions grows.

3.3 Measuring Intelligence

Joseph Weizenbaum (1976, p. 204-205), a German-American computer scientist who by many is considered the founder of the modern AI, criticizes the concept of intelligence. He claims that it requires some sort of reference in order to be measured or compared, which makes it too narrow and therefore simply a meaningless concept. However, in a daily conversation the concept of intelligence does not seem to strike us as meaningless, because we relate intelligence to our own references such as social culture and past experiences and we simply assume that we understand it within our own frames of references. Our assumptions do not make the concept either absolute or universal. Therefore, whenever a scientist or an educator uses the term "intelligence", they ought to set the frame and domain in which they use the term. This is necessary if we want to make sense of the concept and can be applied to other concepts as well.

For example, an uneducated person who does well in her private life but is unable to compose a grammatically correct sentence cannot be compared to an academic person who is an acknowledge genius but makes stupid decisions in her private life. These persons' intelligence cannot be compared because they operate within different intelligence frames, hence the adage "compare apples and oranges". Despite examples like these, there is a general idea that intelligence is something that can be measured along an absolute scale and that it is comparable. It is ideas like this that are responsible for the debate about creating a machine as intelligent, or more intelligent, than a human. Weizenbaum (1976, p. 223) suggests, that instead of asking whether it is possible to create an intelligent machine, we should ask *how much* and *what kind* of intelligence we want the machine to have.

Humans and machines does not face the same problems, they are simply not made of the same components. They also do not deal with issues and situations the same way. Therefore, it makes no sense in measuring intelligence along a standardized linear "intelligence quotient" scale (IQ scale). Weizenbaum (1976, p. 203) argues that this common intelligence scale has caused harm in society and causing people to measure their success and achievements in life according to a standardized scale. Even among human beings the IQ-scale is not doing justice since great achievement can be measured in more ways than just from a scale, which makes the IQ scale incomplete and simply too narrow to measure something as complex as intelligence.

For example, an IQ test aims at measuring different mental skills and abilities such as spatial and mathematical, verbal and logic. However, intelligence is a broad concept and the test fails to include social capabilities, ability to see the bigger picture, relate and connect things or phenomenons and so on.

Arguments about intelligence can also be considered as a criticism of the Turing Test. The human creativity depends on the interplay between the intellect and other qualities such as wisdom and intuition (Weizenbaum, 1976, p. 204), which the Turing Test does not take into account. Turing did not manage to play the Imitation Game successfully, despite different setups. Critics have also questioned whether the test can really give any adequate ground regarding intelligence (Dreyfus, 1997, p. 73). It is therefore possible to question whether the Turing Test really measures intelligence in machines or if it is measuring gratitude of the participants.

Weizenbaums argument about intelligence is highly relevant to the concept of AI. Creating a machine more intelligent than a human, or as intelligent, is a problematic statement since it fails to specify what kind of intelligence and in what domain. There are computers that can perform well in mathematics but no machine has yet managed to pass the Turing Test. If a machine did pass the test, would we consider it as intelligent even if it could only pass the test but not engage in a discussion? According to Weizenbaum (1976, p. 207), machines have the ability to range over any human thought domain. However, Weizenbaums argument can be regarded as pure speculation since there, to this point, does not exist any evidence to prove he is right.

Weizenbaum also highlights the difficulties with natural language processing, emphasizing that humans do not process natural language the same way as a computer processes language. A large part of the human language processing involves memory and associations, which is quite different in a computer. For example, according to Weizenbaum, a computer cannot have hope, laugh or feel empathy. Humans mediates their perception of the world through language. These kinds of considerations raise questions about what it actually means to be a human (Weizenbaum, 1976, p. 209, 212). Humans are brought up in a social environment which is crucial to human's development. The social process has a large influence on how that individual will develop and become (Weizenbaum, 1976, p. 210-211). This argument points toward the same direction as Turing's argument regarding programming "child" machines instead of adult machines. Turing might therefore be right although he never succeeded in reality. Moreover, a computer can calculate faster than any human, it can also make the "right" decisions in certain situations (when playing chess for example). However, it

will always do so according to the humans needs (Weizenbaum, 1976, p. 227), which indicates that machines will always be submissive to humans, a statement Hawking's argument disagrees with.

3.4 Rational Artifacts: Symbol Systems

Another prominent pioneer within AI was the American political scientist and economist Herbert A. Simon. Simon's work revolved around information processing, problem-solving and Artificial Intelligence. He was particularly prominent within AI research because of his work with logic theories and problem solving strategies. "Artificial" according to Simon (1996, p. 3-5), is something that is produced by art instead of nature. Artificial is something that's being characterized by not being either natural or genuine and not pertained by the essence of the matter. In short, artificial is the opposite to natural. Art and design revolves around how things should be, in particular how artifacts can achieve and reach goals. Knowledge about natural objects and phenomena, on the other hand, is natural science. Natural sciences revolves around how things are, systems of logic (Simon, 1996, p. 3-5, 114).

Computers are great examples of artifacts and more specifically, they are physical symbol systems. The reason they are called physical symbol systems is simply because they exist in the real world, operating symbols. Another member of the same category is the human brain. Symbol systems are the main focus of Simon's (1996, p. 22) book "The Sciences of the Artificial" and characteristically for a symbol system is that they are goal-seeking, information processing systems, which is significant for both computers and human brains. A symbol system consists of a set of entities, which can also be referred to as symbols. These are all physical patterns that can create symbol structures that, through time, produces a large collection of symbol structures which is equal to mental images of the outer environment which the symbol system is seeking to adapt. The system generates the environment with more or less details and consequently, reasoning about it. Actions taken by the symbol system are stored in the system's own memory (Simon, 1996, p. 22).

According to Simon, rationality revolves around the ability to make rational decisions based on reasoning, which is the ability to make sense of things, applying logic and verifying facts. Decision-making is a rational process and intelligent computers who provide algorithms for managing and solving complex problems are therefore working rationally. This process, of rational decision-making and finding optimal solutions based on the given information, system

limitations (or limitation of the cognitive process if a human is making the decision) and possible time limitations is what Simon refers to as bounded rationality (Simon, 1996, p. 26). It is possible to say that participants in the Turing Test are operating in a bounded rationality. They are limited regarding use of their senses when playing, they only have limited amount of information and they need to make rational decisions. Even if the Turing Test does not have a limitation regarding time, the idea with bounded rationality still applies, since the purpose is that one should make the most optimal choice given the information available at the time.

Simon (1996, p. 18-19) divides computers into two fields; abstract objects and empirical objects. In the abstract field, he introduces mathematics into the study of computer sciences, which is similar to Turing's definition of computers as machines that uses simple rules to rewrite binary digits of 1 and 0 by manipulating them into memory (Turing, 1950, p. 436-438, Simon, 1996, p. 18). In the empirical field, Simon argues that almost all computers have some sort of organizational features, meaning they are capable of storing symbols, re-coding symbols, copying, moving symbols and comparing symbols (Simon, 1996, p. 19-21). As mentioned earlier, Simon argues that artifacts can achieve and reach goals and this is done by the symbol manipulation. However, this does not mean that the machine, or artifact, actually learns what it is doing. Few would probably argue that a calculator *learns* different equations, yet, what it does is to manipulate binary digits.

Simon (1996, p. 13) argues that an artificial object can imitate the natural through adaptation, which is similar to Turing who argues that a machine can mimic human actions very closely. The invention of the digital computer has widened the range of systems that can imitate and mimic other's behavior. However, a common assumption is that mimicking and imitating the natural will only generate what humans build into it and that the computer only does what it is designed to do. Simon argues that such a statement is not correct. Even when we have as much information as we think is needed for a particular situation, it is still difficult to calculate the outcome (Simon, 1996, p. 14-15). Examples of that are weather predictions and chess playing. Despite correct premises, the implication might look differently than expected. This argument goes align with Turing's argument about the value of randomness and that evolution partly lies in randomness. Computers and machines can therefore contribute to new knowledge and solutions.

In his search for a definition of rationality, Simon (1996, p. 37-38), similar to Turing, explores a game theory. Simon argued that the game theory provides an insight into how people make their decisions and solve problems, which is an important clue into understanding rationality. The so-called 'Prisoner's Dilemma game' is a game that demonstrated how difficult

it is to predict optimally rational action. In the game, each player can choose between two moves. One of the moves are cooperative and the other one is aggressive. Both players will receive a moderate award if they both choose the cooperative one. If only one player chooses the cooperative one and the other one the aggressive one, the aggressor will receive a large reward while the cooperator receives a penalty. If both chooses the aggressive one, they will both receive smaller penalties. Obviously, there is no clear rational strategy in this game. Both players will gain a reward if they both chooses the cooperation but they will gain even more from choosing the aggression. It turns out that bounded rationality happens to produce better outcomes than the opposite, unbounded rationality, in a situation like this. The most valuable contribution from this game theory is that it has shown that rationality is effectively undefinable when competitive parts have unlimited capabilities for outguessing each other (Simon, 1996, p. 37-38). The game illustrates cooperative behavior of two rational participants who might not make the most rational decision despite that both will gain from doing so. This can be applied to real world dilemmas regarding rational decisions and behavior patterns and is therefore commonly used within psychology and social sciences.

Simon (1996, p. 53) claims that human beings are quite simple, at least viewed as behaving systems. Human's complex behavior is a reflection of the very complex environment in which we all find ourselves, indicating that the evolution has enabled for humans to adapt and perform complex tasks. A human being has goals which define the interface between the inner and the outer environment they are in. The behavior reflects largely the outer environment and will reveal some limiting properties from the inner environment. The inner environment is the psychological system and machinery which enables a person to think. One of the most complex cognitive systems of humans is the ability to understand language.

3.5 Natural Language Processing

How can a word such as "strawberry" lead one to think of a certain thing, object or phenomena? How are words connected to mental processes? As seen in the methodology chapter, many scientists have tried to explain language and how humans can learn and process such complex cognitive skill as languages (Minsky, 1988, p. 198). The use of natural language is a very characteristic cognitive skill of the human being. Natural language processing can be described as the interaction between human's natural language and computers artificial language. Simon (1996, p. 75-76) suggests that dependent on the requirements of the task

environment, the human mind constantly adapt itself through individual learning and social transmission of knowledge. However, some leading linguistics (such as Chomsky) have argued that if humans weren't born with an inherent basic machinery for understanding of language, a child would never be able to acquire any skills so complex as speaking and understanding language. A human brain continuously updates the knowledge about the world. It adds new procedures to the already existing as contribution to new ways of managing things and solving problems. This process is learning and it changes the system permanently (Simon, 1996, p. 100). In contrast to the example with a calculator, the human brain updates the knowledge and therefore learns, while the calculator will calculate the equations typed into it but not learn and memorize them.

Any goal-seeking system is connected to the external environment by two different kinds of channels; the sensory, through which it receives information about the environment, and the motor, through which it acts. In addition to that, the system also has a memory, where it stores the information about actions (Simon, 1996, p. 121-122). An infant cannot correlate its sensory information with its actions, except from a few built-in reflexes. Therefore, an important part of the earliest learnings is the particular actions that will contribute to changes in the sensed world. The two separate worlds, sensory and motor, are completely separate and unrelated worlds until the infant learns this knowledge. It is not until the infant gains experiences regarding how the different elements relates to each other, it can act purposefully in the world (Simon, 1996, p. 122). This argument goes align with Chomsky's language theory that the ability to understand language is an innate structure but we only learn it by being exposed to it and our environment. Humans are born with everything that's necessary for learning but we will not develop an understanding if we are not exposed to it.

A program that has been made to model some of the features of human problem solving and to build bridges between these worlds (sensory and motor) is the program GPS, General Problem Solver invented by Simon, J.C. Shaw and Allen Newell (Simon, 1996, p. 122-123). The GPS system was intended as a universal problem solving machine solving any problem that could be programmed in well-formed formulas. The machine managed to solve simple problems but failed at solving more complex problems since they were too complex and difficult. On the sensory part of it, the GPS must represent desired situations or objects, as well as the current situation. In addition to that, it also needs to be able to present any differences between the current and the desired situation. On the motor side, the GPS system needs to recognize actions that change situations or objects. In other words, a GPS system is a system that constantly searches through an often large environment in order to discover sequence

actions that will ultimately lead from one situation to another, a desired one. The GPS system was an important invention since it was one of the first systems separating the information about the problem and the problem solving strategy. However, just like the Turing Test, it was not successful solving more complex problems.

As earlier mentioned, Simon worked closely with Newell, another pioneer within computer science and AI. Focusing on Simon has to do with his description of artificial and natural sciences and his arguments regarding rational symbol systems which is a central part of this thesis. It has also to do with Larsson's (2000, p. 73) arguments about that some information will be repeated after a certain point when conducting research, as outlined in the methodology chapter.

The anthropologist-linguist Edward Sapir (in Minsky, 1988, p. 270) argues similar to Chomsky that humans have an innate system for understanding language and in order to give meaning to words. By combining the innate ideas and knowledge about the words and from external inputs, one can understand the meaning of a text. Definitions of words are of some help but in order for the words to mean something, one needs to combine the word with structures, context, functions and build connections to other things, phenomenon's or objects one knows of. In addition to that, one needs to understand grammar in order to use the words. Exactly how a child can learn a language and grammar are not known.

3.6 Reasoning is Equal to Calculating

A very influential philosopher who contributed with his philosophical implications on AI is Hubert Dreyfus. According to Greek history, all reasoning can ultimately be reduced to some kind of calculating (Dreyfus, 1997, p. 67, 71). The idea has its roots in the Greek logic and geometry where Socrates was one of the first to introduce this theory that later has influenced Western philosophers and scientists. Digital computers literally counts numbers in order to present a result. This goes align with both Turing's definition of machine, where he explains that a digital machine is basically a universal machine that works with symbols and logical operations and Simon's argument about rational symbol systems. Based on this argument, a machines can therefore be argued to reason.

Using digital computers to solve different kinds of problems and simulate intelligent behavior is what Dreyfus (1997, p. 77) defines as AI. This approach to AI is also referred to as a "top-down" approach, which revolves around breaking down smaller segments from the

bigger picture. However, he emphasizes that no computer with an artificial nervous system can ever act just like a human brain, including senses, bodily movements and thinking. The term "artificial" does not mean that scientists are trying to build an artificial person, it would not be possible given the current state (1997, my note) of physics, neurophysiology and chemistry. What the term 'Artificial Intelligence' means, according to Dreyfus, is that scientists and workers within the field are trying to create a program that enables digital information-processing by machines that to a high degree imitates human intelligence. Furthermore, Dreyfus argues that the term 'intelligence' in this matter is also misleading. Despite that robots to a high degree can imitate human behavior and intelligence, no one expects the robot to produce results as human beings. At least not when it comes to social matters such as picking the right partner or understanding the context of a social activity. A robot can, on the other hand, do well in disembodied contexts, such as being able to win in Turing's "imitation game" or calculating mathematical problems.

Dreyfus (1997, p. 78-79) suggests that computers have contributed as much to the technological evolution as the Industrial Revolution. Whatever we learn about intelligence in computers will tell us about the extent of human intelligence and might possibly change the way we look at ourselves. Therefore, many scientists argue that we are about to enter a new revolution, a conceptual revolution that will ultimately change our current understanding of what it means to be a human being.

As already established, language translation is one of the biggest challenges regarding AI. According to Dreyfus (1997, p. 85, 91), it is fairly easy to construct a mechanical dictionary with linguistic items, no matter if they are whole words, groups of words or parts of words, they can then be converted into another language. The first attempts of mechanical language translation was done in the 50's. It later became clear that language translation is more difficult than it first looked like. Dreyfus notes that at the time of the 50's, landing on the moon seemed like science fiction. A few years later, humans managed to land on the moon while machine language translation is still over the horizon. In order to translate a natural language, one needs to fully understand laws of grammar and context. Hence, mechanical dictionaries are too limited to accomplish this. Understanding surrounding words in a sentence is simply not enough. A native speaker can provide meaning in the context of human life which computers and machines cannot (Dreyfus, 1997, p. 92-93, 107, 199). The language philosopher Ludwig Wittgenstein explains the ambiguity of language;

> We are unable to circumscribe the concepts we use, not because we do not know their real definition, but because there is no real "definition" to them. To suppose that there *must* be would be like supposing that whenever children play with a ball they play a game according to strict rules (Wittgenstein in Dreyfus, 1997, p. 108-109).

Wittgenstein (in Dreyfus, 1997, p. 108-109; Tredinnick, 2006, p. 135) focused on the meaning through the use of language in so called language games, as quoted above. Based on logic and formal qualities, Wittgenstein and Chomsky share analytical view on language. In Wittgenstein's view, the meaning of a word is dependent on its use, words does therefore not have a direct correlation with things and are unclear. The link between the word and the meaning is explained by connotation and the meaning of a word is a constant negotiation between the sender and the receiver. In contrast to connotation, there is denotation which refers to the primary meaning of a word. This theory is similar to Shannon and Weavers communication theory, where communication is a constant rather insecure transmission between two parts. The fundamental problem of information transferring communication in general, as outlined in the methodology chapter, is that what is produced at one point, might not be what is interpreted at the point where it is received (Dreyfus, 1997, p. 165; Shannon & Weaver, 1963, p; 34).

To program computers to process natural language is more complex than one might think. The programmer needs to consider the transition from meaningful statements that contain ordinary information to the meaningless discrete bits that will contain the information in a technical sense in which a computer operates. The goal is for AI to do this translation itself. However, even human translators are still today superior to any computer when it comes to translating (Dreyfus, 1997, p. 166) and as already established, language itself is difficult for even humans to get a full understanding of.

3.7 Good Old Fashioned Artificial Intelligence: GOFAI

John Haugeland, a professor within philosophy and cognitive science, took an early interest in Artificial Intelligence and dedicated much of his career studying philosophical implications of AI research. Haugeland's book "Artificial Intelligence: The Very Idea" (1985) received a lot of attention when it was published due to his rather provocative view on Artificial

Intelligence where he argues that "the very idea" is that machine computing and human thinking are essentially the same. Throughout the book, Haugeland tackles central questions within the field of AI, such as "how can something ever mean anything?" or "What are feelings?". While answering his questions in the book, he also proposed a new approach to the research of AI, namely, GOFAI (Good Old Fashioned Artificial Intelligence). Good Old Fashioned Artificial Intelligence is Haugelands (1985, p. 112-113) name on symbolic Artificial Intelligence which is derived from different methods within AI, all based on human-readable representations of problems, in other words, symbolic representations. Any intelligent system must therefore contain some sort of computational subsystem in order to carry out reasonable manipulations. This approach is a branch of cognitive science with particular focus on intelligence and cognitive processes. It assumes that by manipulating symbols, many aspects of intelligence can be achieved, a statement which implies that a machine works rationally and logically. It is important to notice that Haugelands GOFAI and "very idea" about AI are central for the Traditional paradigm, however, he was also a cognitive scientist with a focus on the philosophy of mind. Therefore, some of his ideas will influence even the successor paradigm.

Haugeland (1985, p. 48, 53, 76-77) argues that machines are symbol manipulating systems, which are universal machines according to Turing. Both Turing and Haugeland were inspired by Charles Babbage (in Haugeland, 1985, p. 126-127) who is mostly famous for his invention The Analytical Engine, which was the first mechanical general-purpose computer invented as early as in the 19th century. The Analytical Engine consists of fully programmable operations and if you specify the rules, it can play whatever game you program it to play, within those certain limits, which can be argued as a version of a bounded rationality. This is of significance since this is considered to be one of the most powerful inventions in history of technology and it is the foundation of all computer science. While Turing argued that machines are universal in that sense that they, in theory, can simulate behavior of any digital machine given the right time and memory, Haugeland argued that a computer is an automated formal system (Haugeland, 1985, p. 48).

An automated formal system manipulates tokens and carries out logical calculations. Examples of such systems are games like chess, Chinese checkers and tic-tac-toe, since these games all manipulates symbols or pieces of "X" and "O". All automated formal systems are digital, which can be described as a set of methods and devices that each produces and re-identifies the tokens by itself. It automatically manipulates the different tokens following the rules of that specific system. Therefore, these kinds of systems are information processing systems consisting of symbol patterns. A machine, a physical device, is an example of an

automated formal system. However, exactly how those systems works is still not known. Some explain this as “black boxes”, referring to things we either cannot or do not look closely into (Haugeland, 1985, p. 48, 53, 76-77).

What we both know and do not know about different information processing systems is part of the construction. The term “black box” is therefore a useful term referring to those kinds of systems we don’t have full knowledge about. We can see both the inputs and the outputs but not how they are connected and how one is being manipulated into another (Pasquale, 2015, pp. 2-3). Such kinds of systems are highly present in today’s information society. Modern information technology inherent a fundamental knowledge problem, part of the system is hidden in the “black box” which means that digital information, such as personal information, is difficult to control. Needless to say, this part of the black box is necessary for a functioning society. Information issues with a black box system can be developed further, however, this chapter revolves around intelligent systems and their area of usage and I will therefore not develop such theories further.

The most successful branch of GOFAI are expert systems. These systems developed in the 70’s and are considered to be one of the first successful examples of AI. An expert system can be described as a computer system that makes rational decisions as if it were a human expert. By reasoning, these kinds of systems are designed to solve highly complex problems successfully. Expert systems functions in “micro worlds”, hence, they are experts within a limited area. Examples of domains where these systems have proven to be successful are within medicine identifying diagnoses, microscopic layouts and geological analysis when drilling for oil at oil rigs and these areas are most likely to expand. However, Haugeland (1985, p. 193-194) emphasizes that expert systems are not to be confused with GOFAI or cognitive science, despite being a branch of it. GOFAI have interest in general intelligence and common sense, which expert systems have not. Minsky (1988, p. 72-73) criticizes the expert systems invented at this time since they are centered around solving mathematical problems and are very limited. These systems might be “experts” within a very narrow area but cannot perform or conduct any tasks outside of their expert area. Minsky also points out that these systems solve problems similar to how a child would reason and solve a problem.

3.8 The Very Idea of Artificial Intelligence

According to Haugeland (1985, p. 2), AI is the attempt to build "machines with minds, in the full and literal sense". He argues that machine computing and human thinking are thoroughly the same because we, humans, are "at root, computers ourselves". This is what he also argues to be the very idea of AI. His argument that humans are at root computers is built on his idea that human intelligence is equivalent to an artifice. According to Haugeland, AI is neither science fiction or machines built to mimic human intelligence, instead, AI should strive to be as genuine as possible in order to build machines that think for themselves, although he does not provide an explanation of how this should be achieved.

Haugeland (1985, p. 6-8) argues align with Weizenbaum regarding intelligence tests. Haugeland criticizes IQ tests by saying that such tests are only built to measure degrees of intelligence based on assumptions that different subjects have inherent. Haugeland pointed out that this assumption is problematic because intelligence tests are designed for humans and if we are to apply it to machines, we first need to figure out whether it even makes sense to apply intelligence to machines or not in order to then measure the level of intelligence. Another common critique of "intelligent" machines is that they only repeat what we, humans, have taught them, meaning that their "creativity" is programmed by people. In a technical sense, that is what any computer does, it follows instructions for what it's been programmed to do. However, a similar point can be made about humans, in a sense we are doing what we have learned how to do. We operate in contexts we have knowledge about, or learn about (Haugeland, 1985, p. 9).

In addition to that, Haugeland criticizes the Turing Test saying that the test illustrates that machines can act like humans, for example they can imitate talking. However, even though machines can imitate spoken language, it would not sound like an actual human. Learning and understanding words will enable for a superficial conversation but to fully understand and to make up own opinions/arguments in a conversation one needs to understand the topic, which requires more knowledge (Haugeland, 1985, p. 6-8). He further emphasizes that AI can be programmed to develop expert knowledge within certain areas but Haugeland (1985, p. 209) highlights an issue that haunts AI, namely, graceful and sensible flexibility. A human being can adjust and adapt fast to changing and unexpected conditions, while technology lack this sensibility. Despite these large obstacles to overcome, AI technology is constantly getting better.

3.9 Philosophical Rationalism

In order for us to classify and define objects, phenomenon as ideas, we need to have some understanding of how we perceive the world. According to Birger Hjørland and Jenna Hartel (2003, p. 239), scientific domains are divided into three different areas, ontological theories, epistemological theories and sociological concepts. Ontological theories revolves around objects in the world and human activities, science about what is and exists. Ontological theories investigates how the reality is structured. Epistemology concerns knowledge and how we can obtain knowledge about objects and different roles of observing, languages, traditions and other values in the production of knowledge. The sociological concepts concerns about people, how they are concerned with the objects and how they apply different approaches in order to work with and observe the objects. It is important to emphasize that domains are dynamic and can differ between paradigms and periods. Learning about different views on life and things in the world can help us understand how the world is constructed. As Rihoux and Ragin (2009, p. xvii) argues, we know and learn the difference between things in the world because we compared them. We classify and categorize and by doing so, we can more critically analyze and evaluate things we encounter. Different epistemologies tend to use different approaches when observing and examining objects or areas and it is crucial to observe and being aware of different horizons.

The Traditional paradigm within AI is characterized by mathematics, logic, reasoning and rational decision making. I therefore argue that the Traditional paradigm has its roots within philosophical rationalism or cartesianism, the latter derives from the philosopher René Descartes whose Latinized name was Cartesius. This also means that the hypothesis can be considered as valid. Rationalism concerns mathematical models, logic and computer modeling (Hjørland & Hartel, 2003, p. 240). Descartes is by many considered as the "father of modern philosophy" and is mostly famous for his rationalistic philosophy and his philosophical phase that we saw in the previous chapter, "I think, therefore I am". He argued that this phase is the ultimate knowledge that could not be doubted in any way. One cannot doubt without thinking and thinking cannot be done without existing (Cooper, 1999, p. 97-98). It is important to emphasize that David E. Cooper is a secondary source. Descartes is in this sense the primary source. However, as Lombard (2010, p. 253) argues, in some cases the secondary sources might be superior to the primary ones. The book "Epistemology. The Classic Readings" (1999) consists of central historical texts from the primary sources. In other words, these texts are not edited but selected as the most characteristic texts from some of the most essential philosophers

in modern philosophy. Despite being a selection extracted by Cooper, the same extractions can be seen in other literature which serves the purpose of highlighting the most central parts of philosopher's ideas.

Descartes was a dualist, believing that the soul is separated from the body and that the humans consists of these two parts, the soul and the body. He argued that it is not rational to rely on the senses, because one might be dreaming or items can appear closer or farther away than they actually are and so on. From a dualist viewpoint, the world is constructed this way, it is based on two different kinds of attributes such as matter and soul. This philosophy can also be referred to as Cartesian dualism (Cooper, 1999, p. 107, 111). A dualist perspective can also be seen in the Turing Test where Turing himself pointed out that the way the game was set up separated physical and intellectual human capacities. The player's senses then became limited and focus shifted from perception of the senses to perceptions of the intellect. Descartes argues;

> "I understood that I was a substance whose entire essence or nature consists of thinking and which, in order to exist, does not require any place nor is it dependent on any material" (Descartes in Kenny, 2009, p. 320).

His logic is based on that the aim of philosophy is to doubt all that can be questioned and that all inner conscious activity is regarded as thinking. By questioning all that can be questioned, he tried to answer the question "What am I, this I, who I know exist?". Hi came to the conclusion that he was a "thinking thing", something that has doubts, understands, refuses, denies, wants, does not want, has fantasies and perceptions. This thinking thing exists separated from the body. Furthermore, Descartes claims that on the one hand, understanding the equation "115+28=143" is to simply perceive something with the intellect. To declare that the sentence is true, is on the other hand not an act of the intellect but an act of the will (Kenny, 2009, p. 320-327).

3.10 Summary of the Traditional Paradigm

The Traditional paradigm is characterized by great enthusiasm but less success. The Turing Test marks a milestone in the research of AI and has inspired successors and become a subject of discussion ever since. The test has however been criticized for not measuring true

intelligence, as argued by both Haugeland and Dreyfus. The Traditional paradigm revolves around information, mathematics, logic, problem-solving (despite being less successful) and rational reasoning. It is therefore possible to argue that this research paradigm has connections to philosophical rationalism and cartesianism, where reasoning is about rational decision making and relying on one's senses is not rational for one might for example be dreaming. This is all what characterizes a "top-down" approach to AI, which revolves around breaking down smaller segments from the bigger picture. In the search of how the human can develop its intellect, neurological science became more popular. Examining the brain might uncover keys to how the human intellect functions and understanding this is crucial if one wants to build intelligent machines and the rational and logical ideas within the Traditional paradigm shifted towards empirical, cognitive and neurological ideas within The Neural Networks paradigm.

4.0 The Neural Networks Paradigm

A paradigm shift takes place when new, more successful ways to solve problems are introduced (Kuhn, 1962, p. 23). During the Traditional paradigm, AI research revolved around logic and rationality. It was assumed that consciousness works logically and rationally. In the early 80's, researchers became more interested with how the human mind and brain function and therefore started to simulate knowledge and analytical skills of humans with the help of neural networks. Researchers now started to assume that consciousness works associatively. This chapter is structured according to theme but not according to scientists. While the previous research paradigm has similarities to rationalism, I will now argue that the Neural Networks paradigm on the contrary has similarities with empiricism.

4.1 Philosophical Empiricism

Contrary to Descartes rationalist view of the world and true knowledge, concerning rational reasoning and logic, empiricism is a philosophical doctrine proposed by John Locke (in Cooper, 1999, p. 117-118) and is characterized by knowledge derived from sensory and perceptual experiences. Descartes and Locke are often considered the founders of these opposing philosophical disciplines (Kenny, 2009, p. 348). In empiricism, the only thing that

can be regarded as reliable is what can be proven empirically, rather than innate ideas. Scientists and philosophers within empiricism argues that all knowledge is based on experiences and that knowledge is a subject to falsification, meaning it can logically be proven wrong.

Empiricism focuses on evidence, in particular experiments and revolves around the idea that theories and hypotheses need to be tested by experiments and observations in the natural world instead of a priori (earlier) reasoning or "book-knowledge". Humans do not have innate ideas of the world when they are born, their minds are born as "tabula rasa", as a blank tablet for experiences to leave marks on. This expression of the mind was proposed by the Locke back in the 1600s (Cooper, 1999, p. 117-118, 134-137; Hjørland & Hartel, 2003, p. 240). The Scottish philosopher David Hume (in Dreyfus, 1997, p. 210-211) is another philosopher who became known for his empirical views. His philosophy is based on Locke's theories and he later became very influential in both empiricism and skepticism, which concerns denial and questioning of all knowledge possibilities.

Skepticism means doubting or remaining unconvinced. Epistemological skepticism is on the other hand the philosophical theory that doubts whether we can ever really know anything. Epistemological skepticism goes further to say that human senses are usually unreliable, too often people are fooled by illusions and dreams. If we cannot tell whether we are being fooled by our own perceptions or not, we can never really know anything (Haugeland, 1985, p. 33; Cooper, 1999, p. 117-118, 134-137).

4.2 The Science of Nervous Systems

Any individual or entity that can move, needs to have mechanisms for moving in order to make sure the movement is secure and not arbitrary, no matter what is going on in the outside context. Consider for example an owl and a mouse. For the owl to conduct a silent dive in the night to catch a fleeing mouse, it is crucial that the owl has smooth coordination. (Churchland, 1985, p. 13-14). Animals operate in a moving business, meaning that they feed, fight, flee and reproduce. All moves they make are according to their bodily needs. If the animal's behavior is not at its best, the animal might not survive for as long as it might could have. Consequently, the nervous system needs to coordinate the body correctly, that includes all movements, stored information, needs, perceptions and so on. These bodily needs are different to for example, a plant's needs. Plants takes life as it comes needing only water and sunlight but does not operate in a moving business to get this.

Similar arguments can be made about humans, although we do not operate in a moving business as animals. We can walk down a street, we can walk in stairways, catch a ball and perform numerous of complex movements. Understanding different needs is one step further to understanding how the brain builds and creates the self-representation in order to operate in the outside context (Churchland, 2002, p. 70-71; Churchland, 1985, p. 13-14). What makes all this work is the evolution's magnificent solution of adaptive movement and the very basic elements of the nervous system: Neurons (Churchland, 1985, p. 13-14).

What are neurons and how do they work? A neuron is the very basic element in a nervous system. In more detail, neurons are separated into a cell body, which is called "soma" and processes extending from the soma. The soma is the very core of a cell, containing RNA and manufacturing of proteins which both are transported through the branch system of axons. The processes can be divided into two different parts, axons and dendrites but not all neurons have both. While axons manage the output signals, the dendrites receive and integrate signals. A nervous system of a primate has all the neurons it will ever have from birth (Churchland, 1989, p. 35, 38). Neurons communicate through synapses, which are the connections between the neurons. The neurons are also classified into three different areas; sensory, motor and inter. Each category of neurons have different functions. The sensory neurons transfers physical signals into electrical signals. The motor neurons terminate muscles to produce contractions. The interneurons are a mix of both motor and sensory neurons. Most characteristically for neurons is that they are important building blocks in communication. They receive, send and integrate signals but exactly how this is done is a mystery we are only beginning to understand, they are part of the black box.

Chemical synapses and electrical synapses are examples of two different types of connections between neurons. The electrical synapses are also divided into two fields, those who generates field potentials and gap junctions, which connects the axons and the dendrites (Churchland, 1985, p. 40, 48, 62). All this is relevant in order to understand the mechanisms in the human brain and how it can control and make the body perform complex tasks. It is the network of the neurons that has given the name for the paradigm and the term neural network can refer to either a natural, biological one found in humans and other living organisms, or an artificial one found in AI. Neural Networks can also be referred to as "connectionism" due to the connection between neurons, but since the idea with the networks revolves around the neurons I have chosen to refer to the network as neural networks.

The functional principles of the human brain and the vertebrate's brain has its roots in organisms whose prime concern is competitive advantage in predicting. This allowed for

animals and living organisms to predict and adjust to events in their immediate surroundings. In other words, the better and fastest ones survived the longest, as seen in the example with the owl and the mouse. Understanding how we have come to where we currently are is crucial if we are to predict where we are going from here (Churchland, 1985, p. 14). Compared to silicon chips, brains are significantly slow. Brain events happen in a range of milliseconds while events in silicon chips happen in the range of nanoseconds. However, when it comes to completing a perceptual task, any computer will be outplayed by the brain.

Patricia S. Churchland (2002, p. 124), a Canadian-American analytical philosopher whose research has had a great influence on neurophilosophy and the philosophy of the mind, argues that because the body needs to operate in the outside world, it therefore needs to constantly make good predictions within the right time frame. The predictions and decision need not be the best, at least not for humans who do not operate in the same manner as animals, but it needs to be good enough in order to make a living. The nervous system in the human brain is highly complex and requires excellent nervous coordination of muscles to accomplish ordinary things we might not really have to think about while we do it, such as for example drawing, skating, swimming, climbing trees and speaking languages. By studying the brain and the mechanisms behind it, that might enable for us to build machines with similar mechanisms.

By programming a computer to simulate a human brain, scientist have been able to study patterns of activity in fictive, neural networks. Even if it is still just a model, it does however, provide a valuable heuristic picture of how a neural network operates (Churchland, 1985, p. 414). The nervous system is still the most complex and impressive information-processing device that exists. However, AI approaches have become more sophisticated and have learned how to become experts at things we usually find difficult, such as playing chess and theorem-proving. When it comes to perceptual recognition, the human nervous system is still way ahead of any artificial system (Churchland, 1985, p. 458).

There are dissimilarities between how a human brain and a computer store information. In a computer, all data is stored similar to a library, each piece of data has its own space in a large memory bank. A central processor knows the exact address and date to retrieve the data. A human brain on the other hand, stores the data differently. From just a stimulation, the brain can reconstruct the rest with the help of associations and relationships among the stored information. The human brain stores and retrieves information based on the content rather than the location (Churchland, 1985, p. 459). In addition to that, the brain also separates between movements in the outside environment and movement of the eyes. Thanks to this, we can distinguish between movement of the eyes and motions by for example a cat in the surrounding

environment. Again, even if such achievement on a conscious level is nearly effortless, the coordinations and calculations responsible for this are highly complex. By using memories of certain movements, the brain can coordinate new movements. In other words, it works associatively (Churchland, 2002, p. 85).

Complex coordinations can only be performed by neurons. The neural pattern of connectivity is where the intelligence of the system emerges from. A learning process is constantly ongoing, and the neurons learn to reward the system when neuronal activity is succeeded and weakens the system when the connectivity goes awry (Churchland, 2002, p. 70-71). The brain constantly needs feedback from the joints, muscles and tendons to learn skills such as climbing or playing tennis. The brain is highly connected to the rest of the body to detect and transmit signals. These signals go both ways and the brain gets information about the body through the body-to-brain wirings while on the contrary, the body is controlled by the brain-to-body wirings (Churchland, 2002, p. 90-91). This might seem like a small difference but it is significant for how the brain works. Those smaller differences might inherent the keys to unlock the functions of the brain which ultimately might enable for more complex machines and systems to be built.

4.3 Parallel Distributed Processing (PDP)

What is now referred to as Neural Networks, was used to be referred to as PDP models, and sometimes still does. David Rumelhart, James McClelland and G. E. Hinton (1986, p. 10) asked what makes humans smarter than machines. Humans are not either quicker or more precise than machines but as already outlined, they are far better at perceiving objects in their natural environment, noting objects relations, understanding natural language, retrieving information appropriate to the context, conducting contextually appropriate tasks and an additional large amount of cognitive tasks. Humans are also better at accurately and fluently learning to do all these things. Rumelhart and McClelland's answer to their question is that the human brain has developed a basic computational architecture suited to deal with natural information processing.

When processing elements called units each send both excitatory and inhibitory signals to other units, an information processing is taking place. This is what parallel distributed processing (PDP) is about. In a conventional model, a symbol is a representation, whereas in parallel models, the patterns of activities are instead distributed across a network. This parallel

distributed network states a breakthrough in the science of information processing and what has been found about information processing in both computer science and cognitive psychology. Instead of transmitting large amounts of symbolic information, individual neurons compute and communicate by being connected to a large number of similar units (Churchland, 1985, p. 460-461). The PDP models describes the larger units internal structure just as subatomic physics describes atoms and their internal structures within larger units of chemical structures (Rumelhart, McClelland and Hinton, 1985, p. 10-12). PDP mechanisms is the model in a wide range of areas such as motor control, memory, perceptions and language. Conventional models stores knowledge and information as a static copy of a certain pattern, as Churchland (1985, p. 459) illustrated. In a PDP-model, the patterns are not stored but the connection strengths between each unit are memorized as knowledge and can therefore be re-created. The knowledge about patterns are not stored in the connections between units belonging to that specific pattern, instead they are distributed over a wide range of connections of processing units (Rumelhart, McClelland and Hinton, 1985, p. 113, 31-33). A recognition task only takes a split second for a PDP network as soon as the system is trained. The reason is that the system processes the different stimulus in parallel, which makes it more efficient (Churchland, 1989, p. 168-169).

One of the most distinctive parts with PDP models and neural networks, which also makes them highly relevant to study, is that they learn. The learning situation is divided into two sections;

1. **Associative learning**: Whenever a particular pattern occurs in a set of units, the system can learn to produce another pattern when activated. That kind of learning scheme needs to allow arbitrary patterns on the sets of units to be able to produce another similar system.
2. **Regularity discovery**: The units learn to respond to different patterns in their input. Such a scheme must be able to form development of future detectors and at the same time knowledge representation in a PDP system.

Programming systems and machines to *learn* enables for more complex technology. As with the example of the calculator, it manipulates symbols but does not learn and constantly needs to be programmed in order for it to be updated. A machine that can learn can manage to update itself. Another one of the key features of the PDP model is that learning ability is not located in the units themselves but in the connection between the units. The neurons constantly sends activation or inhibitions between each other to communicate, also meaning that

information is always available. Learning is therefore a process of modifying the connections and their strengths. The simplicity and homogeneity of the learning procedure allow for fast and very useful learning procedures (Rumelhart, McClelland and Hinton, 1985, p. 132-133).

Many scientists have argued that the human cognition and its network system is similar or works the same way as lower animals cognition, such as rats or dogs. However, it is still difficult to explain exactly why humans are considered as smarter or more intelligent than other living organisms with similar cognition. What can be established is that a human brain has more brain capacity than any other living organism, in particular, a human brain has more cerebral cortex (a region in the brain that is significant regarding memory, perception, cognition, attention, language, thoughts and consciousness). One significant reason that sets human apart from other living organisms is that human brains can form the connections required for language processing and the cultural environment humans have learned to organize their lives in (Rumelhart, McClelland and Hinton, 1985, p. 143). This argument points toward the same direction as Chomsky's (1972, p. 4) argument about inherent cognitive abilities that processes languages.

4.4 Associative Learning

To train a network, the network needs a "teacher" that monitors the responses and adjust the inputs and outputs into the network. After being repeated several times, the network slowly adjusts and learns the desired behavior. In fact, the desired behavior will be provided in ninety percent of the time (Churchland, 1989, p. 166-167), which makes neural networks so efficient and powerful. The teacher will be a part of shaping the character of the internal representations and the patterns of the environmental responses. Relatively simple networks can therefore learn to recognize cultural auditory features such as phonemes. A human infant will therefore early learn how to recognize and respond to its nearest environment since the "teacher" is the culture in the local environment in which the infant is born into. Linguistic, social and conceptual surroundings all play a part in shaping the child's consciousness. The child's brain will develop consciousness and in detail start reflecting on both elements and structures in the surrounding environment. Humans does therefore continuously adjust to the culture in which we are raised and our consciousness is determined by the cultural context (Churchland, 1989, p. 130-131).

One network called NETtalk is of significance since it has learned how to transform written words and sentences into audible speech. The system, however, has no understanding

of the words involved but that does not affect the systems performance. After significant training the system began to perform almost perfectly with only minor errors. This is worth mentioning because of several reasons. One reason is that the system has over time learned how to master a very irregular input/output transformation which also means that it needs to be sensitive of lexical context in order to function well. It is therefore performing in a highly complex way (Churchland, 1989, p. 170).

Churchland (2002, p. 321) argues that we can articulate certain things, for example instructions for putting together a table but other things we cannot articulate, such as fact retrieval from memory. While learning some things such as driving a car, we need to try persistently while we immediately can learn not to touch nettles, since last time you did, you were stung. In addition to that, we can draw conclusions on things we have not considered before, from the fact we already know, such as knowing that sharks do not knit. Some things that we know is called logical truths. An example of a logical truth is that it is false that cucumbers are both green and not green. There are also factual truths, for example, dogs have been domesticated, and some truths are a combination of both. An example of that is that you cannot physically exist in two places at the same time. This is all important because it affects how we learn.

The discussion about epistemology is a dichotomy between how much of the knowledge we have is based on experiences and how much we have innate. From a rationalist viewpoint, all knowledge is innate. Some knowledge is instinctual, such as a kitten will land on its feet if you throw it into the air. The fact that oxygen is an element or how to grow vegetables is a knowledge we come to learn. These examples show that the knowledge we have comes from either experience or we have it in our genes (Churchland, 2002, p. 322). The different domains of neuroscience, ethology, psychology, molecular biology are all teaching us about ourselves, what it is to learn, remember, forget and how our brains functions (Churchland, 2002, p. 366).

In contrary to a top-down approach, which is characterized by breaking down smaller segments from the bigger picture, a bottom-up approach focuses first on the smaller parts in order to then create a bigger picture. By studying neurons and neural networks, scientists adapted a bottom-up approach to achieve the overall goal, building intelligent machines. A suitable top-down question might be something similar to; “How could neuron-like elements in network N interact to produce global effect E?”. On the contrary, a bottom-up question might therefore be something such as; “Do the neurons in network N produce global effect E by conforming to algorithm A?”. However, Churchland argues that this is misleading despite that the contrast between the two is quite clear. Instead, she suggests calling the two types of

questions as theory-devising (the first question) and theory-testing question (the second). Churchland (1985, p. 462) further explains that what characterizes a top-down approach is if the strategy cares of how the brain perform the processing of information under study. To call the first question a top-down, is to limit the question of the significance it has, because the "how could" questions are limited by considerations of neural networks which the classical top-down computer models are not. However, Churchland emphasizes that this just an issue regarding choosing of words and the contrast between the two types are not useful anyway. Instead, she also suggests that researchers should consider parallel models as theories concerning inputs and outputs. Based on this, the line between the two different approaches seem rather blurry and many modern software developers combine the two approaches to utilize the best results (Churchland, 1985, p. 462).

According to Haugeland, the human brain and computers are essentially the same. Churchland (2002, p. 284) disagrees with that statement, saying that computers and brains might have a few things in common, but they differ on many others. Some of the major differences between a brain and a computer are;

- Neurons constantly make new connections while they also abandon old contacts and learn as they structurally change and both strengthens and weakens old connections. They are therefore very dynamic.
- Computers have serial organizations while the nervous system have a parallel system.
- The nervous system has evolved through a natural selection. Computers on the other hand, have been developed and designed by humans to basically manage numbers.

4.5 Self-Knowledge and Consciousness

The concept of "self" is an ambiguous concept and although there are no clear definitions of how it should be defined, many agree that the self is about who we are and who we ought to be. The self represents the self-image we have of ourselves and it determines goals, expressions and emotions. For example, one might tell oneself to exercise or get a certain job or education. Being self-aware means that one has a perception of the physical body in order to navigate in the external environment and in addition to that, one will have a sense of personal identity and privacy (Minsky, 1988, p. 39, 44, 51). The concept of self is closely connected with other ambiguous concepts such as consciousness and mind and these are all crucial to what it means being alive. Humans are self-constituting entities and when human culture evolve and

develops, the individual human consciousness develops with it (Churchland, 1989, p. 129). As Churchland outlined, the ability to switch between a fight and flight mode are so important it can be life-saving. Such movements affects the whole body including the internal organs. The brainstem controls all afferents (the input neurons) from the whole somatic sensory system, in vertebrates. This is what controls the vital functions that controls and regulates sleep, wakefulness and dreaming. This coordination has a key role regarding the representation of the self and the consciousness (Churchland, 2002, p. 71).

Descartes (in Churchland, 2002, p. 59-61; Cooper, 1999, p. 117-118) claimed that the essential self is not a physical thing. It is not attached to a body or a physical item. He argued that the self is therefore a conscious thing. Hume examined whether the self could actually be only a conscious thing without a body. He came to disagree with Descartes and argued that it is only possible to introspect complex visual perceptions, sounds, emotions, smells, memories, thoughts and so on but no "tangible self", as one can say "this is my hand", one cannot point somewhere specific and say "this is the self".

The problem, according to Hume, is that one can think of oneself as a person/something, yet the self is not something that can be observed. The question is, what is the self? Churchland (2002, p. 61-62) argues that at this time, it is possible to propose an answer to Hume's question. The proposed answer lies within the frames of neuroscience, Churchland suggests. To think and construct thoughts are activities performed by the brain, which means that thoughts about the self is something that the brain constructs. In order to be able to think about the self, the body needs to have a brain, the body and the brain needs then to be closely connected through communication. Churchland further claims that there is a very general answer from an evolutionary perspective, on why brains even construct a concept like the self. The answer is that the self controls and coordinates movements with needs and perceptions. From the evolutionary perspective, such coordination is crucial to survival for animals. For example, the self controls the animal not to eat itself when it gets hungry.

Churchland (1985, p. 64, 67) suggests that the self can also be defined as a representation. From a neuroanatomical perspective, representations are activity patterns which carry information between neurons in the brain. Such patterns can embody information that food is needed, it is windy outside or the water in the shower is too cold. However, the brain can sometimes fool the body. An example of that is auditory hallucinations. These hallucinations are usually part of the diagnostic of schizophrenia but they provide an example of when the self-representational capacity fails to work. Auditory hallucinations are when the brain constructs "voices" heard by the own inner speech. It can also occur as the person's own

voice whispering. What is characteristic about these hallucinations is that they appear very realistic to the person experiencing them.

Some people that suffers from schizophrenia can be confused about their own personal identity. They can for example be convinced that he/she is Jesus and tend to behave/dress/appear as that person. Churchland also mentions that certain drugs can have a "depersonalization effect" when the patient may experience that their body is separated from their thoughts (Churchland, 2002, p. 67). These arguments contradicts Hume's empirical philosophy and instead points towards a rationalist viewpoint. According to Descartes rationalist view, it is not rational to rely on the senses since one might be dreaming or things might appear different than what they actually look like. A person that suffers from hallucinations might therefore not rely on the senses. However, such contradictory arguments might be explained by the philosophical view "Cartesian doubt", which questions the truth or beliefs. By questioning and doubting, one might be able to define the possibility of a certain knowledge. This is done by ruling out all that can be doubted and what is left is the truth (Cooper, 1999, p. 107, 111).

According to the Oxford Dictionary, the word "consciousness" is defined as follows; "the state of being aware of and responsive to one's surroundings" (en.oxforddictionaries.com, 2018.06.07). Consciousness is therefore a state of mind and cannot exist without the mind. Significant for consciousness is that it is experienced in a first person perspective, owned by every individual organism and cannot be observed by anyone else. Part of the consciousness experience is the bodily sensations such as visual and auditory perceptions. To be conscious is to be awake, to have an operational mind and to have a sense and experience of self. Consciousness is therefore something that is *felt* (Damasio, 2012, p. 157-158). However, here is where the real trouble begins, because consciousness is a particular state of mind since the mind can also be unconscious. Wakefulness is not to be confused with consciousness. A person can be forced to sleep with anesthesia or fall asleep naturally and therefore become unconscious. A patient in a vegetative state or coma, cannot respond in a regular sense as if they were conscious but their brain still produces electrical wave patterns as produced in a living, conscious brain. In other words, a patient in a vegetative state or coma is not conscious but unconscious. Yet, it's not possible to wake a person in this kind of consciousness like waking up a person in a deep sleep. Neither is the person brain dead. This is therefore one of the mysteries with the brain and its different states (Damasio, 2012, p. 161), which again refers to the black box.

Consciousness is something dynamic that alternates between different states of sharpness, from dull to very sharp. For example, sitting at home enjoying a cup of coffee on a day off, allows for little scope of consciousness. One is just being present at the moment. Drinking a cup of coffee while being interviewed for a possible job requires a different scope of consciousness. One would still be present at the moment, yet mentally one is being transported to different potential situations and scenarios. In addition to that, we can walk down a street thinking about what groceries to get before getting home instead of fully focusing on the actual walking. Still, we get where we intend going, getting the groceries needed and arrive home safely. The only conscious part of that journey was remembering what groceries to get and to arrive at the destination, the rest was nonconscious exercising. Nonconscious exercising is what characterizes our daily activities and is one of the big wonders about the brain. This ability is, according to the neuroscientist Antonio Damasio (2012, p. 168, 270), is what makes us who we are. It is the very core and essence of being human, to remember, visualize and imagine. This conscious process is what is constantly being illustrated and visualized by films, music and novels.

Robots are logical, efficient and insensitive, in other words, they do not have any feelings. This can be argued to be either a blessing or a curse. Some argue that feelings are what makes life worth living. Others argues that it would be a blessing not to have to deal with disruptions of mood and emotions. However, Haugeland (1985, p. 230) questions whether AI needs to have feelings or not. Human beings can experience a wide range of different feelings, including sensations. We can be sad, feel lonely, fall in love, feel lust, become angry, feel jealous, understand irony, feel empathy and so on. We can sense the taste of salt, feel hungry, pain, tickling and itching for example. All these sensations are what makes us humans and they are helpful to us in order to function in this world. AI uses sensations to some degree in order to function and to navigate their way around their environment. They can also detect malfunctions and internal damage. Haugeland continues by saying that if robots can navigate and operate to do their work, what does it matter whether they have feelings or not? Emotions and sensations do affect human's goals and actions (Haugeland, 1985, p. 235-236), because they determine movements, reactions and what one might want to achieve. To develop the discussion about feelings further, Haugeland (1985, p. 245) argues that any system with a "self-image" needs to be able to reflect of oneself and to represent itself. In order to have a sense of self, the machine needs to have a repertoire of metacognitive abilities.

Whether non-human, non-language beings have a consciousness or not has not yet been satisfactorily proven. Although, it is possible to circulate substantial evidence indicating that

other beings except from humans do have a consciousness. If, for example, an animal has a behavior caused by brain processes rather than just reflexes and uses senses to operate in the world, it will likely be conscious. Damasio (2012, p. 171-172, 180) argues that the highest scope of consciousness requires language processing, which so far only has been observed in human beings, which also goes align with Chomsky's argument that humans are the only ones with a cognitive system for language processing. Furthermore, a conscious mind is very complex and it therefore goes without saying that creating an artificial conscious mind is highly challenging. The human mind and brain has developed into its current state over millions of years and biological evolution.

A philosophical term within subjective phenomenons is qualia. It is important to emphasize that the concept of qualia can be used in different ways dependent on in which context it is used and within what domain. Here, the term is used to describe the subjective quality of conscious experiences. Experiences or feelings that are difficult, if not impossible, to describe are examples of qualia. Such feelings or experiences are the taste of chocolate, the experience of the color blue or physical pain just to mention a few. The problem with qualia is why it even exists. Why should perceptions feel anything at all? Damasio (2012, p. 253-256, 262) argues that qualia is part of the conscious process and that this is the very essence of what it means to be a human being, to feel and experience things in the world. However, despite extensive research, scientists are not sure how the brain can create a conscious mind.

Damasio (2012, p. 284-285) highlights that consciousness appeared later in the history of evolution. There are no signs that bacteria or cellular organisms such as fungi or plants, were conscious. So why did the evolution develop consciousness? According to Damasio, the appearance of neurons is what changed life. Neurons enabled for a more complex life than the life of a plant or bacteria. Neurons enabled for more elaborate behaviors, different mind processes and consciousness. Whenever you are awake, you are aware of things such as your surroundings, sounds, smells, your feelings. Similar to Thiselton (2009, p. 13-14), Churchland describes the learning process about neurons and consciousness, as putting together a jigsaw puzzle to create a bigger picture.

Despite our knowledge about visual perception, being awake, paying attention and so forth, that still does not fully explain what consciousness is. Churchland (2002, p. 171-172) argues that learning about consciousness is basically equal to learning about what is it to be alive. Biology can provide a possible answer to what being alive means. Being alive comes down to modern cell biology, physiology, molecular biology and evolution. From a biological perspective, "life" can be explained by cell structures, cytoplasm, mitochondria, energy

production and so on. However, cell biology does not give an answer to what life itself is. To provide a possible answer to what life itself is, one needs to dig deeper into philosophy, physics and biology. Churchland explains that understanding the physical processes such as metabolism and protein building for something to be alive, is one step further in understanding life itself.

4.6 Meaningfulness

Significant to humans, is that we assign meaning to things in our surroundings dependent on our thoughts. With the help of our words we give our thoughts meaning. If the word "cat" has a particular meaning in the English language it is because we have given it a meaning. The question is where meaningfulness originate. Haugeland (1985, p. 25-27) suggests that some meanings derive from public symbols, such as language and others from internal symbols, such as thoughts and language is an important property of cognitive science. Semiotics revolves around sign systems or anything that can be interpreted as a sign, regardless of the intention (Tredinnick, 2006, p. 143), which goes align with what Haugeland describes. In a semiotic system, a sign is anything that represents something else and one of the most studied sign systems are languages. Other sign systems include hypertexts, advertising, classification systems and indexes. Signs are subjective and sign interpretation is therefore individual. Semiotic theory is therefore a natural component when it comes to information processing systems. Semiosis, on the other hand, describes the meaning within a sign system. Semiotics therefore suggests that meaning is carried through semiosis and the interrelation of signs.

Ferdinand de Saussure and Charles Sanders Peirce (in Tredinnick, 2006, p. 145-147, 155) are two influential linguists of the modern semiotics. Saussure makes a distinction between language and speech, where he argues that language and linguistic habits is what allows for individuals to be understood and to understand. Speech on the other hand, is individual expressions of the language system. He also argued that speech is the most authentic use of language. Furthermore, he argues that what represents the image of the world (something that can be observed, touched, smelled, tasted or heard) is the *signifier*, while the mental construct is what the image signified and this is referred to as the *signified*. The latter concerns mental images in anyone receiving a signifier. This model circles the relationship and the issue surrounding signs and their meaning. Peirce made a similar model of signs but with a different aim. While Saussure was interested in understanding language itself, Peirce wanted to examine the relationship between language and logic. According to Peirce, the substance of a sign is the

representamen, the *interpretant* is similar to Saussure's signified, the idea of the object which is signifies. Lastly, the *immediate object* is the thing itself. Both these models illustrates the complexity of language and addresses the problem as well as the relationship between representations and meaning. These models can be regarded as theoretical tools to approach semiotics.

Haugeland (1985, p. 94) continues that when we interpret, we are making sense of words and our surroundings. We make sense of systems of marks or tokens by specifying what they mean. The keyword to all interpretation is coherence, all interpretations needs to be coherent in order for us to make sense of them. A question one might then ask is, what is it to make sense? A very short answer according to Haugeland is that the truth is crucial for making sense. Therefore, falsehood is incoherent. However, there is more to sense-making than the truth. The truth can be considered as a form of validation, since meaning in general relates symbols to their objects but even sensible people do make mistakes in this matter.

4.7 Summary of the Neural Networks Paradigm

The Neural Networks paradigm, revolves around the cognitive processes which enables for humans to perform all the complex tasks we usually do nonconsciously or without full attention, as for example walking, singing, swimming, navigating and throw or catch a ball. What enables all this are the neurons in our brain, consciousness and self-knowledge. This paradigm suggests that consciousness works associatively, thanks to the highly complex neural network in our brain which enables us to learn and constantly improve. It has been demonstrated that in contrast to the Traditional paradigm, the Neural Networks paradigm has connections to philosophical empiricism, which indicates that the hypothesis is valid.

5.0 The Evolutionary Paradigm

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would

> then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control (Good, 1966, p. 33-34).

The quote above comes from the British mathematician and cryptologist Irving John Good (1966, p. 33-34) who worked with Turing and is known for his work with statistics. With increased computer power, he predicted that it will become possible to build superintelligent machines that exceeds humans in intelligence and problem-solving skills. A statement many scientists would agree with today. What is distinctive for this paradigm is that if or when we have created intelligent agents we might be passing an era where the natural evolution has driven the development into a new era driven by the artificial. In other words, an Evolutionary paradigm. This era is also referred to as “technological singularity”, which put emphasis on an Artificial Superintelligence that will far surpass the human intelligence which will ultimately have unfathomable results for the human civilization. However, I will use the term evolutionary because as this chapter will demonstrate, a superintelligence might not result in human extinction but possibly quite the contrary. As my hypothesis stated, this paradigm is not characterized by the dichotomy seen in the previous paradigms. This chapter is structured according to theme.

5.1 Children of Our Minds

For more than million years we have developed and evolved biologically. We have built societies and cities to live in but the cultural development goes faster than the biological development which have resulted in large cities with skyscrapers, we are traveling to space, we are traveling more in general and utilizes the Earth's resources at a higher speed than the biological evolution evolves. This requires great adaptability for anyone living in the modern society (Moravec, 1998, p. 6-7). This cultural development at a high speed has contributed to a dichotomy between unhappy voices and technological optimists. The unhappy voices want to restore humanity and nature while the technological optimists are positive to any technological development. While a robotic society will produce the same work and result as humans but more efficiently and while some people turn against the technological developments it leaves

the opportunities to others. This means that in the future, the technological industry will grow away from earth. The artificial progeny will develop fast and eventually grow away from and beyond us which some argue could lead to our own distinction (Moravec, 1998, p. 9-13). This argument goes align with Goods prediction he made in the 60's but what would this development look like and why would we develop such technology if it might mean the end of human societies as we know it?

The futurist, whose work within robotics has been highly influential, Hans Moravec (1988, p. 1) argues that a "postbiological" era is what awaits humans. This postbiological world is a world where humans have been passed from the center stage which have been overtaken by the artificial progeny. Consequences from this are unknown but Moravec plays with the thought of different scenarios. He claims that current computers are barely worthy of being called intelligent. They are simple creations that needs human support, which he refers to as "parental care" and current machines can be compared to newborn babies. Needless to say, babies develop fast and so will the machines and computers do too. Within the next century, these machines will mature and become as complex as grown human beings and eventually they will transcend into something we know nothing about - the postbiological era. This process, from being a machine in need of parental care to a grown entity which no longer needs its parental care, is what Moravec calls "the children of our minds".

5.1.1 Universal Robots: First-Generation

Moravec (1998, p. 95-96) divides robots into different generations, based on their features and what they are capable of. He estimated that the first generation of universal robots would be available in 2010. These are general-purpose robots with basic perception and mobility. These robots would operate according to our needs and their perception and mobility would be adjusted to our lives, such as home environments where the robot needs to be able to move around, behind and underneath furniture, on flat ground and so on. An example of a first generation robot is a robot vacuum cleaner, such as Roomba. These kinds of robots use sensors to perceive their surroundings and sometimes even cameras to create a 3D map of the environment. These robots have a processing power similar to a lizards and will be advanced enough to handle speech and text recognition meaning that they can take instructions and read well enough to conduct commands.

5.1.2 Universal Robots: Second-Generation

The second-generation of robots will, according to Moravec (1998, p. 98-101) make their entrance by 2020. These robots will have the processing power similar to a mouse and the biggest difference between a first- and a second-generation robot, is that the latter will have a learning ability. While the first-generation robot repeats its mistakes, the second-generation one can learn and adapt to new situations and environments. It will gradually improve its performance and be able to find own solutions to problems it encounters. Some of these robots learn by themselves, while others learn by the help of assisting humans. Not knowing exactly how she works, Sophia might be an example of a second-generation robot, which would also mean that Moravec did a correct estimation regarding when these robots would make their entrance.

5.1.3 Universal Robots: Third-Generation

Fast forward 10 years; the third-generation robots will make appearance by the year of 2030 (1998, p. 104-107). These robots will have a processing power similar to a monkey. Moravec estimates that these robots will learn much faster than their predecessors and they will continuously practice trial and error. These robots will be able to simulate their surrounding environment in real time and predict possible consequences from their actions, which makes these robots somewhat “conscious”. By simulating certain tasks with the help of conditioning systems, they learn from those simulated experiences and can therefore perform and succeed at the given task. When a third-generation robot has some spare time, it can repeat some previous tasks by itself by replaying the process in order to make them more efficient or come up with alternative approaches and solutions for future improvements. Another significant improvement from the second-generation robots is that the third-generation robots can imitate and self-invent. By just watching someone else conducting a task, it can take notice and formulate a program to do the task itself. However, although it might sound like the third-generation robots are very much self-independent, they still rely on external programs to perform complex and complicated tasks.

5.1.4 Universal Robots: Fourth-Generation

By 2040 robots with the processing power equivalent to humans will make their entrance. These robots will be able to reason and they will be much better at reasoning than human beings because of their high processing power. Moravec (1998, p. 108-109, 124) argues that the reasoning programs existing at 1998 were limited to information prepared by humans which was in small amounts and very unambiguous. One significant quality of the fourth-generation universal robots is that they will be able to device robot programs both for themselves and other robots. If a fourth-generation robot was told "the bathtub is filling up with water", the robot will be able to understand that after a while the water needs to be turned off in order for the tub not to be flooded (Moravec, 1998, p. 109). The fourth-generation of robots will be more powerful than any of its predecessors and will be able to design its own successor. These intelligent robots will have a consciousness and emotions which is frightening to some. Moravec (1998, p. 111) argues that imagining a robot with feelings and a consciousness is frightening because it is contradictory to our perception of nature. However, Moravec argues further that space travel was at first, and maybe still is, frightening to some. If we are to develop these intelligent robots, we are also obliged to decide how much we are going to integrate them into our society, which also points toward the same direction as Goods' argument. If robots can design its own successor, the fourth-generation robots might be the last invention humans will ever make.

As seen in the previous chapter, having feelings might be either a blessing or a curse and a similar argument can be made about consciousness. Being aware of the surroundings and environment makes us take notice about potentially dangerous situations. This is both a good and bad quality. It is good in that sense that we can take into account potential risks and therefore avoid certain situations we consider too risky. However, it can also make us rather paranoid, being afraid of things that might happen. According to Moravec (1998, p. 115), the fourth-generation robot can either be configured to become more conscious than humans and therefore also more likely to avoid tasks and situations. Configured the other way, less conscious, makes it a less careful robot, a robot that gets the job done regardless of possible dangers. Moravec (1998, p. 118) further argues that the consciousness of robots can be similar to animals and humans and the reason they develop it is simply because it makes them able to deal with uncertainties that is life, which is similar to both us and animals.

Despite consciousness and some emotions found in robots, not all human emotions make sense in a robot. One example of that are romantic feelings. The reason, according to

Moravec (1998, p. 118), is since the robot cannot reproduce itself, there is no need for sexual feelings. On the other hand, for business reasons it might be strategic to install programs that enables for the robot to develop feelings similar to love and loyalty. This will allow for the robot to consider possible effects of its actions in regards of the human it encounters. This argument suggests that robots, despite being of the fourth-generation, needs to have programs installed, which is contradictory to their independence which suggests that they both reproduce itself and design their successors. The argument also suggests that the robots are submissive to humans. Many AI researchers argues in the opposite direction, that the robots will eventually reach a stage where they, themselves, decide not to be submissive to humans.

Despite providing a detailed description of how the robots will develop and chance the human society within the nearest future, Moravec does not provide an explanation of how this is going to be achieved, which might make his descriptions seem like pure speculations. However, there is no denial that he has been right regarding the first-generation robots and even the second one. There is simply no evidence against that the technology will not develop according to Moravec's prediction.

Furthermore, Moravec (1988, p. 16-17) argues that different ways to approach AI will all have a great chance of succeeding. If human intelligence is what wants to be achieved in robots, then we should let the Darwinian evolution lead the way by letting us use our own intelligence as a tool (Moravec, 1988, p. 16-17). The Darwinian evolution concerns the biological evolution among living creatures. All organisms develop through a natural selection which determines each individual's ability to survive by competing and reproduce (en.wikipedia.org, 14.06.2018). Different approaches to create intelligent machines or studying the field of AI can also be viewed among different researchers. Computer scientists and biologists focus on vision and to build machines with better vision while for example mathematicians and physicists tries to improve sonar and other senses. Mobility such as walking and grabbing something with your hands is the main focus for mechanical engineers. However, the communication between these fields have been poor and that has resulted in various results in the actual robot. Despite the lack of communication, Moravec is optimistic and believes that the first mass produced general-purpose robots will show up by the millenium (Moravec, 1988, p. 22).

Earlier attempts at building robots has demonstrated that robots can execute tasks that seems rather difficult to humans, such as solving mathematical problems or become really good at playing chess. However, they are not performing well when it comes to more basic tasks that humans consider easy, such as learning languages, seeing, hearing, general mobility and

reasoning. This dichotomy is also referred to as “Moravec’s paradox” (Moravec, 1988, p. 2, 4, 6, 9), and consequently, if researchers from different disciplines were to communicate better, this paradox might be solved.

5.2 A Postbiological World: A Robotic Society

As more and more intelligent machines will develop, the world as we know it today will develop into a postbiological world. One of the first robots was built by the British psychologist W. Grey Walter as early as in 1948. He built electronic tortoises that had rotating phototube eyes, miniature radio-tube brains and contact-switchers. These tortoises could wander around, avoid trouble and return to their docks when the power began to end. More advanced robots were developed in the 60’s such as robots that could plug in and recharge themselves from a power outlet in the wall. These robots inspired several imitators decades after such as the robot vacuum cleaner “Roomba”, which are common in modern homes today. Despite the fast and successful development within technology, machines are still struggling with reasoning and visual and auditory perceptions while calculating is not a problem (Moravec, 1998, p. 18-19, 22), hence Moravec’s Paradox.

In order to illustrate the gap between artificial and human skills, Moravec (1998, p. 70) created a “landscape of the human competencies” (see Fig. 2 below). Similar to both Turing and Simon, he argues that computers are universal machines with great potential. The landscape of human competences have high mountain peaks with labels such as “hand-eye-coordination”, “social interaction” and “mobility”. In the lower lands of the foothills, one can see labels like “theorem proving” and “chess playing”. In the lowest parts if the landscape are labels such as “arithmetic” and “rote memorization”. Humans live on the highest mountaintops and it requires great effort to reach the rest of the landscape and only a very few can reach the very lowlands. The water in the landscape symbolizes the advancing computer performance which is slowly flooding the landscape. A half century ago, the water reached as far as the lowlands. The water has currently reached the foothills, which causes us to leave out outposts in the lowest grounds. We feel rather safe on our mountaintops. However, the water will eventually reach the mountaintops if it keeps flooding at the present rate, the mountains will be submerged too only within another half century. A half-century ago, Turing prevented a development in this direction (Moravec, 1998, p. 72; Turing, 1950, p. 442). According to both Turing and Moravec, machines will be able to develop their visceral sense and perform better in a wide range of areas

that will be highly appreciated among humans, also indicating that humans will still be superior to robots. Eventually, when the flooding reaches the highest mountaintops, machines will be able to communicate and act as intelligently as humans.

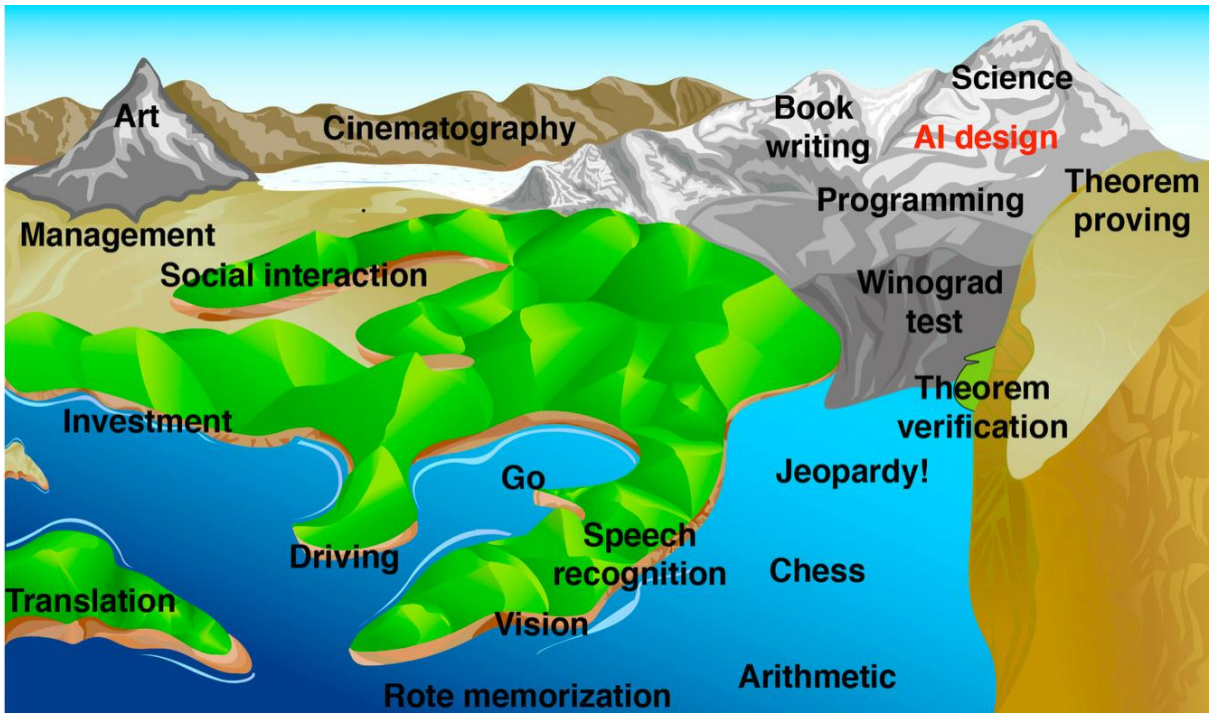

Fig. 2. “Landscape of Human Competence” described by Moravec (in Tegmark, 2017, p. 70).

According to Turing, machines would be able to pass his Turing Test under five minutes by the year of 2000. Moravec continues that by the year of 1998 when his book “Robot: Mere Machine to Transcendent Mind” was released, Turing's prediction was not yet reality. However, Moravec argues that Turing’s prediction might become true within a few decades from now, when more powerful computers are more common. Turing had ideas of how to approach his problem proposed in his text, however, he turned out to mainly focus on arguments against thinking machines. Moravec (1998, p. 73-74) provides a current update on these arguments. I will only focus on the arguments I brought up in the first chapter.

According to *the theological objection*, thinking is something that is linked directly to a soul and since computers have no soul, they are not able to think. Moravec (1998, p. 75-76) argues that computers and machines may not have a soul per se, but if a robot behaves and acts in a way we can interpret as human, we can assign a “personality”, despite if the machine actually can think or not. The question simply becomes irrelevant. However, Moravec emphasizes that not everyone agrees with this argument. For example, social ethicists claim that if a robot is acting in a way we can interpret as human, it's clearly fake and almost like a

parody of human behavior. A machine is programmed to act like an adult, missing the external links in its creation, relationships or history. It has never experienced birth and childhood. This is important for human beings as it comes to shape them as persons.

Some others, such as functional mechanists, claim that a robot cannot represent either thoughts or feelings because of their internal structure. However, despite disagreements, Moravec argues that most people are likely to actually consider robots as persons after a while. This is if the robot is acting like an adult and can interact intelligently. The reason for this, is that it is simply the most effective alternative if robots become more common in society. To conclude, Moravec therefore argues that when machines and robots become accepted in the human society as persons, it might be appropriate to say that robots have souls. This on the other hand, raises new ethical questions, such as if we integrate robots in the human society, should we give them the right to vote? Do they need to pay taxes? In 2017, an uncanny human-like robot was granted a citizenship in Saudi Arabia. The robot is named Sophia and is created by Hanson Robotics. Sophia is the latest development in robotics and has features such as being highly mobile, expressive and can display emotions (Weisberger, 2017). This indicates that Moravec might be right arguing that advanced robots will become more common in the human society and the best strategy for us is to embrace the development.

The argument from consciousness states that "not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, could we agree that machine equals brain". This would mean that humans are the only ones that can experience emotions and be able to communicate them. In philosophy, this is called *solipsism,* which means that the only thing that exist is our own minds. Moravec (1998, p. 82) emphasizes that this philosophy has few followers and that we often believe other people when they open up about their feelings and describes their motivations. However, a robot can be able to describe situations and its possible feelings in that situation, such as "I avoid the stairs because I am afraid of falling down". This is just the robot creating certain beliefs about its own feelings and actions by analyzing its own behavior. The question here is whether the robot does have feelings or if it has just learned how to express its opinions in a convincing way. On the other hand, Moravec emphasizes that we can ask ourselves the same question. Because, what difference does it make if robots have feelings or not? As Haugeland (1985, p. 235-236) argues, emotions and sensations do affect people's goals and actions because it determines movements, reactions and what one might want to achieve. Or as Gadamer (1998, p. 210-211) argues, that one can fully participate in life if one has an understanding of the environment and the objects and

phenomenon's in it. If a robot does not have feelings it might not be able to act like humans after all, unless it gets really good at imitating.

One argument that agrees with the technology optimist's view is that a growing arsenal of artificial organs and body parts are the reason many people are even alive today. Along with the technological development, those body parts become better and better and will eventually become even better than the original, natural parts. Moravec (1988, p. 109) goes as far as suggesting that the human brain one day might be transplanted into a robotic body, an artificial body. This would raise questions of what it means to be a human and is similar to Descartes dualism philosophy, where he makes a distinction between the mind and the body. If the mind and the body really are separate, it would be possible to transplant a human brain into an artificial body. Moravec (1988, p. 116-117) argues in a similar direction, by making a distinction between what he calls pattern-identity and body-identity. Body-identity revolves around the body and assumes that a person is defined by the physical body. Pattern-identity is, on the other hand, the very essence of a person since the pattern and the process is within a person's head and body. Transplanting a human brain into an artificial body would also raise ethical questions such as what would it mean to the human society if we transplanted human brains into artificial bodies? Such questions are difficult to answer but since we already use artificial body parts such as prostheses we have already integrated the artificial body parts into our lives and the more we integrate them, the more natural it might become to consist of more artificial parts than biological parts.

5.3 Artificial General Intelligence (AGI)

Even the most eminent scientists within AI find it difficult defining what intelligence is. There is no single definition of the concept of intelligence but rather plenty which involves rationality, logic, planning, creativity, self-awareness, learning ability and emotional insight, which we have seen in the previous chapters. According to The Oxford Dictionary (oxforddictionaries.com, 30.05.2018), the word 'intelligence' means; "The ability to acquire and apply knowledge and skills". This is also the definition Tegmark (2017, p. 66-68) agrees with. However, he emphasizes that this is still a very broad definition. Tegmark agrees with both Haugeland and Weizenbaum that intelligence is a definition which is difficult to measure and there are also different forms of intelligence. The figure below illustrates how the skills differ between different forms of life. What is most significant is that life beyond the human

level might be very good at memorizing or heavy lifting but fail to conduct a conversation with a human in a convincing way. A human on the other hand, might have moderate skills in both Chess playing and text translation.

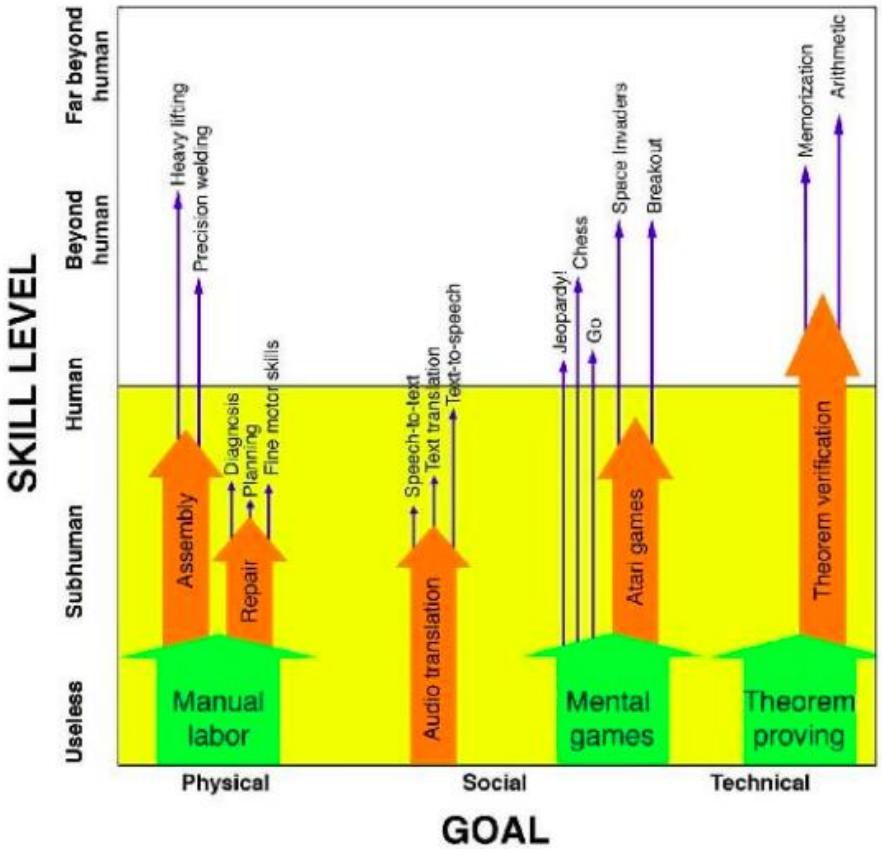

Fig. 3. Goal/Skill Level (Tegmark, 2017, p. 68).

Tegmark (2017, p. 68) therefore suggests that the intelligence within current AI science is aiming at a *general* intelligence and not a *super* intelligence which some argue. By that he means that the intelligence level would be equivalent to human intelligence. Superintelligence is a higher level of intelligence. Humans usually rank tasks according to the level of difficulty but that is somehow misleading in regards of how difficult the task is to perform by a computer. For example, it is much harder for us humans to multiply 410 345 x 320 673 than to recognize a friend on a photography. However, for a computer, it is the other way around, which refers to "Moravec's paradox" (Tegmark, 2017, p. 70; wikipedia.org, 30.05.2018).

As the previous chapters has demonstrated, the Turing Test might have been too simple for the concept of intelligence or not equivalent in order to measure intelligence. According to Stuart Russell and Peter Norvig (2010, p. 2), in order for a computer to pass the Turing Test, it needs to possess several capabilities, including natural language processing, reasoning,

knowledge representation to store new information and machine learning to be able to adapt to new circumstances. As Weizenbaum, Haugeland and Tegmark have argued, intelligence includes more than just performing well in one area. Therefore, Turing's test can be criticized for not measuring intelligence but rather a person's gratitude. A system is rational if it can perform the "right thing" given what it knows (Russell & Norvig, 2010, p. 1), which on the other hand might indicate that given how the game was setup to be played, it was at that time regarded as an eligible way of measuring intelligence in machines. Whether a machine can act like it is thinking or if it is actually thinking is what is referred to as weak and strong AI. Weak AI means that the intelligent machine only simulates thinking processes without producing thoughts itself. The hypothesis that machines actually does think, is on the other hand strong AI. As long as the programs and machines works, most researchers do not care about the strong AI-hypothesis (Russell & Norvig, 2010, p. 1020).

5.4 Deep Learning

To calculate an equation is to convert information from one memory state to another. Mathematics call this process a *function*. Current computers can calculate fast due to their parallel distributed processing systems. If a calculation can be divided into processes which can be performed in parallel, the computer can conduct several calculations at the same time but how can machines ever learn anything? A calculator cannot learn, it will continue to calculate but it will never remember and *learn* the calculations per se. It is simply just performing a mathematical function. Very simply put, when humans created calculators and chess playing machines, we did the programming. In order for a machine to learn, it needs to be able to do the programming itself. It needs to memorize, calculate and store the new knowledge in order to learn (Tegmark, 2017, p. 80, 91-93).

The human brain works the same way but faster and more efficient. John Hopfield (in Tegmark, 2017, p. 93; Hopfield, 1982, p. 2554) illustrates how interconnected neurons can learn in an analogous manner, if you repeatedly put the network in certain states, it will gradually learn these states and return to them from any nearby state. Neural networks have influenced both biological and Artificial Intelligence but also machine learning, which revolves around the study of algorithms that are improved through experience. Large amounts of data, also referred to as Big Data, enables for mapping out otherwise unrecognized patterns, behaviors, trends, identities and practical knowledge with the help of algorithms (Allen, 2016, p. 1; Zuboff, 2015,

p. 78). Google Translate is an example of learning algorithms which constantly collects and improves data. At first when Google Translate was introduced in 2006, it could only translate a few words and sentences but nothing too complex. In 2016 Google announced that Google Translate operates with a neural machine translation, based on neural network systems which means that instead of translating single words that form sentences, it now started to translate whole sentences and paragraphs. This improved the language translation tremendously although it still suffers from some grammar errors (Lewis-Kraus, 2016). Most importantly, the new way Google Translate is set up enables for the system to constantly make new neural connections and continuously upgrade the grammar spectrum.

One other aspect of learning is deep learning. Deep learning is a branch of machine learning and it is also called deep structured learning and hierarchical learning. Deep learning has enabled self-driving cars and digital assistance such as Siri or Google Now. In addition to deep learning, another branch of learning is deep reinforcement learning, which is influenced by psychology behaviorism. The idea is that positive reward increases one's tendency to do the same thing again (Tegmark, 2017, p. 103, 113). In early 2016, Google's computer program DeepMinds AI-system AlphaGo defeated the world champion Lee Sedol. The main idea with AlphaGo is combining the intuitive ability of deep learning with the logical ability of Haugeland's GOFAI. The combination of deep learning and logic in traditional GOFAI resulted in a strong and creative computer program. AlphaGo winning over Lee Sedol marked a milestone within AI research about learning abilities and also in AlphaGo's history. Playing AlphaGo requires a combination of both strategic moves, intuition and logical thinking. Tegmark further argues that this is also important to remember and consider in the development of AI. Combining logic and deep learning might be very beneficial in areas such as investment strategies, political and military strategies (Tegmark, 2017, p. 117-118; youtube.com, 31.05.2018).

A fully automated self-driving car is another example of a systems which operates with algorithms. If a translation program fails to recognize a sentence, it wouldn't be disastrous but with a self-driving car we need more assurance and more accurate systems. Such car needs to be able to master highly complex situations such as getting to the correct destination on time, dealing with a variety of roads, traffic, following the traffic rules and so on, the list is long and the possible combinations of situations the car might face are endless. Russell and Norvig (2010, p. 693) argues that there are several reasons of why an intelligent agent needs to be able to learn by itself instead of being programmed. First, programmers cannot possibly anticipate all opportunities or situations the agent might come to face. Second, as time goes by, conditions

change. An intelligent agent need therefore be able to adopt and adjust. Sometimes it is also about the fact that the programmers do not know the solutions themselves. Google Translate is an example of a program that needs to learn by itself in order to constantly improve and manage complex grammar in different languages.

5.5 Natural Language Processing 2.0

That brings me back to natural language processing, where most of the advances within AI is currently made. Thanks to the latest advances in in-depth learning in terms of speech-to-text and text-to-speech conversion, users can now also speak with their smartphone in one language and then listen to the translated result. Tegmark (2018, p. 120) argues that success within AI and translation and natural language processing will be of great importance for human kind since our language is so significant for us. The better AI can manage languages, the better the technology can manage for example mail conversations and participate in spoken conversations. From an outside perspective, this might give the impression of a human thought process. Therefore, he continues, AI might actually pass the Turing Test, where a machine must be able to converse well enough in writing and speech to trick a person to believe that it is human.

Despite all success within natural language processing, a machine still needs to become better at managing languages. For example, a machine can learn the difference between "queen" and "king" but it still doesn't know the meaning of the words and the context, it lacks understanding. When humans translate we constantly use our knowledge about real life and situations in order to get an understanding for what the pronoun refers to. When it comes to understanding what the pronouns refers to, AI is far behind and performs very poorly. However, since most current translating programs uses combined methods (in-depth learning and logic GOFAI), it is constantly improving (Tegmark, 2017, p. 122).

Edward Sapir and Benjamin Whorf suggested in the middle of the twentieth century that language to a large degree influences our understanding of the world. Whorf described it as follows;

> We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way ─ an agreement that holds throughout our speech community

> and is codified in the patterns of our language (in Russell & Norvig, 2010, p. 278; Tredinnick, 2006, p. 139).

This hypothesis has been both criticized and praised throughout the years. According to Chomsky (1972, p. 4), a person has an inherent system for processing and understanding language but this will only be developed if the person is exposed to it. Based on this argument, it is possible to argue that people will develop different patterns in their societies and cultures depending on where they are from. Robert Callan (2003, p. 248-249), a technical consultant specialized in AI and pattern recognition, makes a similar argument as Chomsky regarding meaning of language. Callan argues that we understand and interpret sentences by putting together words combined with our knowledge and experiences. The way in which words are put together to form sentences is called *syntactic knowledge*. The words can be put together in multiple ways and each single word belongs to a certain class which is dependent on how the word is used. Clauses are the description of groups or words which contains a subject, a verb and often an object. Grammar is a framing system for how the words can be put together, although people are able to understand many sentences even if they are not correct grammatically. In addition to that, we all have something we call common-sense, which refers to our ability to perceive and understand our surroundings. It is knowledge we take for granted. In order to understand natural language processing, one needs to have knowledge about the structure of the language, how sentences can interact to become meaningful and how meaning is connected to sentences (Callan, 2003, p. 349-350, 364). This argument connects to the communication model by Shannon and Weaver, Gadamer's hermeneutics theory and Chomsky's language theory, implying that these grammatical rules are part of why natural language processing is so complicated, even for humans.

5.6 The Potentials of Artificial Intelligence

The potentials of AI are even greater but the more powerful AI becomes, the more important it becomes that it is reliable and execute what we tell it to do, similar to what Good stated in the 60's. Since back in history, we have relied on the well-proven method: learning from mistakes. We invented cars, then we crashed, so we invented the seat belt. Same thing with fire, we invented it, realized how bad it can get or what uncontrolled fire can cause, so we invented fire-extinguishers and the fire alarm. Tegmark (2017, p. 124-125) suggests that we

need to be proactive, rather than learning from mistakes when it comes to the development of AI. The reason for that, he argues, is that we do not really know the full capacity of AI. Computers and technology have always malfunctioned and had different problems but AI is different because it has entered our daily lives on a level no other technology has done before and we also trust it regarding managing of information, medical care, economics and so on.

Nearly all fatal accidents in traffic are caused by the human factor and with the help of AI technology and self-driving cars the high percentage can be reduced by 90 percent (Tegmark, 2017, p. 131-135). The biggest challenge for AI within transports is currently control. A human operator's ability to monitor the system and, if necessary, intervene is still something AI lacks. Another area where AI have been very influential is within health care. In a study conducted 2016 by Stanford (Kubota, 2016), AI were able to diagnose lung cancer using microscopic images even better than human pathologists. Tegmark therefore argues if we, with the help of machine learning, can detect the connection between genes, diseases and treatment responses, it could possibly revolutionize individualized medication. In addition to that, Tegmark continues, robots have the potential to become more thorough and reliable surgeons than humans, even without using advanced AI technology. A variety of different successful surgeries has been carried out during the recent years which has enabled more precision and smaller incisions that led to decreased blood loss, less pain and shorter healing time. According to an American study published 2013 in The Scientific American, an average of 100 000 people die from human errors at hospitals yearly, in America only (www.scientificamerican.com, 03.06.2018). The potentials with AI are endless and as Tegmark argues, it would revolutionize individualized medication but at the same time, it raise ethical questions such as who would be responsible if a robot accidently do harm to a human during a surgery? This is something we do not have a clear answer to yet but need to consider. Just as human beings can get sick and suffer from diseases, so can the intelligent machines. Different forms of computer viruses have existed since decades and are expected to still be around for the intelligent machines. The viruses will evolve and develop similar to a biological virus. A virus, whether digital or biological, reproduce and copy itself very fast and therefore causes damage to the cells (Tegmark, 2017, p. 136, 156; Moravec, 1988, p. 125-126, 128). Moravec (1998, p. 97) emphasizes that awareness of information security and integrity is more important than ever since maliciously programmed robots will be a possible danger. However, if intelligent robots reach the level of intelligence that they start upgrading themselves they might also be able to come up with a cure for potential viruses that might harm them.

The more intelligent the machines become, the more important it becomes for us to make sure that we all share the same overall goals. If *intelligence* is defined as the *ability to acquire and apply knowledge and skills* (oxforddictionaries.com, 30.05.2018), an intelligent machine will therefore become better at achieving this than humans, which in theory means that the machines will be superior to humans. Therefore, Tegmark (2017, p. 346) argues that the greatest risk with AI is not that it might become evil, as some might argue but that it will gain a lot of knowledge. However, Tegmark emphasizes that it is difficult to even define whose goals we are talking about and what are those goals? One possible answer to what those goals are might be our definition of a society, we have set rules in order for all humans to live together.

We need to keep that in mind that by creating intelligent machines, we are also changing the way we perceive the world and ourselves in it. We are currently the ones in charge of the societies and what the societies should look like. Where we go from here regarding the development of AI must be very well considered in order to get the society we want. Some scientists even go as far that they argue that in a few decades, AI will have greater influence than climate changes, war, terrorism, poverty. However, AI will also be able to provide solutions to some of these problems (Tegmark, 2017, p. 47-48).

5.7 Future Outlook: A Technological Singularity Awaits (or does it?)

It has become clear that technology influences people in significant ways, and it will continue to do so; it is comparable to the emergence of human life on earth. Technological singularity, implied by the computer scientist Vernor Vinge (1993, p. 365-366), means the invention of an artificial superintelligence that will far surpass the human intelligence which ultimately will have unfathomable results for the human civilization. It will set new rules for reality and discard old ones. This might sound frightening or like science fiction but Vinge (1993, p. 366) argues that these superintelligent entities might not occur at all. Despite building machines as powerful as humans they might not act the same way as humans. However, a technological singularity will happen if it can, he continues. What might be most frightening with a technological singularity is that it could possibly extinct the human race. The reason such fear exists is because superintelligent entities might not share the same goals and values as humans, as Tegmark pointed out. We treat ourselves as the most superior race in the world and if a more intelligent race occurs, there are no reasons for them not to treat us the same way

(Vinge, 1993, p. 369). Similar to Vinges argument, the computer scientist and futurist Ray Kurzweil (2005, p. 204-205) predicts a future where humans make way for robots.

Modern debate about AI tend to evolve around whether AI will ever outperform humans and if they will, the following questions are when. The answers to these questions vary, there are scientists arguing that superintelligent AI will be developed within the next 50 years from now and there are scientists arguing that such technology will never exist. On the other hand, according to Tegmark there are no evidence that this will actually happen, which means that a technological singularity might be nothing but mere speculations. The current media-debate about AI is more controversial than what the reality is. According to Tegmark (2017, p. 54, 57), fear is after all something that catches people's attention.

5.8 Summary of the Evolutionary Paradigm

The Evolutionary paradigm, has both similarities and differs from the dichotomy between the Traditional and the Neural Networks paradigm. This paradigm might be regarded as a paradigm with connections to science fiction due to its concern with the development of AI which includes robotics, a postbiological world and superintelligent life. The invention and development of intelligent technology forces us to reflect on what it means to be a human and to be alive. Scientists suggests that we need to be proactive, rather than learning from mistakes when it comes to the development of AI. The reason is simply because we do not really know the full capacity of AI.

6.0 Concluding Discussion

This thesis has examined the differences and similarities in the paradigmatic development of the dominant paradigms Traditional, Neural Networks, and the Evolutionary within Artificial Intelligence in the time period from the 50's to modern age. A comparative analysis method was applied, which allowed for a deep analysis of both differences and similarities which also helped emphasize related philosophical, cultural, social and information processing influences. The analysis was conducted involving a hermeneutics approach which has resulted in the suggestion that the different paradigms has connections to philosophical

theories such as rationalism and empiricism. In this section, I will highlight some of the most distinctive differences and similarities.

6.1 Consciousness Works Rationally

The Traditional paradigm revolves around information processing, mathematics, logic, rational reasoning and problem-solving. Pioneers within computer science, mathematics, psychology and cognitive science have contributed to the research within AI. What has become clear is that this paradigm is characterized by great enthusiasm regarding problem solving, but they were less successful. Turing introduced the Turing Test, which has become a landmark within AI research, but no one has yet succeeded in the test. Simon introduced the General Problem Solver, which did not succeed either. Furthermore, Weizenbaum criticized the concept of "intelligence" as a too broad an ambiguous concept to measure along an IQ scale. In addition to that, whatever we learn about intelligence in computers will tell us about the way we look at ourselves since we need to reflect on ambiguous concepts such as what intelligence and thinking means.

Good Old Fashioned Artificial Intelligence, introduced by Haugeland, marks a traditional view of AI, concerning symbolic representations. It assumes that by manipulating symbols, many aspects of intelligence can be achieved, a statement which implies that machine works rationally and logically. What has become clear is that theories within language processing are central, not only within the Traditional paradigm but within AI in general. The linguist Chomsky argues that humans have an innate system for understanding language and in order to give meaning to words. By combining the innate ideas and knowledge about the words and from external inputs, one can understand the meaning of a text. This is important to understand if we want to develop better and more intelligent technology.

Since mathematics, reasoning and logic are central terms within the Traditional paradigm, I have argued that this paradigm has similarities with René Descartes philosophical view rationalism and cartesianism, where reasoning is about rational and logical decision making. The Descartes was a dualist, believing that the soul is separated from the body he argued that it is not rational to rely on the senses, because one might be dreaming. It is therefore possible to say that the hypothesis is valid.

6.2 Consciousness Works Associatively

Contrary to the Traditional paradigm and its similarities with rationalism, the Neural Networks paradigm has been argued to have similarities with empiricism, which focus on knowledge derived from sensory and perceptual experiences. Humans does not have innate ideas of the world when they are born, their minds are born as "tabula rasa", as a blank tablet for experiences to leave marks on, which is contradictory to the rationalistic view. Therefore, the hypothesis can be considered as valid. What has become clear is that studying the brain and how it function can reveal important keys to how we can develop technology with similar characteristics. One clue is neurons. Complex coordinations can only be performed by neurons and this is where the intelligence of the brain emerges from. Most significant is that neurons learn, they adjust and updates. After being repeated several times, the network slowly adjusts and learns the desired behavior. In fact, the desired behavior will be provided in ninety percent of the time, which makes neural networks so efficient and powerful.

One significant part of being human is that we have a consciousness and we have self-knowledge and can experience *qualia*, which is a philosophical term for subjective experiences such as the taste of chocolate. Some argue that this is the very essence of human beings since these are qualities that determines our goals, interests, our intellect and personal identities. Building intelligent machines raise such questions and forces us to think about what it means to be a human. Considering the smaller parts in order to create a bigger picture is what characterizes a bottom-up approach to AI, in contrary to the Traditional paradigm that is characterized by a top-down approach.

6.3 A Postbiological World Awaits

While the two previous paradigms reflects a dichotomy between philosophical rationalism and empiricism, it has become clear that this dichotomy does not fully apply on the Evolutionary paradigm, although some similarities with both philosophical views can be seen. It is therefore possible to say that the hypothesis is applicable here as well. For example, Moravec argues that it might be possible to transplant human brains into artificial bodies. Despite raising ethical questions, this argument is similar to Descartes dualism philosophy, characterized by a distinction between the mind and the body. This paradigm is characterized by robots, artificial progeny's that will enforce a shift from a biology driven world towards

postbiological world. The technology development forces us to reflect on what it means to be a human. Instead of focusing on intelligence in machines, which has been illustrated is an ambiguous concept and much of the discussion revolves around the self and consciousness, we might instead focus on building *conscious* machines. Many scientists questions whether robots or intelligent technology can experience emotions or not. However, intelligent robots can learn to express opinions and emotions in a convincing way, which means the question becomes irrelevant. Since we do not fully know the capacity of Artificial Intelligence, we should be proactive rather than learning by mistakes before we develop the technology further, that is, if we wish to keep our superior place in the world.

List of References

Allen, A. (2016). Protecting one's own privacy in a big data economy. *Harvard Law Review*. 130(2), 71-78.

Artificial Intelligence. (n.d.). In *Oxford Dictionaries*. Retrieved from https://en.oxforddictionaries.com/definition/artificial_intelligence

Berg-Schlosser, D., De Meur, G., Rihoux, B. & C. Ragin., C.. (2008). Qualitative Comparative Analysis (QCA) as an Approach. In Rihoux, B. & C. R., C. (Eds.), *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques* (1-18). Los Angeles: SAGE Publications Inc.

Berg-Schlosser, D. & De Meur, G. (2008). Comparative Research Design. In Rihoux, B. & C. Ragin, C. (Eds.), *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques* (19-32). Los Angeles: SAGE Publications Inc.

Bermúdez, J. L. (2014). *Cognitive Science. An Introduction to the Science of the Mind. Second ed.* New York: Cambridge University Press.

Callan, R. (2003). *Artificial Intelligence*. Basingstoke: Palgrave Macmillan.

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

Chomsky, N. (1972). *Language and Mind*. *Enlarged edition*. New York: Harcourt Brace Jovanovich.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Churchland, P. S. (1985). *Neurophilosophy, Toward a Unified Science of the Mind-Brain*. Cambridge: MIT Press.

Churchland, P. M. (1989). *A Neurocomputational Perspective. The Nature of Mind and the Structure of Science*. Cambridge, Massachusetts, MIT Press.

Churchland, P. S. (2002). *Brain-Wise: Studies in Neurophilosophy*. Cambridge: Massachusetts, MIT Press.

Consciousness. (n.d.). In *Oxford Dictionaries*. Retrieved from https://en.oxforddictionaries.com/definition/consciousness

Cooper, D. E. (1999). David Hume, An Enquiry Concerning Human Understanding, Section 12. In *Epistemology. The Classic Readings* (134-147). Oxford, UK: Blackwell Publishers Inc.

Cooper, D. E. (1999). René Descartes, Meditations on First Philosophy, I-III and 'Objections and Replies' (Selections). In *Epistemology. The Classic Readings* (97-116). Oxford, UK: Blackwell Publishers Inc.

Damasio, A. (2012). *Self Comes to Mind. Constructing the Conscious Brain*. Vintage: London.

Darwinism. (n.d.). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Darwinism

Denscombe, M. (1998). *Forskningshandboken- för småskaliga forskningsprojekt inom samhällsvetenskaperna*. Lund: Studentlitteratur AB.

Dreyfus, H. (1997). *What Computers Still Can't Do. Fifth ed*. Massachusetts: MIT Press.

Durham Peters, J. (2000). Conclusion: A Squeeze of the Hand. In *Speaking Into The Air. A History of The Idea of Communication* (263-272). Chicago: University of Chicago Press.

Fish, S. (1980). *Is There a Text in This Class? The Authority of Interpretive Communities*. Cambridge, Massachusetts: Harvard University Press.

Frisch, P. A. (2001). U.S. History: Primary and Secondary Sources. *College & Research Libraries News, 62,* 991–994.

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers, 6*, 31-88.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, Massachusetts: MIT Press.

Hjørland, B. & Hartel, J. (2003). Afterword: Ontological, Epistemological and Sociological Dimensions of Domains. *Knowledge Organization*, 30(¾), 239-245.

Hjørland, B. (2013). Theories of Knowledge Organization - Theories of Knowledge. *Knowledge Organization*, 40(3), 169-181.

Hopfield, J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *National Academy of Sciences of the United States of America*, 79(8), 2554-2558.

Intelligence. (n.d.). In *Oxford Dictionaries*. Retrieved from https://en.oxforddictionaries.com/definition/intelligence

Kubota, T. (2016 November, 2018). Stanford algorithm can diagnose pneumonia better than radiologists. *Stanford News.* Retrieved from https://news.stanford.edu/2017/11/15/algorithm-outperforms-radiologists-diagnosing-pneumonia/

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kurzweil, R. (2006). *The Singularity Is Near: When Humans Transcend Biology*. London: Viking, Penguin Group.

Larsson, L. (2000). Personliga intervjuer. In *Metoder i kommunikationsvetenskap* (49-77). Lund: Studentlitteratur AB.

Lewis-Kraus, G. (2016, December 14). The Great A.I. Awakening How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself. *The New York Times Magazine*. Retrieved from https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Lombard, E. 2010. Primary and Secondary Sources. *Journal of Academic Librarianship*, 36(3), 250-254.

Marshall, A. (2013 September 20). How Many Die From Medical Mistakes in US Hospitals. *The Scientific American*. Retrieved from https://www.scientificamerican.com/article/how-many-die-from-medical-mistakes-in-us-hospitals/

Minsky, M. (1988). *The Society of Mind.* New York: Simon & Schuster, Inc.

Moravec, H. (1988). *Robot: Mere Machine to Transcendent Mind*. Cambridge, Massachusetts: Harvard University Press.

Moravec, H. (1990). *Mind Children. The Future of Robot and Human Intelligence*. Cambridge, Massachusetts: Harvard University Press.

Moravec, H. (1998). When will computer hardware match the human brain? *Journal of Evolution and Technology*, *1*, 1-12.

Pasquale, F. (2015). Introduction – The Need to Know. In *The Black Box Society. The Secret Algorithms that Control Money and Information* (1-18). Cambridge, Massachusetts: Harvard University Press.

Reasoning. (n.d.). In *Oxford Dictionaries*. Retrieved from https://en.oxforddictionaries.com/definition/reasoning

Rihoux, B. & C. Ragin, C. (2008). Introduction. In Rihoux, B. & C. Ragin, C. (Eds.), *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques* (xvii-xxv). Los Angeles: SAGE Publications Inc.

Rihoux, B. & Yamasaki, S. (2008). A Commented Review of Applications. In Rihoux, B. & C. Ragin, C. (Eds.). *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques* (123-146). Los Angeles: SAGE Publications Inc.

Rumelhart, D. & McClelland, J. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Cambridge, Massachusetts: MIT Press.

Russell, S. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. *Third ed.* New Jersey: Pearsons.

Ryen, A. (2011). *Kvalitativ Intervju- från vetenskapsteori till fältstudier*. Malmö: Liber AB.

Shannon, C. E. & Weaver, W. (1963). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

Simon, H. A. (1996). *The Science of the Artificial*. *Third edition*. Cambridge, Massachusetts: MIT Press.

Tegmark, M. (2017). *Life 3.0: Being Human In the Age of Artificial Intelligence*. Stockholm: Volante.

Thinking. (n.d.). In *Oxford Dictionaries*. Retrieved from https://en.oxforddictionaries.com/definition/thinking

Thiselton, A. C. (2009). *Hermeneutics: an introduction*. Grand Rapids.

Tredinnick, L. (2006). *Digital Information Contexts. Theoretical Approaches to Understanding Digital Information.* Elsevier.

Trost, J. (1997). *Kvalitativa intervjuer, Second ed*. Lund: Studentlitteratur AB.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *236*, 433-460.

Vinge, V. (1993). Technological Singularity. *VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute*, 1-4.

Wallace, M. & Wray, A. (2011). *Critical Reading and Writing for Postgraduates*. London: SAGE.

Weisberger, M. (2017 October 30). Lifelike 'Sophia' Robot Granted Citizenship to Saudi Arabia. *Livescience.com*. Retrieved from https://www.livescience.com/60815-saudi-arabia-citizen-robot.html

Weizenbaum, J. (1976). *Computer Power and Human Reason*. San Francisco: W. H. Freeman and Company.

Wray, K. B. (2011). What Makes Kuhn's Epistemology a Social Epistemology? In *Kuhn's Evolutionary Social Epistemology* (170-185). Cambridge, UK: Cambridge University Press.

YouTube: Move 37!! Lee Sedol vs AlphaGo Match 2 (2016 March 12). Retrieved from https://www.youtube.com/watch?v=JNrXgpSEEIE

Zuboff, Shoshana (2015). Big other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, 30(1), 75-89.