GÖTEBORGS UNIVERSITET

INSTITUTIONEN FÖR FILOSOFI, LINGVISTIK VETENSKAPSTEORI

ARTIFICAL INTELLIGENCE

The implications of current technological developments for Harry Collins epistemological theory and his critique of the possibilities of a general AI.

Marita Isaksson

Uppsats/Examensarbete:	15 hp
Program och/eller kurs:	Magister vetenskapsteori
Nivå:	Avancerad nivå
Termin/år:	Vt/2018
Handledare:	Johan Söderberg
Examinator:	Dick Kasperowski

Abstract

Social constructivism and Harry Collins epistemological theory put focus on the sociology of scientific knowledge, scientific practice and the nature of expertise. From his theory, inspired by Ludwig Wittgenstein's 'forms of life' and David Bloor's strong programme, Collins has from the nineties until today argued against the possibilities of creating a general artificial intelligence, which can be compared with any cognitive functions that a human accommodates. Collins epistemological theory withholds aspects of knowledge basics, transfer and communication that are said to be unique and/or can only be performed by humans, which together with the problem of socialization constitute the demarcation line between a human and what can be achieved by an artificial intelligence, AI.

Two areas of AI-research that in recent years have reported progress are autonomous vehicles and games. Autonomous vehicles are supposed to, in the near future, be able to handle different types of environments, interpret conflicting signals and 'make' ethical decisions. Game-playing AIs are said to be able to create their own set of rules and perform a basic communication. If these advancements are correct the AIs seem to challenge Collins critique of a potential general AI by exhibiting social sensibility, being able to adapt to different social and/or cultural settings and improvise when needed plus developing a (limited) language in order to communicate.

In the paper I clarify what is at stake in the AI-debate, and intend to empirical test Harry Collins' theories, and if his epistemological theory is affected by the reported and anticipated progress in the AI-research. My findings are that Collins' epistemological theory is severely affected due to him, in his argumentation, slipping towards naturalism in order to construct a solid demarcation line, safeguarding against moving goalposts and a resistance to acknowledge the possibility of an AI being able to absorb anything equivalent to human social experience.

Uppsats/Examensarbete:	15 hp
Program och/eller kurs:	Magister vetenskapsteori
Nivå:	Avancerad nivå
Termin/år:	Vt 2018
Handledare:	Johan Söderberg
Examinator:	Dick Kasperowski
Nyckelord:	Harry Collins, artificial intelligence, epistemology

Syfte:	To clarify what is at stake in the AI-debate, and how, and if the demarcation line Harry Collins drew, between human and AI, is unchanged. This in order to empirically test if his epistemological thesis still is valid.
Teori:	Discourse study.
Metod:	Philosophical investigation based on Harry Collins' critique of the possibility of a general AI in *Artificial Experts Social Knowledge and Intelligent Machines*, and *Tacit and Explicit Knowledge*.
Resultat:	Harry Collins' epistemological theory is severely affected due to him, in his argumentation, slipping towards naturalism in order to construct a solid demarcation line, safeguarding against moving goalposts and a resistance to acknowledge the possibility of an AI being able to absorb anything equivalent to human social experience.

Table of contents

Introduction

The past two years' reports in media, from involved companies, and research-labs have emphasised advancements of AI-research in the areas autonomous vehicles and game-playing AIs. To mention two examples; In the beginning of 2016 AlphaGo, a version of Google's AI DeepMind, won a game of Go for the first time, a game regarded by many impossible for an AI to conquer; During 2017 test-areas for autonomous vehicles started popping up in many countries around the globe, e.g. DriveMe (Test Site Sweden 2018) in Gothenburg, Sweden, where 100 self-driving cars are planned to drive on public roads during 2018. Technical innovations such as these urge us to ask questions in terms of where the use of AI-technology is heading, and what impact it will have in a future where AIs might become more intelligent[1] than us?

With the possibility of a human-level AI moral and ethical issues needs to be taken into account, such as liability and safety of human beings. Safe-making the human race, and way of living, are placed on top of the agenda in many areas of society at present, e.g. the European Parliament's Committee on Legal Affairs' report with recommendations to the Commission on Civil Law Rules on Robotics to present EU-wide rules on the topic of robotics and AI (European Parliament 2017) and the report *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* co-written by 26 authors from academia, civil society, and industry (Brundage et al. 2018). There are several practical aspects that need to be addressed in a future inhabited by technology that might be smarter, and better, than a human in completing a work-task. These are important to handle, think through thoroughly, and make sure rules and laws are explicit enough to cover a multitude of possible scenarios. But it need to be noted that questions like these assume that artificial intelligence is possible, at least at such a high level of performance that it might present risks to the safety and wellbeing of humans, and this is not something scientists agree upon ever will happen.

In this paper the fundamental question that will be investigated is if an artificial intelligence comparable to the human intelligence is possible. The question will be addressed by placing Harry Collins' critique of the AI-project in one ring-corner and the research-status of the current AI-project(s) in the other. Collins is chosen due to his position in the epistemological chicken debate, where he assigns uniqueness to the human and social explanations. In the debate Collins was criticized by Bruno Latour and Michel Callon, advocates of the Actor Network Theory, ANT (Latour& Callon 1992). Collins claims on the possibility of the AI-project presuppose his epistemological theories and empirical investigations of contemporary AI-research allow new entry points to this debate over human uniqueness. Here lay my interest, and reason for narrowing down to investigate Collins'

[1] In the case of a super intelligent AI it is already assumed that a human level AI is possible. Scientists who believe that an AI with the same level of intelligence as a human is possible usually also withhold that a super intelligent AI is achievable.

critique of, and disinclination to, the possibility of a human level AI. If Collins is wrong in his statement that a human level AI is not possible it will have consequences for and/or severely damage his epistemological claims.

As theory for the paper I aim at applying a (critical) discourse analysis approach by analysing the writings of Harry Collins together with the presented result of AI-research, and placing Collins' arguments against the possibilities of a general AI, and his epistemological theory, in the practical context of todays research in AI, conducted within the areas of autonomous cars and game-playing AIs. The discourse analysis is inspired by the content analysis presented by Simone Natale and Andrea Ballatore (2017), which in many aspects follows the foundation of critical discourse analysis as presented by Norman Fairclough in *Critical Discourse Analysis: The Critical Study of Language* (1995). My analysis does not follow Fairclough's three critical discourse analysis' dimensions; 'text', 'discursive practice' and 'social practice'. Instead I choose to use the framework presented by Marianne Jørgensen and Louise Phillips (2002) by which it is suggested that 'the researcher can delineate the different discourses, focusing on the following: the aspects of the world to which the discourses ascribe meaning; the particular ways in which each of the discourses ascribes meaning; the points on which there is an open struggle between different representations; and any understandings naturalized in all of the discourses as common-sense.' (Jørgensen & Philips 2002, pp. 145) Hereby I follow the route suggested by Jørgensen and Phillips of placing focus on the interplay between the discourses, as presented by Collins and the AI-research, to sift out the (social) consequences if one of the two discourses was accepted instead of the other.

Collins was one of the main critics of the possibilities of AI during the nineties with the book *Artificial Experts: Social Knowledge and Intelligent Machines*, which in 2010 was followed by his updated critique in *Tacit and Explicit Knowledge*. Due to Collins' objective my focus will be put on strong/general artificial intelligence, which simplified is a human level AI, one that can be compared with any cognitive functions that a human being may have.

Two influential, and the most prominent, philosophers in the epistemological critique of the AI-project, together with Collins, during the nineties and until today are John R. Searle and Hubert Dreyfus. I will present their theories, and also a brief list of arguments focusing on the difference between (human) brains and computers, in the section 'Critique of the AI-project'. This is done in order to clarify Collins' position and situate his critique in the AI-discussions, and also how he relate his critique to Searle's and Dreyfus'.

With the paper I am going to clarify what is at stake in the AI-debate, and how, and if the demarcation line Collins drew, between human and AI, has evaporated or might have been moved. This in order to

highlight and sift out the possible remains of Collins critique and how (if) his epistemological thesis still is valid.

My hypothesis is that the reported progress in AI-research, the last couple of years, challenge Collins' arguments against a possible general AI by exhibiting social sensibility, being able to adapt to different social and/or cultural setting and improvise when needed plus developing a (limited) language in order to communicate.

I will take a closer look at primarily two areas of AI-research, autonomous vehicles and game-playing AIs, and analyse if the way these function and perform refute the uniqueness of humans and the need for socialisation as stated by Collins. Autonomous vehicles are chosen on the basis of the promise that they will be able to function in all environments, overcome obstacles such as harsh weather, conflicting signals, and also to be able to make complicated decisions in situations where an ethical dilemma is present, e.g. to choose between driving into a group of people or crashing into one person.[2] Game-playing AIs are chosen since they are presented as being able to create their own set of rules and perform basic communication.

[2] The subject of machine ethics is the topic of the website Moral Machine, which provide a platform for ' 1) building a crowd-sourced picture of human opinion on how machines should make decisions when faced with moral dilemmas, and 2) crowd-sourcing assembly and discussion of potential scenarios of moral consequence.' (Moral Machine 2018)

Background

Driven by fantasies, possibilities and promises the project of creating artificial intelligence has moved from more theoretical to more practical oriented work in the past decade. What started out with experimental machines a few centuries ago, where it became possible to test 'hypotheses about the mechanisms of thought and intelligent behaviour and thereby demonstrate mechanisms that formerly existed only as theoretical possibilities' (Buchanan 2005, pp. 53), has evolved into a multi-billion industry where the focus seems to be directed towards 'production' of new and more advanced products.

To understand what is at stake, and the origin of thoughts, a brief historical flashback is needed. During the 1700th century Gottfried Wilhelm von Leibniz, Blaise Pascal and Wilhelm Schickard designed calculating machines that mechanized arithmetic. Leibniz has been described as a philosopher who 'seemed to see the possibility of mechanical reasoning devices using rules of logic to settle disputes' (Buchanan 2005, pp. 53).

The innovations inspired Charles Babbage to start working on a method that could be used mechanically and would remove the human error factor. Babbage invented the 'Difference Engine', completed in 1832, to compile mathematical tables and in 1856 he presented the 'Analytical Engine', which is regarded as the world's first general-purpose computer.[3]

In the early 20th century innovations in electronics and computers gave rise to new visions of possibilities. Contemporary movements in society and science influenced research in computer science, and computers were regarded as experimental devices to understand the nature of intelligent thought and action. Buchanan quotes Allen Newell and Herbert A. Simon:

> 'Any rational decision may be viewed as a conclusion reached from certain premises.… The behaviour of a rational person can be controlled, therefore, if

[3] 'On the analogy of a modern digital computer, the design principle of the *Analytical Engine* can be divided to:

1. *Input*. From 1836 on, punched cards […] were the basic mechanism for feeding into the machine both numerical data and the instructions on how to manipulate them.
2. *Output*. Babbage's basic mechanism was always a printing apparatus, but he had also considered graphic output devices even before he adopted punched cards for output as well as input.
3. *Memory*. For Babbage this was basically the number axes in the store, though he also developed the idea of a hierarchical memory system using punched cards for additional intermediate results that could not fit in the store.
4. *Central Processing Unit*. Babbage called this *the Mill*. Like modern processors it provided for storing the numbers being operated upon most immediately (registers); hardware mechanisms for subjecting those numbers to the basic arithmetic operations; control mechanisms for translating the user-oriented instructions supplied from outside into detailed control of internal hardware; and synchronization mechanisms (a clock) to carry out detailed steps in a carefully timed sequence. The control mechanism of the Analytical Engine must execute operations automatically and it consists of two parts: the lower level control mechanism, controlled by massive drums called barrels, and the higher level control mechanism, controlled by punched cards, developed by Jacquard for pattern-weaving looms and used extensively in the beginning of 1800s.' (Charles Babbage 2017)

> the value and factual premises upon which he bases his decisions are specified for him.'
>
> (Simon 1974 as cited by Buchanan 2005, pp. 54)

An important influence from this time is Norbert Wiener. In his research he noted that feedback is a key feature of all life forms and that plants, as well as animals, change their actions in response to their environment, ideas that would develop into the field of cybernetics.

> 'Cybernetics is interdisciplinary in nature; based on common relationships between humans and machines, it is used today in control theory, automation theory, and computer programs to reduce many time-consuming computations and decision-making processes formerly done by human beings.'
>
> (Encyclopædia Britannica 2017)

With the cybernetics the thoughts of 'thinking machines' where articulated. As Jean-Pierre Dupuy writes when he presents the cybernetics' first thesis, '[…] to think is to compute as a certain class of machines do – amounted to analysing and describing what it is to think […]' (Dupuy 2000, pp. 4-5).

During the fifties the research program of artificial intelligence started to develop and with that the focus was shifted from, what Dupuy describes as, 'the mechanization of mind' that had been the focus of the cybernetics, to 'the anthropomorphization of the machine'. The question asked instead became; 'Can a machine think?' (Dupuy 2000, pp. 5)

A key figure was Alan Turing who with his paper *Computing machinery and intelligence*, 1950, proposed the 'Turing Test', an imitation game that should function as a test of a machine's ability to 'think' and exhibit intelligent behaviour equivalent, or not possible to distinguish from, that of a human. The Turing Test has influenced debates over artificial intelligence until present days.

Between 1946–1953 ten cybernetic conferences where held, known as the 'Macy Conferences', in New York that are associated with the development of cybernetics. A few years later, 1956, the 'Dartmouth Summer Research Project on Artificial Intelligence' was held in New Hampshire, a workshop that is considered as the formal start of artificial intelligence as a field of research.

A few years earlier, 1952, Arthur Lee Samuel, working for IBM, wrote a chess-playing program that supposedly was the first self-learning program, and thereby the first AI program.[4]

Three other meetings focusing on AI are viewed as central in the establishing of artificial intelligence

[4] Samuel is said to have coined the term 'machine learning' 1959.

as a ‘full-fledged field of research’; ‘Session on Learning Machines’ in 1955, ‘Summer Research Project on Artificial Intelligence’ in 1956, and in 1958 the symposium ‘Mechanization of Thought Processes’[5] (Nilsson 2010, pp.73). At the gatherings two papers[6] were presented that have become fundamental to most of the later work done in enabling computers to ‘see’ something that is of importance in developing autonomous vehicles. During the following decades improvements were made in pattern recognition, speech, language processing and ‘understanding’, and robotics.

The early AI-research conducted is not to be perceived as without setbacks, or as evolutionary continuous line of progress. During the years the research had to deal with decreased funding and disbelief from fellow scientist and society, these periods are called ‘AI-winters’. Larry Hardesty notes that differences in the approach towards the research had a historical context, and that the controversies were influenced by the progress in technological research and the current ‘state of research mind’ (Hardesty 2017).

By the beginning on the nineties fresh approaches to programing were established. Earlier work was, by philosopher John Haugeland, given the name GOFAI, Good Old-Fashioned Artificial Intelligence, and described as; ‘uses heuristic search and discrete collections of symbolically represented facts and rules’ used by those handling ‘logical representations and logical reasoning methods’ (Nilsson 2010, pp. 413).

Contemporary research in artificial intelligence

When closing in on today’s AI-research my focus will be placed on autonomous vehicles and game-playing AIs in order to present what is regarded as ‘milestones’ in each of these AI-research areas.

The nineties saw the emergence of the World Wide Web, games becoming more realistic using 3D-graphics, robotic pets being introduced, and the first RoboCup[7] taking place. During the 21st century advancements has been made in e.g. digital personal assistants, robotics, natural language processing, and image analysing.

[5] Session on Learning Machines’ was held in conjunction with the Western Joint Computer Conference in Los Angeles, ‘Summer Research Project on Artificial Intelligence’ at Dartmouth College 1956, and the ‘Mechanization of Thought Processes’ at the National Physical Laboratory, UK.

[6] Wesley Clark and Belmont Farley of MIT’s Lincoln Laboratory concerning experiments conducted that explored the networks of neurons, especially in pattern recognition, wrote one of the papers. The experiments were mostly simulated on computers and the findings came to be called neural networks. Two other papers on the same subject with a different approach was written by Gerald P. Dinneen respectively Oliver Selfridge, also from MIT's Lincoln Laboratory. (Nilsson 2010, pp. 74)

[7] RoboCup is an annual robotic competition. The intention of the competition is stated on their webpage ‘[...] to use RoboCup as a vehicle to promote robotics and AI-research […]’ (RoboCup 2018)

In 1995 the researchers Dean Pomerleau and Todd Jochem, from Carnegie Mellon University, carried through a tour named 'No Hands Across America', from Pittsburg to San Diego. It was done using a car equipped with a computer running the software RALPH, Rapidly Adapting Lateral Position Handler. RALPH used video images to navigate and steer the vehicle while Pomerleau and Jochem handled the throttle and brake. This was one of the first (almost) autonomous vehicles. RALPH controlled the vehicle for at least 98 % of the trip, which was reported as a success, '[…] they demonstrated that with the right sensors and programming, a computer could indeed steer a car over thousands of miles of freeway driving.' (Baker 2017)

In 2005 STANLEY, created by Stanford University in cooperation with the Volkswagen Electronic Research Laboratory, ERL, won the 2005 DARPA (Department of Advanced Research Projects Agency's) Grand Challenge, a 132-mile course through the rough terrain of the Mojave dessert. To 'sense' the environment STANLEY was outfitted with cameras, radar, and laser rangefinders together with on-board software to command the steering, braking, and acceleration (Russell & Norvig 2010, pp. 28). The year after DARPA Urban Challenge took place, which tested autonomous vehicles ability to drive in traffic, obeying traffic rules, avoiding pedestrians, other vehicles and other obstacles. Boss, a car constructed by Carnegie Mellon University in partnership with General Motors Corporation[8] won the challenge (Spice 2007).

In 1994 Chinook, a checkers playing computer program developed at the University of Alberta, won the world champion title in a match against Marion Tinsley. Even though Chinook's victory was admirable it did not use any form of machine learning. The creators programmed all of Chinook's knowledge. A few years later, 1997, Deep Blue, IBM's chess-playing program, defeated the world champion Garry Kasparov, and became the first computer to win a chess match. Programming Deep Blue required a different approach than had been used to program Chinook. The reason being that in chess there are considerably more possible moves and to program all would have been impossible. Deep Blue was able to analyse 2-2.5 million positions per second, many more possible moves than any human, an advantage named 'brute force computing' (Campbell, Hoane & Hsu 2001, pp. 59). Deep Blue's victory was looked upon as remarkable since chess playing was 'once thought to epitomize human intellection' (Bostrom 2014, pp. 603–604), as Douglas Hofstader wrote;

> '[…] the way humans represent a chess situation in their minds is far more complex than just knowing which piece is on which square, coupled with knowledge of the rules of chess. It involves perceiving configurations of several related pieces, as well as knowledge of heuristics, or rules of thumb, which pertain

[8] Today General Motors Company.

> to such higher-level chunks. Even though heuristic rules are not rigorous in the way that the official rules are, they provide shortcut insights into what is going on on the board, which knowledge of the official rules does not. This much was recognized from the start; it was simply underestimated how large a role the intuitive, chunked understanding of the chess world plays in human chess skill.'
>
> (Hofstader 1979, pp. 603–604)

I will return to today's research further ahead in my paper, and I herby leave the historical background. Before continuing it should be emphasized that AI-research has not been a straightforward business. Worries have been vividly illustrated in works of fiction as well as presented by opponents to the AI-field. Concerns dealing with an unknown, and maybe frightening, future, as well as euphemistically reports in media have most likely had impact on the research in artificial intelligence's possibilities, such as the refutation of funding and support, as well as the lack there of, from politicians, fellow scientists and society. The development of AI could be analysed from how expectations has influenced the research in both positive and a negative ways, as Mads Borup et. al. notes on the subject of science and innovation;

> 'As such, future-oriented abstractions are among the most important objects of enquiry for scholars and analysts of innovation. Such expectations can be seen to be fundamentally 'generative', they guide activities, provide structure and legitimation, attract interest and foster investment. They give definition to roles, clarify duties, offer some shared shape of what to expect and how to prepare for opportunities and risks. Visions drive technical and scientific activity, warranting the production of measurements, calculations, material tests, pilot projects and models. As such, very little in innovation can work in isolation from a highly dynamic and variegated body of future-oriented understandings about the future.'
>
> (Borup, Brown, Konrad & Van Lente 2006, pp.285–286)

Simone Natale and Andrea Ballatore identify three dominant patterns in the construction of the AI myth;

> '(1) the recurrence of analogies and discursive shifts, by which ideas and concepts from other fields were employed to describe the functioning of AI technologies; (2) a rhetorical use of the future, imagining that present shortcomings and limitations will shortly be overcome and (3) the relevance of controversies around the claims of AI, which we argue should be considered as an integral part of the discourse surrounding the AI myth.'
>
> (Natale & Ballatore 2017, pp. 1)

Shortly lingering by the these patterns it seems obvious that especially the second pattern is prevalent in the past and present reporting and writing of progress and success in developing autonomous vehicles.

I have chosen not to investigate further into the sociology of expectations in my paper, but find it important to keep in mind that the social impact on the AI-area's hype is an essential factor of how the field has evolved, and in the way media and companies present progress.

Artificial intelligence and learning

Different ways of learning and how the software improves its performance and skills is central in programming artificial intelligence agents (software), especially in the subset machine learning. Russell and Norvig presents three types of feedback that determine three main types of learning; unsupervised learning, reinforcement learning and supervised learning. Unsupervised learning is when 'the agent learns patterns in the input even though no explicit feedback is supplied.' With reinforcement learning the agent 'learns from a series of reinforcements – rewards or punishments.' In supervised learning the agent 'observes some example input-output pairs and learns a function that maps from input to output.' They also mention semi-supervised learning, which is described as learning when we are 'given a few labelled examples and must make what we can of a large collection of unlabelled examples.' (Russell & Norvig 2010, pp. 695)

The mutual component in all AI-learning is the software's preconditions, which depends on four factors:

- 'Which component is to be improved.
- What prior knowledge the agent already has.
- What representation is used for the data and the component.
- What feedback is available to learn from.'

(Russell & Norvig 2010, pp. 694)

Recent AI-research has combined the different ways of learning, e.g. hybrid reinforcement learning[9] that works by using a multitude of agents that each has a specific task to fulfil, and a top agent who, from the suggestions feed by all the other agents, decides the outcome or action handles the control of

[9] Hybrid Reward Architecture for Reinforcement Learning, developed, and named, by researches from Microsoft Maluuba and McGill University in Montreal, Canada. (van Seijen et. al 2017)

all the agents.[10] Another example is deep reinforcement learning[11], which combines reinforcement learning and deep learning (a description of deep learning follows bellow).

Terminology

To initiate my investigation a brief rundown of the terms used in the discussion of AI and, for my paper central, in Harry Collins epistemological theory and argumentation against the possibility of creating a general AI will be done.

Artificial Neural Network (ANN)

'Neural nets are a means of doing machine learning, in which a computer learns to perform some task by analysing training examples.' (Hardesty 2017) The idea of neural nets are modelled loosely on the human brain, and when the term first was introduced, 1944 by McCullough and Pitts, research was conducted using the term in both neuroscience and computer science. The neural nets of today are usually organized in layers and data moves through them in only one direction. This is called 'feed-forward'. The net is made up of nodes and each node can have several connections in the layer below. The 'down-under' connections feeds the node with data and the data is, by the node, feed-forward to the above node(s) that it is connected to. Each node processes the data that it receives from the layer below, but does not automatically send everything ahead.

Deep learning

Deep learning is a part of the broader term machine learning, and uses lots of layers in a neural network to analyse data at different abstraction (Vincent 2016). Deep learning software attempt to mimic the way our brains work; 'the activity in layers of neurons in the neocortex, the wrinkly 80 percent of the brain where thinking occurs' (Hof 2013). The software models are described as 'loosely related to information processing and communication patterns in a biological nervous system, such as neural coding that attempts to define a relationship between various stimuli and associated neuronal responses in the brain.' (Olshausen1996 as referred to by Wikipedia 2018)

[10] Hybrid reinforcement learning was used to make top-score at playing Ms. Pacman during June 2017. Microsoft. (Linn 2017)

[11] Deep reinforcement learning was used in Google's Go-playing DeepMind version, AlphaGo, which won a five-game match against a professional Go-player in March 2016.

Deep learning needs a lot of data and time to process the data, how long depends on the work that it is set to do. Each layer has a specific task to perform and, as with neural nets, a training period is needed where either a human or the system itself corrects the program with the goal of it to get better at the assigned task.

Machine learning

Machine learning is a 'discipline concerned with the implementation of computer software that can learn autonomously' (Encyclopædia Britannica 2017). Several approaches to machine learning exist. A machine learning software is able to learn from data and also to make predictions on data, which makes it useful e.g.in situations where programming explicit algorithms will make the performance of the software less successful. Two simple examples are email filtering and programs aiming at detecting potential data breach.

Mimeomorphic and polimorphic action

Collins defines two types of actions, mimeomorphic action and polimorphic action.

'Actions that are intentional but nevertheless mimic the world of mechanical cause and effect are called mimeomorphic actions.' (Collins 2010, pp. 55)

'Polimorphic actions are actions that can only be executed successfully by a person who understands the social context.' (Collins 2010, Preface ix)

Mimeomorphic actions are described as those where a human act mechanically, a situation where humans can be seen as continuous with the world of mechanisms, and Collins hold that humans can sometimes choose to act in this way even if their actions are intentional[12].

Polimorphic actions are described as actions 'that require different behaviours for successful instantiation depending on context and require different interpretations of the same behaviour depending on context', (Collins 2010, pp. 125) and humans are said to use collective tacit knowledge[13] to handle them.

[12] The definition of intentional in Collins' context is 'meditated through the world of meaning'. (Collins 2010, pp. 55)
[13] I will return to collective tacit knowledge, CTK, and describe it more thorough. In short CTK is described as having to do with the way society is constituted and a form of tacit knowledge.

Narrow/weak artificial intelligence

Narrow or weak[14] artificial intelligence is designed to do a specific narrow task. The narrow/weak AI works towards a predefined goal, and is limited by its algorithms in performing its task(s). Examples of narrow/weak AIs are robots used for manufacturing, chat-bots and digital personal assistants.

Social cartesianism

Social cartesianism is by Collins defined as the essential difference between humans and animals. No sharp distinction between animals and humans is drawn, but rather the difference is in abilities: ‘ The issue is the marked difference in abilities between a species that possesses fully developed languages and cultures and one that does not.’ (Collins 2010, pp. 126)

Socialness

Socialness according to Collins is described as ‘the ability to absorb ways of going on from the surrounding society without being able to articulate the rules in detail.’ (Collins 2010, pp. 125) Socialness is when personal traits correspond to significant cultural differences.

Strong/general artificial intelligence

A strong, general or full artificial intelligence is one that can be compared with any cognitive functions that a human being may have and is thereby in its essence not different from a real human mind. This implies that a general AI has mental capabilities and functions that are of the same functional status as a human brain. The AI is in principle not limited by its algorithms and is able to learn, adopt and act flexible by interacting with surroundings such as databases, physical environment, etc.

> ‘According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have

[14] Weak artificial intelligence can also be a definition of the thought held by the group of people who think that a general, strong, artificial intelligence is not possible. The terms weak and strong artificial intelligence where introduced by John Searle in *Minds, brains, and programs* 1980.

> other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.'
>
> (Searle 1980, pp. 417)

Some researcher, e.g. Marvin Minsky and Nils Nilsson, use the term human-level AI, which is described as 'machines that think, that learn and that create' (Russel & Norvig 2010, pp. 27).

Critique of the AI-project

Doubts of and arguments against the AI-project have been around from the start. Of interest for this paper are the arguments that speak against artificial intelligence, and are presented, from an epistemological point of view, especially arguments putting forward the uniqueness of the human being as the reason for the impossibility of an artificial intelligence.

Arguments against the AI-project can be divided into those who say that an AI *cannot* be constructed and those who argue that it *should not*. *Should not* arguments focus on ethical and political questions in relation to research and innovations in AI, e.g. if it is ethically tenable to conduct research in the area of artificial intelligence, or if society should subsidize innovations that will have as a consequence that different occupational groups lose their jobs. My focus here is on the former, the *cannot* argument(s). These arguments focus on questions of social ontology and epistemology, which are the ones Harry Collins is concerned with, and that are in line with my research questions.

Arguments against artificial intelligence

Harry Collins is not alone to argue against the possibility of creating general AIs. I will briefly present two other critiques, John Searle's and Hubert L. Dreyfus', together with a short overview of some of the arguments against the possibility of a disembodied AI. What these all have in common is that they might seem similar to Collins' arguments but he rejects them in parts.

The need of intentionality – John Searle

John Searle argue that computational processes lack something that humans (and animals) have, namely intentionality. He presents two propositions, (1) and (2), from which he draws three consequences, (3), (4) and (5), that are central to his arguments.

(1) 'Intentionality in human beings (and animals) is a product of causal features of the brain.
(2) Instantiating a computer program is never by itself a sufficient condition of intentionality.'

(Searle 1980, pp. 417)

Searle assumes that (1) is an empirical fact about the casual relations that actually takes place between mental processes and brains. He clarifies his stance by stating that this presents the idea that certain brain processes are sufficient for intentionality.

(3) 'The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program. This is a strict logical consequence of 1 and 2.
(4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1.
(5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. This follows from 2 and 4.'

(Searle 1980, pp. 417)

In his line of argument Searle states that the only way a machine could think is if it would be possible for it to have '[…] internal causal powers equivalent to those of brains', which he think is not possible since he believe that no computer-program is '[…] by itself is sufficient for thinking'. (Searle 1980, pp. 417)

Searle describe intentionality as follows:

> 'Intentionality is by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world. Thus, beliefs, desires, and intentions are intentional states; undirected forms of anxiety and depression are not.'
>
> (Searle 1980, pp. 424)

Searle's thesis contains the central term 'understanding' that accordingly '[…] implies both the possession of mental (intentional) states and the truth (validity, success) of these states' (Searle 1980, pp. 424), and in he put focus only on the former. A computer-program does not have knowledge of what the symbols they use mean. Nilsson writes '[…] computational processes lack "aboutness" […]' (Nilsson 2010, pp. 383). Here rests the difference between a human and a machine in Searle's argument, because when a human use a word s/he knows what it is about.

A thought experiment is used to illustrate his arguments concerning understanding, 'the Chinese Room', and Nilsson presents a short summery of the main question at stake together with Searle's answer.

> 'Searle's question, essentially, is "Can it be said that the room (containing Searle, the rules, and the batches of Chinese symbols) `understands' Chinese?" Searle

> claims the answer is "no" because all that is going on is 'formal symbol manipulation' without understanding what the symbols mean.
>
> (Nilsson 2010, pp. 386)

John Searle versus Harry Collins

Searle's and Collins' argument differ in that Searle refers to internal states of the single individual while Collins place the uniqueness of the human intelligence in the collective interaction in society. The knowledge, 'construction' of knowledge, and exchange of knowledge that come about when humans socialize in society and different scientific areas all depends on the socialization, which Collins stress to be fundamental in human learning, knowledge transfer, communication and interaction.

In relation to the Chinese Room Collins think that Searle's deduction 'that the mere reproduction of an action, […] does not demonstrate that consciousness or understanding or meaning is involved' (Collins 2010, pp. 127) is correct. He does not agree with Searle on the goal of his inquiry though, since his own aim is to investigate whether the though experiment provide 'the conditions under which humans, who normally accomplish fluency in language only through internalizing tacit knowledge through socialization, could accomplish something indistinguishable from fluency by referring only to a lookup table and doing nothing more than string transformation' (Collins 2010, pp. 127).

Unlike Searle Collins traces the conclusion from his description of how context is provided in life by both conversation and what is happening in the world, 'questions and answers are sensitive to context and, […] language changes' (Collins 2010, pp. 129).

Searle has continued to be active in the debates concerning AI and in short his arguments at present are that consciousness is a biological phenomena and as such no stranger than the body or water. But it is a biological phenomenon that we have not been able to describe epistemological yet[15]. In principle he sees no obstacle in us being able to build an artificial brain, one that thinks as a human, but what needs to be done first is that we need to figure out how the brain works, and this has not been done, because to think requires 'a brain or something with equivalent causal powers to the brain'. (Searle 2015)

[15] Searle hold that ontological is consciousness subjective. (Searle 2015)

The need of global processes and being in the world – Hubert L. Dreyfus

Hubert L. Dreyfus wrote the paper *Alchemy and Artificial Intelligence* after he had spent the summer at the RAND Corporation[16] in Santa Monica, California. It was written as a reflection, and a critique, of the current reported successes in digital computing and the expected future. Dreyfus, like Collins, place his focus on the *cannot* argument(s) and to what extent a digital computer can be programmed to exhibit simple intelligent behaviour characteristic of e.g. children. In the paper he argues that the reason for computers having difficulties to simulate cognitive processes is because they exclude three forms of information processes that are said to be fundamental for humans, fringe consciousness, essence/accident discrimination, and ambiguity tolerance. (Dreyfus 1965, pp. iii).

Fringe consciousness is described as a two-pieced ability where a person at the same time is able to focus on the details in a certain situation, or task, and simultaneously be aware of and take into account the overall situation, or task. What is lacking when AI-research fails is accordingly the cases where global awareness is necessary, and the 'counting out' ability is not enough, or feasible.

Essence/accident discrimination is described as the ability to distinguish the accidental from the essential, and as illustration Dreyfus quotes Max Wertheimer's writing in *Productive Thinking*: 'The process [of structuring a problem] does not involve merely the given parts and their transformations. It works in conjunction with material that is structurally relevant but s selected from past experience […] [40:195].' (Dreyfus 1965, pp. 29)

Ambiguity tolerance is described as the ability to handle and reduce ambiguity, e.g. in using a natural language. This ability is said to presuppose the two other abilities, where fringe consciousness is said to make a person aware of cues in the context, and a sense of what is important in the context allow the person to ignore what is irrelevant. What ambiguity then allows the person to do is to use the information and narrow down on the remaining spectrum of possible parsings and meanings as much as is required 'without requiring the resulting interpretation to be absolutely unambiguous.' (Dreyfus 1965, pp. 34–35)

The combination of the forms of information processing permits what is called perspicuous grouping, which Dreyfus, inspired by Wittgenstein, describes as 'understanding which consists in seeing connections' (Dreyfus 1965, pp. 45). According to Dreyfus the processes are essential for intelligent behaviour but not something computers are able to make use of. It is the way that a brain compared with computer processes information that is the weak spot. He defines four things that a machine,

[16] 'The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is non-profit, nonpartisan, and committed to the public interest.' (RAND 2018)

which is to be counted as equal to human performance, must be able to do; ’1) Distinguish the essential from the inessential features of a particular instance of a pattern; 2) Use cues which remain on the fringes of consciousness; 3) Take account of the context; 4) Perceive the individual as typical, i.e., situate the individual with respect to a paradigm case.’ (Dreyfus 1965, pp. 46)

Nilsson emphasize that Dreyfus does not think that creating an AI is impossible as such but that an AI is not possible due to the way computers works. The human intelligence is derives from us ‘being in the world’, not because we are guided by rules, and this is linked to the definition of a body. Nilsson cites a conversation he has had with Dreyfus; ‘[…] a model of our particular way of being embedded and embodied such that what we experience is significant for us in the particular way that it is. That is, we would have to include in our program a model of a body very much like ours with our needs, desires, pleasures, pains, ways of moving, cultural background, etc.’ (Nilsson 2010, pp. 392)

This is in line with what Dreyfus wrote when foreboding about the future of AI development; ‘Significant developments in artificial intelligence in the remaining two areas must await computers of an entirely different sort, of which the only existing prototype is the little-understood human brain.’ (Dreyfus 1965, pp. iii)

Hubert L. Dreyfus versus Harry Collins

Compared to Collins it occurs as if Dreyfus, at least in principle, leaves a door open for the potential AI, given that a new type of computers, or rather programs, is innovated. Even if they have similarities, e.g. stressing the context and taking into account the ‘whole’ picture, they differ in important parts. Collins major point is the social connections, the social networks that humans are entwined in, and while Dreyfus also highlights the context it is not to the same extent as Collins. Dreyfus seems to place the individual in the centre rather than the collective, as Collins does.

In *Tacit and Explicit Knowledge* Collins criticise Dreyfus, and the influence he has had, stating that ‘it reflects much less about the nature of knowledge itself than is usually imagined.’ The reason he writes is that Dreyfus only describe ‘one or two special kinds of human skill and it entirely ignores the most fundamental subdivision of human expertise: that between expertise of the sensory motor kind and expertise of the social kind.’ (Collins 2010, pp. 124)

Arguments against a disembodied AI

There are several other objections to artificial intelligence, and several of these argue that a

disembodied AI is not possible. Nilsson presents a list of arguments focusing on the differences between brains and computers.

- ‘Computers have perhaps hundreds of processing units whereas brains have trillions.
- Computers perform billions of operations per second whereas brains need perform only thousands.
- Computers are subject to crashes whereas brains are fault tolerant.
- Computers use binary signals whereas brains work with analogue ones.
- Computers do only what their programmers tell them to do whereas brains are creative.
- Computers perform serial operations whereas brains are massively parallel.
- Computers are constrained to be ”logical” whereas brains can be “intuitive.”
- Computers are programmed whereas brains learn.’

(Nilsson 2010, pp. 392–393)

Some of the stated differences Collins would agree upon, and they fit into his theory, but neither is sufficient to his arguments, as has been foreboded in his responses to Searle’s and Dreyfus’ theories. The ones that could fit into his theory are that brains are creative, are massively parallel, can be intuitive and that they learn. Although Collins stresses that there are differences between a computer/AI and the human brain it is not the functions that are put in focus but rather the way humans (brain and body) are socialized and part of the collectivity. The differences listed above are illustrative in the sense of functions and abilities but as arguments against a potential general AI they are far from enough for Collins.

Harry Collins epistemological theory and artificial intelligence

Collins’ arguments against the possibility of creating an artificial intelligence stems from his theories of knowledge and expert(s). In *Changing Order: Replication and Induction in Scientific Practice*, first published 1985, he, inspired by David Bloor’s the Strong Programme and the later Ludwig Wittgenstein presents his Empirical Programme of Relativism, EPOR, consisting of three stages:

1) ‘Demonstrating the interpretative flexibility of experimental data.
2) Showing the mechanisms by which potentially open-ended debates are actually brought to a close - that is, describing closure mechanisms.

3) Relating the closure mechanisms to the wider social and political structure.'

(Collins 1992, pp. 25–26)

Collins empathizes that the solution to the problem of induction is achieved by a shift of focus from 'how we could be certain *in principle* about induced regularities' to 'how we actually come to be certain about regularities *in practice*' (Collins 1992, pp. 6). With the 'sociological resolution of the problem of induction' he distances himself from his predecessors, by placing the explanation-power in the social structures rather than in safe-proving the structures of how science is conducted. The influences from the later Wittgenstein, and his term 'forms of life', are visible in Collins thoughts; in his focus on language as a vital part of socialization as well as his focus on the 'how we come to be certain'.

The social structures are said to answer to the *experimenters regress*, which occurs when science uses the reproduction of experiments to prove that they are possible to reproduce, and in the definition of what is the 'correct' way to carry out experiments in a new field of research. According to Collins the regress is broken and closure is reached by negotiation within the appropriate scientific community, rather than scientific reason. 'There is, then, no set of 'scientific' criteria which can establish the validity of findings in this field. The experimenters' regress leads scientists to reach for other criteria of quality.' (Collins 1992, pp. 88)

The central problem, and the one that constitute the bridge to Collins critique of artificial intelligence, is called the socialization problem, and Collins defines two aspects that need to be settled for the problem to be solved. The first deals with how humans approach and handle questions and answers. Collins holds that this is structured by the unfolding context; to ask, and reply to, a question and to follow up with further questions, or answers, the 'speaker' needs to be sensitive to what is going on around her/him. The second aspect focus on 'the right way to do things', which only is possible to capture through experience, something that, together with its application can vary from e.g. country to country.

Knowledge

A central pivot is the difference Collins outlines between explicit and tacit knowledge, a line of arguments started in *Artificial Experts*[17] and refined in *Tacit and Explicit Knowledge.*

[17] In *Artificial Experts* Collins urges that the conventional status hierarchy of knowledge undervalues practical abilities, a statement that also is made to favour the skills a person with sociological or philosophical training have, and what makes them 'unique' in completing the picture of knowledge.

Collins emphasize that all explicit knowledge rests upon tacit knowledge, and that no concept of tacit knowledge would exist without one of explicit knowledge. With tacit knowledge he thinks that the mistake made is the failure to separate different types of tacit knowledge, and separates out three. 'These have to do, respectively, with the contingencies of social life (relational tacit knowledge), the nature of the human body and brain (somatic tacit knowledge), and the nature of human society (collective tacit knowledge) – RTK, STK, CTK.' (Collin 2010, pp. Preface x)

Collins hold, inspired by Michael Polanyi's texts, that it is a mistake to believe that tacit knowledge is harder to understand, or obscure, compared with explicit knowledge. Tacit knowledge is normal life, and what humans and animals have been doing '[…] without anyone *telling* anything to anything or anyone' (Collin 2010, pp. 7). Collins does not find anything strange with things being done but not being told stating that all knowledge is either tacit or rooted in tacit knowledge[18], and explicit knowledge does not have any significance if the tacit is missing. But, he continues, the idea of tacit could not occur, as being special, until the explicit, or explicability became to be granted as the ordinary. Note that the relation is not symmetrical. The dependence is in the first case where the explicit, with Collins words, is parasitical on the tacit handles *knowledge itself*. The later, where the tacit is parasitical on the explicit, handles *the idea* of the tacit.

Explicit knowledge, strings, communication and language

' "Explicit" is something to do with something being conveyed as a result of strings impacting with things.' (Collins 2010, pp. 57)

Collins holds that the basic level is strings that together with entities are described as the interaction between physical objects. Strings are said to be flexible; a string can be an entity, and an entity a string, depending on the situation, and they always contain information but are not always used to transmit information. Information in this context is described as '[…] a physical feature of a string that refers to the number and arrangements of its elements.' (Collins 2010, pp. 16) The terms 'string', 'strings' and 'elements' are used interchangeable, with the explanation that they are 'just entities'. Strings are said to be able to affect entities in four ways:

> (1) 'A string is a physical thing, so it can have a physical impact.
> (2) A string is a pattern, so it can impress, print, or "inscribe" a similar pattern on an entity in many different ways.
>
> […]

[18] This line of thought is also founded on thoughts presented by Michael Polanyi.

(3) a string can communicate "mechanically," as when a new piece of code is fed into a computer or a human reacts to a sound in a reflex-like way; and
(4) a string can communicate by being interpreted as meaningful by a human.'

(Collins 2010, pp. 16–17)

Collins describes (1) and (2) as less fundamental than (3) and (4), with which a change can come about in a way that can cause the affected entity to do something or give it the ability to perform something that the entity could not do before. (1), (2) and (3) are said to be possible for other entities than humans, but (4) is unique for humans.

In (3) and (4) communication is a vital part, and inspired by Wittgenstein Collins writes:

'A communication takes place when an entity, P, is made to do something or comes to be able to do something that it could not do before as a result of the transfer of a string.' (Collins 2010, pp. 20–21)

Stating that communication is not always simple, Collins presents five enabling conditions for communication illustrated with an example of a simple multiplication. The fifth is of importance in his argument, and the fourth shows us were he places the line between the computer/software and a human. The fourth condition is described as; 'The transfer of a string plus a significant fixed physical change in the receiving entity gives rise to a communication.' (Collins 2010, pp. 31) Image an old computer at which a calculation-program has been installed and the multiplication instructions have been typed but the computer cannot proceed due to lack of memory. If the computer shall be able to do the sum a new memory unit must be installed.

The fifth condition is described as; 'The transfer of a string plus a significant flexible and responsive physical change in the entity gives rise to a communication.' (Collins 2010, pp. 31) This condition requires a fluency in a language, and only applies to humans, or humanlike things. Language is said to be '[…] always changing as the circumstances and societies in which they are located change.' (Collins 2010, pp. 30) Condition five differs from condition four as the change cannot be a fixed one, but a rather is a flexible ability. 'Somehow, the ability that has to be transferred to engender fluent language use has to be flexible–it has to be an ability that can respond to social cues and contexts.' (Collins 2010, pp. 31)

Here Collins connects his epistemological theory, and the socialization problem with his arguments against the possibilities of AI, which he further elaborates by withholding that the difference between strings and language is a 'major principle' because: 'A language *cannot* be transformed. A language can only be *translated*, and translation always involves the risk of irremediable loss or change of *meaning*.' (Collins 2010, pp. 25)

Stings on the other hand are said to be possible to transform, and to transform back, without loss of

information. Collins holds that information in this context is a term that belongs to the physical world, and if any loss of information takes place in practice it can be more or less fixed with the help of techniques from the physical sciences. He presents three stages of which language translation, and plain conversation, within one natural language consists.

(1) Inscription – 'In "telling" the attempt is made to represent lived meaning with the inscribed string.' (Collins 2010, pp. 27)
(2) Transmission and transformation – 'This is the move of the string from one person to another that always involves *transformation* of the string from one form to another.' (Collins 2010, pp. 28) Collins says that this is the domain of ordinary cause and effect.
(3) Interpretation – 'This is the attempt to recreate meaning from the string—to interpret it.' (Collins 2010, pp. 28)

The third is regard as unique to humans, or at least humanlike beings.[19] Language is said to be full of meaning, whereas strings are defined as essentially meaningless, and Collins considers meaning to be something 'that relates to the changing ways people live in society.' (Collins 2010, pp. 26)

It seems to me that Collins here is on a similar route as Searle. In the interpretation that Collins withhold as unique to humans the intentionality that Searle withhold appears to be essential. To create meaning (from the string) could be said to be one version of an intentional state (the feature by which the mental state is directed at or about).

Collins believes that explicit knowledge can be transferred by the use of strings in the 'right circumstances', which are defined in five conditions for communication, and of these the first three are possible to explicate but condition four and five 'comprise changes in the receiving entity rather than changes in the string.' (Collins 2010, pp. 81) Four meanings of explicable is presented:

> (1) 'Explicable by elaboration – A longer string affords meaning when a short one does not.
> (2) Explicable by transformation – Physical transformation of strings enhances their causal effect and affordance.
> (3) Explicable as mechanization – A string is transformed into mechanical causes and effects that mimic human action.
> (4) Explicable as explanation – Mechanical causes and affects are transformed into strings called scientific explanations.'
>
> (Collins 2010, pp. 81)

[19] Collins writes that it might be the case that dolphins, chimpanzees or some other animals have the interpretive abilities as humans to some degree. (Collins 2010, pp. 25)

Knowledge and the possibility of creating artificial intelligence

Collins epistemological theory withholds aspects of knowledge basics, transfer and communication that are said to be unique and/or can only be performed by humans. Together with the problem of socialization they constitute Collins' 'unbreakable' line between a human and what can be achieved by the AI-project. Lets shortly recapture these before continuing to Collins' arguments for the need of a body/brain and tacit knowledge.

When discussing strings Collins put forward that the level that only humans can master is the fourth; 'a string can communicate by being interpreted as meaningful by a human.' This is related to the central place language plays in his theory. Of the conditions for communication presented it is the fifth condition; 'the transfer of a string plus a significant flexible and responsive physical change in the entity gives rise to a communication' that is withheld as applicable to humans. The reason being that to be, or become, fluent in a language require a flexibility, and ability to respond to social cues and contexts, which, according to Collins, is something computers lack.

One way to understand what Collins is aiming at here is to think about when you learn a new language. You might master the grammar, pronunciation, and have a large vocabulary, but when you visit a country were the language is the mother tongue you will notice that the way people speak differs in in different regions, and/or in cultural or social contexts, and a joke every native laughs at does not make any sense to you. There is in the translation of a language, as Collins states, always a risk of loss or change of meaning. Translation of a language, and plain conversation, takes place in three stages and it is the third, interpretation, that Collins hold to be unique to humans. Here again a connection is made to socialization as Collins states that language is full of meaning and is intertwined in the changing ways people live in society.

The relation between the body, the world and knowledge

In *Artificial Experts* Collins introduces the body and brain as vital 'organs' for a human's ability to gain knowledge, and he differentiate between behaviour-specific acts and nonbehaviour-specific acts. The acts are similar in the way that they can be accomplished either self-consciously or unself-consciously, but (some) nonbehaviour-specific acts are not possible to mimic (or make explicit). These are called 'embodied' and 'embrained' knowledge, and Collins thinks that what is special are that this knowledge cannot be taken out of the respective organs. Embodied/embrained knowledge differ from what Collins calls encoded knowledge, which is knowledge a human might do without thinking but one that can be 'broken and its contents read' (Collins 1990, pp. 219).

In *Tacit and Explicit Knowledge* the embodiment/embrained thesis is developed and the 'minimal embodiment thesis' and the 'social embodiment thesis' are introduced. The minimal embodiment thesis builds on the distinction between animals and humans, and Collins do not think that animals are able to be, or become, 'social parasites' and are only, and always, an individual thing. To be a 'social parasite', requires only a minimal (human) body, writes Collins, but the human brain is vital.

The social embodiment thesis states that only humans have tacit knowledge, and this is a matter of the way human bodies are with 'the kind of brain that is associated with the language speaking species and the larynx, lungs, and ears or prostheses that can take their place' (Collins 2010, pp. 135). The collectivity's language is said to only develop its characteristics as a result of the members of the collectivity practices.

Collins also introduces the sociological uncertainty principle: 'When a system is completely understood, it is too late for all practical purposes.' Collins 1990, pp. 220) Preceding the principle is the assumption that a mistake is done when a scientific description of the world is taken to be the world itself. There is, says Collins, no problem with capturing what is stated in textbooks and insert them into a computer-program (expert system). The difficulty lays in trying to describe, and encode, the day-to-day practice of a scientist.

'Where the principle applies, only knowledge from low down the hierarchy is useful. This knowledge is not found in texts or computers but only in skilful people.' (Collins 1990, pp. 220)

Tacit knowledge

In *Tacit and Explicit Knowledge* Collins presents a three Phase Model of tacit knowledge; relational tacit knowledge, RTK, somatic tacit knowledge, STK, and collective tacit knowledge, CTK. The presented definition of tacit knowledge is '[…] knowledge that is not explicated' (Collins 2010, pp. 1).

The first phase, relational tacit knowledge (weak tacit knowledge), RTK, is described as knowledge that could be made explicit by a physical transformation of strings (the second meaning of explicable). Collins thinks that the reason RTK is not made explicit turns on 'the way societies are organized', rather than depend on the nature and the location of knowledge or the way humans are constituted. In RTK both sender and receiver have enough cultural similarity from the beginning for a string to transfer the intended meaning if the string is proportionality the needed length. Collins believes that RTK, in principle, is possible to be made explicit.

Somatic tacit knowledge (medium tacit knowledge), STK, has to do with 'properties of individuals' bodies and brains as physical things' (Collins 2010, pp. 85–86). Collins holds that this is the kind of

tacit knowledge that humans have in common with animals and other living things, such as trees and plants. STK is possible to be made explicit by a string being 'transformed into mechanical causes and effects that mimic human action (the third meaning of explicable). He holds that this transformation is an outcome of research done by scientists, and identifies two major subdivisions of somatic tacit knowledge; somatic-limit tacit knowledge and somatic-affordance tacit knowledge. The first is described as knowledge that is tacit because of our bodily limits. The later is knowledge that turns on the 'special physical nature of the body (and brain)'. Both subdivisions can be made explicit by mechanical causes and affects are transformed into strings (the fourth meaning of explicable). Somatic-limit tacit knowledge can also be made explicit by the third meaning but this is not the case with somatic-affordance tacit knowledge, at least not yet but it might be possible in the future says Collins. This implies that at present somatic-affordance tacit knowledge only affects humans, since 'only the human body and brain are made of suitable substances' (Collins 2010, pp. 113).

Collective tacit knowledge (strong tacit knowledge), CTK

Collective tacit knowledge, CTK, is of most interest in Collins arguments against AI, and has to do with 'the way society is constituted'. According to Collins CTK we do not know how to make explicit, and we are not able to predict how to explicate CTK in any of the four senses of explicable. CTK is located in society, and Collins mentions social sensibility as a characteristic of a human in possession of CTK. Social sensibility is described as ability to adapt to different social and/or cultural setting, and to improvise in common situations in ordinary life.

> 'Bike riding in traffic, car driving in traffic, and dancing all require learning with a degree of flexibility so that the style can be changed to fit different circumstances, such as riding or driving in different countries and dancing in different settings.'
>
> (Collins 2010, pp. 123)

The flexibility is said to be a matter of the fifth condition of communication; 'The transfer of a string plus a significant flexible and responsive physical change in the entity gives rise to a communication'.

Collins holds that CTK depends on specific features of humans, as is the case with somatic-affordance tacit knowledge, but it is not the individual that is the locus of knowledge but instead the collectivity, the individual shares the collectivity's knowledge. 'The special thing about humans is their ability to feast on the cultural blood of the collectivity in the way that fleas feast on the blood of large animals.' (Collins 2010, pp. 131)

Collins illustrative definition of humans is as parasites, and our uniqueness is that our brains 'afford parasitism in the matter of socially located knowledge'. In Collins theory human bodies are not to be looked upon as boundaries for knowledge but instead the collective knowledge is seated in the collectivity of brains, which Collins states is 'just a large- scale version of my brain'. Individual brains are described as connected to the neurons of every other brain in the way that the individual brain is in touch with through the five senses. 'The collectivity of brains is just a large-scale version of my brain– it is just a bigger collection of interconnected neurons–and, as with synapses, the weights of the connections change whenever social and technological life is rearranged' (Collins 2010, pp. 132).

Even if language is said to play an important role it is not withheld to alone be the most efficient. Collins also believes that physical activity is a step towards, and into the locations, where linguistic fluency is learned.

While the other phases of tacit knowledge are held as possible to explicate CTK it is not that simple, and Collins believes that before a solution to the socialization problem, which can make explicable the way humans acquire CTK, is presented we are not able to explicate CTK, and thereby not able to develop an AI.

> 'We can describe the circumstances under which it is acquired, but we cannot describe or explain the mechanism nor build machines that can mimic it. Nor can we foresee how to build such machines in the way we can foresee how we might build machines to mimic somatic tacit knowledge.'
>
> (Collins 2010, pp. 138)

Summary

Collins arguments against the possibility of creating a general AI are interwoven with his epistemological claims. I will return to this in the conclusion but let me shortly relapse what parts that constitute the demarcation line before moving on to the status of today's research in artificial intelligence. Collins believes that some sorts of tacit knowledge can be made explicit and by that, at least in theory, be adapted by machine/computer software. The tacit knowledge that cannot be made explicit is CTK, collective tacit knowledge. CTK is located in society, tacit knowledge that we have collectively and that changes with society and the use of language. A main point in Collins argument is that he makes a difference in the role of, what he calls, 'the typical body' and 'the individual body', and this is the focus in the minimal embodiment thesis. To be as social parasite, on the collectivity, only a minimal body is necessary but the human brain is vital, he writes, thereby putting the human brain in the central position for the connection, and communication in the collectivity. From Collins

point of view the difference between a human being and an artificial intelligence is ontological and it does not matter what e.g. Moors Law predict, thereby stated that it does not matter how powerful or advanced a computer chip or a software become, a computer will never be able to gain CTK. This is a different view than the ones Searle and Dreyfus propose. For Searle the hindrance for us to create an AI is not ontological but epistemological, we do not know what the biological constitution of consciousness is (yet), and since we do not have the epistemological facts at hand we cannot say that an AI is or is not possible. For Dreyfus the obstacle to overcome is to create computers of an entirely different sort, of a sort that have a model of a body similar to the human body with human experiences of 'being a human'.

AI – the goals reached 2018

In the following section focus is put on the two chosen segments of AI-research, autonomous vehicles and game-playing AIs. The two areas of research are chosen with their reported progress and Collins theory in mind. The area of autonomous vehicles are supposed and reported to in the near future be able to handle different types of environments, navigate in difficult weather, interpret conflicting signals and make ethical decisions, etc. with increased safety compared with a human driver. Game-playing AI's performance are at presented tested and used in a secluded game-environment but the research is of interest since these AIs are reported to be able to create their own set of rules and perform a basic communication. Both areas appears to challenge Collins critique of a potential general AI by exhibiting social sensibility, being able to adapt to different social and/or cultural setting and improvise when needed plus developing a (limited) language in order to communicate.

Autonomous vehicles

The benefits of autonomous vehicles are anticipated to be several, impacting traffic safety, equity, the environment and economy, e.g. quoting Tesla's and Volvo's websites:

> 'All Tesla vehicles produced in our factory, including Model 3, have the hardware needed for full self-driving capability at a safety level substantially greater than that of a human driver.'
>
> (Tesla 2018)

> 'We believe that mobility should be safer, sustainable and more convenient. For Volvo Cars, technology should make people's lives easier. That's why our approach to autonomous driving is all about the people that will use them.'
>
> (Volvo Car Corporation 2018)

Tesla has already the autopilot-function installed in several models, and Volvo is currently running their Drive Me project in Gothenburg, Sweden. (Volvo Car Corporation 2018) The project is said to be a pre-step in Volvo's development of safe autonomous driving. Volvo is not unique in tackling the test-process by letting the public try out their technology, e.g. a similar project is the Ann Arbor Connected Vehicle Test Environment', AACVTE (Ann Arbor Connected Vehicle Test Environment 2018) initiated by University of Michigan Transportation Research Institute.

When talking about autonomous vehicles it should be remembered that automatic processes have been common in the car industry for several years. A table of layers of autonomy can be of use for

discriminating at what level of automation the industry is today, and which goal(s) that are aimed at. SAE International defines in standard J3016: *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*, six levels (0 – 5) of driving automation (SAE International 2014).

Level 0: No Automation – all major systems are controlled by the human driver.

Level 1: Driver Assistance – certain systems, e.g. cruise control and automatic braking, may be controlled by a driver assistance system but the human driver monitor the driving environment and all other parts of driving.

Level 2: Partial Automation – the driver assistance system can execute both steering and acceleration/deceleration, but it requires the human driver for safe operation.

Level 3: Conditional Automation – the automated driving system can execute all aspects of driving, but the human driver is expected to respond appropriately to a request to intervene.

Level 4: High Automation – the automated driving system handles all aspects of driving in some scenarios, but not all.

Level 5 – Full Automation – the automated driving system handles all aspects of driving in every situation that can be managed by a human driver.

In level 0–2 the human driver monitor the driving environment, but in level 3–5 the automated driving system monitor the driving environment in different degrees. Autonomous vehicles today are at level three, and slowly crossing over to level four, which is expected to happen in the next couple of years. (Union of Concerned Scientists 2018)

Different brands of autonomous vehicles are more or less using the same technology, although the field of research is done with different approaches; car manufactures combining hardware development with software development, universities combining software development with prototype-vehicle building, companies specializing in software, etc. For example Tesla uses a neural net program that processes vision, sonar and radar using cameras, radar and ultrasonics (ultrasound) to navigate. A sensor that Tesla does not say they use but that also is frequently used is LIDAR, Light Detection and Ranging (or Light Imaging, Detection, And Ranging), which is an instrument principally consisting of a laser, a scanner and a specialized GPS receiver. The sensors are used to mimic what a human driver can see.

Some companies, e.g. Mcity, University of Michigan, and Volvo are working with technology that add interaction and/or communication making the software ‘better’ than a human driver. Mcity is testing V2V, vehicle-to-vehicle, communication that is wireless and make it possible for vehicles to share data, e.g. location, speed and direction. To make the communication technology working optimally it

is also installed in infrastructure, vehicle-to-infrastructure, e.g. traffic lights. The technology is expected to allow cars '[…] to see beyond what is immediately in front of them – sensing a red light around a blind curve, or automatically braking for a car that runs a stop sign.' (Fingas 2017)

'In a typical driverless vehicle prototype, cameras, lidar and radar devices serve as "eyes." But these sensors can't detect obstacles beyond their line of sight–like the stopped car behind the blind curve, for example. Connected technology changes that.' (Moore 2017)

AIs who plays games

To begin with two central terms needs to be described, game tree and the minimax algorithm.

A game tree is a representation of the positions in the game, the game states, presented as nodes in the tree and possible actions as edges. The tree starts at the root, which represents the state at the beginning of the game, followed by the first level representing the possible states after the first move, followed by the second level representing the possible states after the second move, etc. Monte Carlo Tree Search (MCTS) focuses its analysis on the most promising moves by running many game simulations.

> 'At first, the simulations are completely random: actions are chosen randomly at each state, for both players. At each simulation, some values are stored, such as how often each node has been visited, and how often this has led to a win. These numbers guide the later simulations in selecting actions (simulations thus become less and less random). The more simulations are executed, the more accurate these numbers become at selecting winning moves. It can be shown that as the number of simulations grows, MCTS indeed converges to optimal play.'
>
> (Burger 2016)

The minimax algorithm states that at 'each game turn, the AI figures out which move would minimize the worst-case scenario' (Burger 2016). The algorithm is useful when, and requires that, the complete game-tree is known.

Go, Ms. Pacman, Gathering game, and Wolfpack hunting game

For a very long time the game Go was considered to be a game not possible for a computer to conquer, the background being the complexity of the game and the amount of possible moves. 'There's limited

data available just from looking at the board, and choosing a good move demands a great deal of intuition.' (Byford 2016)

In March 2016 Google's AlphaGo, a version of DeepMind, beat Lee Se-dol, the second-highest ranking professional Go player in a match of five games of which AlphaGo won four. In Go there are more possible moves than in chess at each state, a wider branching factor, making it more difficult to search the game tree for a sufficient depth, and it has turned out to be more difficult to design evaluation factors for Go. The approach therefore had to be another than with Chess where the program searched the game tree as far as possible, usually to a depth of six moves or more, and after this an evaluation evaluated the quality of the nodes at that depth. The evaluation function then replaced the subtree below that node with a single value summarizing this subtree after which it proceeded in a similar way as the minimax algorithm and adopted the move that led to the least bad worst-case scenario.

AlphaGo consists of two different components, a tree search procedure and convolutional networks (a special type of neural networks). The convolutional networks 'guide' the tree search procedure, and they are learned. A total of three networks were trained; two policy networks and one value network, all using the current game state as input.

> 'The value network provides an estimate of the *value* of the current state of the game […]. The input to the value network is the whole game board, and the output is a single number, representing the probability of a win. The policy networks provide guidance regarding which action to choose, given the current state of the game. The output is a probability value for each possible legal move (i.e. the output of the network is as large as the board). Actions (moves) with higher probability values correspond to actions that have a higher chance of leading to a win.'
>
> (Burger 2016)

First the policy networks was trained on positions from games played by human experts, then in order for the program to be able to win, and not only predict moves from its opponent, the networks played against each other. The outcome of the played games was used as a training signal (deep reinforcement learning). The value network was also trained on, 30 million, game positions, and the final result of the training is described as special: '[…] it suggests a mixture of intuition and reflection. The value network provides the intuition, whereas the simulation result provides the reflection.' 'AlphaGo uses a mixture of the output of the value network and the result of a self-play simulation of the fast policy network: value of a state = value network output + simulation result.' (Burger 2016)

The classic arcade-game Ms. Pacman, by Atari, was one of the latest hindrances for game-playing AIs

to overcome. By the beginning of June 2017 Microsoft/Maluuba's AI[20] scored the ultimate top-score, 999 990 points. In Ms. Pacman the goal is to obtain points by eating pellets and at the same time avoid being submerged and killed by ghosts. If the player eat one of the special power pellets the ghost turn blue for a short time, and the ghosts can then be eaten, which award the player with extra points. At each level of the game there are bonus fruits that can be eaten for extra points. To get to a new level all the pellets has to be eaten. The win is described as unique due to the way the AI works. Linn quotes one of the research managers, Harm van Seijen, '[…] the best results were achieved when each agent acted very egotistically – for example, focused only on the best way to get to its pellet – while the top agent decided how to use the information from each agent to make the best move for everyone.' (Linn 2017) The programming-architecture used approximately 160 different agents, each focused on one specific task, and each working in parallel with other agents. In this way the overall task of mastering Ms. Pacman was divided into small pieces. On 'top' a central agent, 'oracle', controlled the movements of Ms. Pacman, and this agent's task was to decide how to use the information from each subagent in order to make the best (safest and most rewarding) move for Ms. Pacman, which was done by assigning each object, a pellet, a ghost or a fruit, a fixed weight, where the pellets and the fruit were given positive weight and the ghosts negative weight. 'There's this nice interplay between how they have to, on the one hand, cooperate based on the preferences of all the agents, but at the same time each agent cares only about one particular problem, […]. It benefits the whole.' (van Seijen as cited by Linn 2017)

The Maluuba AI uses Hybrid Reward Architecture, HRA, which takes as input a decomposed reward function and learns a separate value function for each component reward function. Because each component typically only depends on a subset of all features, the overall value function is described as smoother and easier to approximate by a low-dimensional representation, which enables effective learning. (van Seijen et. al 2017).

Google's AI, DeepMind, has been used in several other game playing and learning sessions. During the beginning of 2017 researches put DeepMind to test with a fruit Gathering game and a Wolfpack hunting game. The goal of the test was to investigate DeepMind's willingness to cooperate with others. Both games have a level of cooperation and the outcome of the game depends on the player's willingness to either act alone or to cooperate. The researchers describe their choice of games and outset as follows; 'We introduce sequential social dilemmas that share the mixed incentive structure of matrix game social dilemmas but also require agents to learn policies that implement their strategic intentions.' (Leibo, Zambaldi, Lanctot, Marecki & Graepel 2017, pp. 1)

In both games two AI agents were asked to compete against each other. In the Gathering game the

[20] Maluuba is Montreal-based company acquired by Microsoft in the beginning of 2017.

agents should try to gather as many apples as possible. The outcome was that everything went smoothly as long as enough apples were around, but when the supply of apples became sparse the agents turned aggressive, using laser beams to knock (tag) the other out of the game for a set period. No extra points, or rewards, were given when an agent knocked the other over, nothing more than the remaining one bought itself time to gather apples without inference.

In the Wolfpack hunting game the goal was to catch a prey (a third player). Reward was given when the prey was captured. If more than one wolf caught the prey reward was proportionately given to all wolfs in the capture radius. The prey was seen as dangerous, a lone wolf catching the prey was at risk losing the carcass to scavengers, and players working together had a better chance of protecting it, and get a higher reward. Different to in the Gathering game the agents opted for collaboration, either one agent found the other and together started the hunt of the prey, or one agent found the prey first and waited until the other agent arrived before starting the hunt.

The researches also noted a difference between less complex iterations of DeepMind and more complex forms. The more complex the agent was the more prone it was to use violence in the Gathering game and opt for cooperation in the Wolfpack game. (Science Alerts 2017) The behaviours of the agents are described as humanlike in an interview with one of the team researchers, Joel Z Leibo.

> ‘This model also shows that some aspects of human-like behaviour emerge as a product of the environment and learning. Creating AI agents that co-operate with others could lead to systems that can develop policies and real-world applications, he continued.’
>
> (Burgess 2017)

AI today, and in the expected future, and Harry Collins' demarcation line between humans and AIs

Collins' arguments against the possibility of creating an AI follow a red tread of tacit knowledge and socialization. In *Artificial Experts* the un-crossable line between human and AI was drawn more or less between explicit and tacit knowledge. In *Tacit and Explicit Knowledge* Collins redefines his arguments and the line is placed 'in' tacit knowledge between motoric tacit knowledge, MTK, and collective tacit knowledge, CTK. He does seem to make a reservation that he does not think it in any foreseeable future will be possible to make CTK explicit and accordingly for a machine/computer to gain the ability of CTK. (Collins 2010, pp. 138)

Let me begin with the current state of AI-research and put it in relation to Collins' thoughts, before moving on to Collins' epistemological theory and the possible weaknesses I have noted in it, and finally closing by connecting with his critique of the possibilities of AI by, in parts, reviving the epistemological chicken debate. By the end of this section it should become evident if Collins' arguments stand the test of todays AI-innovations, and if not which consequences the possible flaws in his arguments have for his epistemological theory.

Socialness and AI

The minimal embodiment thesis is Collins' safeguard against a general AI. Even if we are able to make CTK explicit and transfer it to software the AI will not be able to fulfil the minimal embodiment thesis. If the minimal embodiment thesis is not to be taken literal, for example by saying that a brain does not have to consist of the same 'material' as a human brain does, but harbour the same functions and abilities, then the point of departure for the AI-project is brighter[21].

The social embodiment thesis, stating that only an entity with the human kind of brain, one that is associated with the language speaking species, has the ability to become social parasites, blocks off the AI project almost as definite. But here a loophole might be available, and that is because of the focus put on language speaking species. If we create a species that is able to speak a language, as a human, and having the flexibility in its adaptation to different social and cultural settings, CTK, then we in theory could be able to create an AI. But this requires that we skip the minimal embodiment thesis or at least does not read it literally to need a human brain.

[21] This assumption would place Collins very close to the theory proposed by Searle that if we solve the problem of how consciousness comes about, then we in principle would be able to create an AI.

With this said only the minimal embodiment thesis can be read as the demarcation line between a human brain and an AI. The minimal embodiment thesis does have implications for Collins epistemological thesis though, which I will return to shortly. Of importance here is that the social embodiment thesis appears to be able to 'overrule' given that we are able to make CTK explicit. For Collins both thesis are important since they together, as well as the connection between them, safeguard against the threat of moving goalposts where the success of the AI-research e.g. is said to have reached the goal of a general AI just because the AI is able to mimic certain aspects of tacit knowledge. At the same time Collins makes sure that human intelligence is not 'just' the matter of ability, CTK, and neither a part of the human body but a complex relation between the physical pre-condition(s) and something that according to him cannot be rationalised.

If we overlook the minimal embodiment thesis anyway and take a closer look at the next obstacle, socialness: How close are AI-researchers to crack that code? Lets start by recapturing Collins thoughts about strings. Collins holds that it is in the fourth way strings are able to affect entities, 'a string can communicate by being interpreted as meaningful by a human', that is resided for humans. He declare that computers are said to primarily occupy the third way, 'a string can communicate "mechanically," as when a new piece of code is fed into a computer or a human reacts to a sound in a reflex-like way' (Collins 2010 pp. 16–17), but that computers often are mistaken to be able to occupy the fourth way as well.

In both the third and fourth way communication is vital, which leads to the assumption that Collins thinks that computers in some aspect are able to communicate. Although not in the same ways as humans due to them not being able to fulfil his five conditions for communication, and especially not the fifth condition; 'The transfer of a string plus a significant flexible and responsive physical change in the entity gives rise to a communication'. The fifth condition has as pre-condition a fluency in (human) language, which is described as always changing in relation to changes in society. The fluctuating nature of language therefore defines condition five as a flexible ability not possible to be fixed, which is another important section of Collins' argument against the possibility of AI, because the only ways for condition five to take place is through socialization.

With addition of the three stages language is translated Collins ads further weight to his arguments, stating that he does not believe the third stage of translation, 'the attempt to recreate meaning from the string—to interpret it', is reachable for a computer since they 'only' deal with transformation of strings, and strings are said to be 'just a physical thing' whereas language is believed to be full of meaning. What computers are able to do though are to mechanical mimic acts that humans perform self-consciously as well as unself-consciously.

An example of a situation where Collins thinks the flexibility of the knowledge, his fifth condition of communication, is vital is when a car-driver navigates in traffic. This ability is of fundamental importance to the AI-field of autonomous vehicles. Today vehicles are said to nudge at level four of automation and are anticipated to sometime in the near future reach level five, full automation. If autonomous vehicles reach level five, and if the vehicles are able to communicate with the infrastructure, including pedestrians, cyclists, etc. in a sufficient way then this could mean that they are able to fulfil Collins' fifth condition of communication. Then it can be argued that the autonomous vehicles in their limited world could achieve a level of socialness between each other, interchanging information and being affected by their 'society', even if it is not a human way of interaction. They would then be social parasites in a autonomous vehicles' world, using a signal-based language that might be possible to translate into human language. Would this do for a general AI? Well I would say no, due to the responses from the car not being self-learning neither possible to act flexible more than under very strict rules. These limitations of the autonomous vehicle will be placed on them by the developers, and regulated by laws, all to make sure the vehicles are safe for humans. If the strict control, laws and regulations of the industry are overlooked it could be imagined that the vehicles started to use their communication skills to help each other, tipping each other where the police controls are so that the car 'behaves' appropriate, taking shortcuts on pedestrian streets and in other ways breaking rules when estimated safe, and they had something to win (as human car drivers do today). But hopefully we will never meet an anarchistic self-driving car deciding to break the rules and do a burnout just for the fun of it on the freeway. So if the regulations and traffic rules are correctly, and without bugs, implemented in the algorithm, and the vehicles self-learning abilities are set to a minimum, and very limited, then an autonomous vehicle is not threatening Collins argument even it might be able to act as if it is.

DeepMinds game-playing AIs' learning in the Gathering and the Wolfpack games appears to be shifting some ground as well. What has happened, if only at a very modest level, is that the AIs' approach and solutions create rules of their own and act in ways that can be perceived as humanlike. The language part is here lacking in an articulated form, but in the Wolfpack game the agents performed a basic communication when deciding to hunt as a team, a behaviour that the AI-agents had learned via trial and error. The important aspect is that the AI-agents learned, they were not instructed to behave in a collaborative, or non-collaborative way.

Collins considers meaning in the context of language to be something 'that relates to the changing ways people live in society', and it can be argued that at a very basic level this is what the AI-agents in these games do when they react to increase and decrease in assets, apples or points given when not losing the prey to scavengers.

Changes in society, adaption and AI

Collins hold changes in society e.g. use of language, cultural differences, etc. to be an obstacle that an AI will not be able to handle. With autonomous vehicles the example with vehicles navigating in different traffic-scenarios probably will be solved and accordingly his example will loose its correctness. The industry is not fully there yet though, and even if the industry beam out how close they are to level five of automation they still have to completely master level four, and obstacles might occur along the way.[22] DeepMind has proved to be able to adapt to the (game-)environment and to change behaviour, without any set rules of how to, by it self, and its actions has been described as humanlike, and to have emerged as '[…] a product of the environment and learning' (Leibo 2017 as cited by Burgess 2017).

During the summer of 2017 the media reported of Facebook deciding to close down one of their AI projects after two AI-programs appeared to talk to each other in a self-constructed language.[23] Facebook's project is one of several working in the area of language and AIs. At OpenAI another project has been conducted, and these AIs are encouraged to find ways of communicating in order to improve their skills.

> 'Our hypothesis is that true language understanding will come from agents that learn words in combination with how they affect the world, rather than spotting patterns in a huge corpus of text. As a first step, we wanted to see if cooperative agents could develop a simple language amongst themselves.'
>
> (Abbeel, Mordatch, Lowe, Gauthier & Clark 2017)

The way the work is conducted is that it aims for the AIs to invent a simple language, which is grounded and compositional. With 'grounded' meaning that the words are tied to something 'directly experienced by a speaker in their environment', and compositional meaning that the speakers 'can assemble multiple words into a sentence to represent a specific idea' (Abbeel, Mordatch, Lowe, Gautier & Clark 2017).

To train the AI's multi-agent reinforcement learning problems are used and the experiments are represented as cooperative rather than competitive. OpenAI's project is interesting especially with Collins' hesitation regarding the possibility of AIs and socialization in mind, but the language-projects still have to post more reports of progress before they can be taken into account.

[22] Volvo has for example noted that Volvo's autonomous cars have problems with kangaroos. (Zhou 2017)

[23] What happened in short was that the programs had been given the task of negotiating between themselves and to improve their bartering. During this learning process the AIs developed a shorthand language that was not readable for others than the chat-bots. (Griffin 2017)

Changes in society require the flexibility that Collins includes in CTK and the examples above can be interpreted as having at least a start of CTK. In *Tacit and Explicit Knowledge* he present an example 'Badly broken text that is and isn't easily repaired', and writes that there are three remarkable things with humans being able to 'repair' the broken text and understand what it says.

> 'The first remarkable thing is that most readers can read the second passage so easily and the overall demonstration shows that the reading is accomplished via meaning. The second, still more remarkable thing is that, though my spell- checker highlighted almost every word with a jagged redline indicating a mistake, the copy editor of this book will not even think about correcting either paragraph. That is an effect of understanding meaning at a still higher level—the meaning of a whole group of paragraphs. The third remarkable thing is that you, dear reader, have actually undertaken to try to repair both passages, known when it was time to give up in the case of the first passage, and known that you should persevere in the case of the second passage in spite of any initial difficulty.'
>
> (Collins 2010, pp. 115–116)

Collins argues that humans have the ability to do all three things but things around us cannot do this kind in respect of us. With computers, pocket calculators, etc. he writes that these are 'social prostheses', meaning 'entities that can take the place of a social being because the rest of the social organism fills in the gaps' (Collins 2010, pp. 116).

This brings to mind NEIL, Never Ending Image Learner. NEILs work is to automatically extract visual knowledge from Internet data, and the effort is to 'It is an attempt to develop the world's largest visual structured knowledge base with minimum human labelling effort.'(Chen, Shrivastava & Gupta 2013) It is possible to image a similar database with words, which would be able to provide an unsupervised AI with a basic knowledge similar to that humans learn during childhood. The AI would have to keep the knowledge up-to-date and flexible by interacting with its environment. This could be looked upon as a brute force way of knowledge since the AI would have all the world's interpretations of a single word available, and it would probably use enormous amounts of processing power. But if the brute-force-argument is put to a side then the relation between the AI, and the word-knowledge-database has a similar relation as the asymmetry-relation between a human and the things around us that Collins holds is de facto.

Reviving the epistemological chicken debate and the implications of the current state of AI on Collins' epistemological theory

During the epistemological chicken debate Bruno Latour and Michel Callon described Collins' and Steve Yearley's claim to one of a tug-of-war between natural realism and social realism, and that scientists should alternate to social realism when they play the role of sociologists explaining science from natural realism when they are scientists (Latour & Callon 1992, in Pickering, pp. 346). Collins and Yearley confirm this in their reply when they provide their 'prescription' in the world of philosophical insecure objects namely to 'stand on social things–be social realists–in order to explain natural things' (Collin & Yearley 1992, in Pickering, pp. 382).

According to Latour and Callon Collins and Yearley think that any 'sociologist who stops being a social realist would be a traitor, since he or she would abandon the fight, or worse still, help out the other side' (Latour & Callon 1992, in Pickering, pp. 346). Latour and Callon hold that they are viewed as traitors of this kind because they 'give back to nature the role of settling controversies.' (Latour & Callon 1992, in Pickering, pp. 346)

To be able to refute the possibility of creating a general AI Collins places the minimal embodiment thesis and the social embodiment thesis as the demarcation line between the human body/brain and the potential artificial intelligence. The thesis' constitute the line that no computer program will be able to cross and therefore is a general AI not possible. It is not an option, as I have showed earlier, to skip the embodiment thesis, if we do the theory starts to weaken and the demarcation line Collins want to draw between humans and AIs cannot be firmly placed.

With his definition of the thesis' Collins seems to commit the same 'treason' of which he accused Latour and Callon of committing, namely of bringing natural realism into the sociological realm (of knowledge). The reason being that the minimal embodiment thesis presupposes that (human) knowledge starts in human biology, in the physical constitution of a human. Whether it is only the human brain or to what extent other parts of the human body are prerequisite is left open. Also the social embodiment thesis holds a pre-requirement of the human biological setting, the need of a minimal human body. The implication is hereby that Collins is taking a naturalistic point of departure, before continuing his line of arguments by distancing himself from any, and all, naturalistic explanations by trying to make sure that all arguments are socially anchored.

When the two thesis' are placed in the context of Collins' epistemological theory it is difficult to overlook this 'slip' towards naturalism. The 'slip' consisting in the founding of knowledge in physical pre-deposition of the human body and brain rather than, as Collins empathizes in social interactions

and the collectivity. With the minimal embodiment thesis, but also the social embodiment thesis, Collins shift his position by taking a step a closer to natural realism, which make him dependent on the natural sciences if his demarcation line between a human brain and a AI shall stand solid. In the context of Collins epistemological theory this implies that human knowledge is founded in, and can be explained by, natural science.

This is not the only part that appears unstable in Collins thesis', and which brings the epistemological chicken debate to mind. If we hold that AIs are able to produce, contribute, or even create knowledge, as can be alluded with the quote from the Go-community below, then it seems as if AIs are more than mere things, and that the dichotomy that Collins hesitate between people and things (Collins & Yearley 1992, pp. 386) is not solid.

> 'But AlphaGo could also open up new avenues for the game. Members of the Go community are as stunned with the inventive, aggressive way AlphaGo won as the fact that it did at all. "There were some moves at the beginning – what would you say about those three moves on the right on the fifth line?" American Go Association president Andy Okun asked VP of operations Andrew Jackson, who also happens to be a Google software engineer, at the venue following the match. "As it pushes from behind?" Jackson replied. "If I made those same moves…" Okun continued. "Our teachers would slap our wrists," Jackson agreed. "They'd smack me!" says Okun. "You don't push from behind on the fifth line!" "We're absolutely in shock," said Jackson.
>
> "There's a real question, though. We've got this established Go orthodoxy, so what's this going to reveal to us next? Is it going to shake things up? Are we going to find these things that we thought were true – these things you think you know and they just ain't so?" '
>
> (Byford 2016)

Collins holds that things may contribute to our social experience, but that the things are not able to absorb anything equivalent to my social experience. (Collins 2010, pp. 117) With autonomous vehicles this is happening, at present in portions since the vehicles still have the possibility for the human driver to interfere. Quoting Tesla's web-page on the feature 'Full Self-Driving Capability' that today can be ordered on a new Tesla;

> '[…] enabling full self-driving in almost all circumstances, at what we believe will be a probability of safety at least twice as good as the average human driver. The system is designed to be able to conduct short and long distance

> trips with no action required by the person in the driver's seat. […] All you will need to do is get in and tell your car where to go. If you don't say anything, the car will look at your calendar and take you there as the assumed destination or just home if nothing is on the calendar. Your Tesla will figure out the optimal route, navigate urban streets (even without lane markings), manage complex intersections with traffic lights, stop signs and roundabouts, and handle densely packed freeways with cars moving at high speed. When you arrive at your destination, simply step out at the entrance and your car will enter park seek mode, automatically search for a spot and park itself.'
>
> (Tesla 2018)

For the car to be able to perform in 'Full Self-Driving Capability' it have to able to absorb the surroundings, adapt to traffic rules as well as other drivers mistakes and navigate different environments (from snow to traffic jams). The 'Full Self-Driving Capability' also has capabilities that are similar to a digital personal assistant as it is able to 'read' your (digital) calendar and act from what is written in it.

In a recent report from the DeepMind team the researchers present that they have been able to develop an AI that have capabilities which resembles the navigational abilities of a human, and that in the future could be used in autonomous vehicles. The AI has at present only been tested in maze-environment, but the researches write that their findings are 'providing a foundation for proficient navigation' (Banino, Barry & Kumaran 2018).

The autonomous vehicles driving on the streets today do not talk to each other, and they will most likely not 'talk' to each other in any human language, but their communication will be possible to translate into a human language. The cars will probably not be able to interact in an epistemological discussion, or write a paper on a subject, but they will be able to communicate between themselves, and the infrastructure. This communication will be of meaning to them to the extent of their task.

With DeepMind and the results from the Gathering and Wolfpack game the case of learning and knowledge is slightly different. DeepMind agents are not provided with a set of rules but instead they '[…] learn for themselves to achieve successful strategies that lead to the greatest long-term rewards. […] Also like a human, our agents construct and learn their own knowledge directly from raw inputs, such as vision, without any hand-engineered features or domain heuristics.' (Silver 2016)

Even if we grant that Collins is correct in his thought that 'things' 'do not participate in the world of language in the way that humans participate; they participate only in the world' (Collins 2010, pp. 117), and regard AIs as things, they will most likely impact not only the way we humans experience

society but also our knowledge of the world, ourselves and knowledge as such. Collins fundamental thesis that knowledge rests in society, the crucial role of socialisation and how experts come to shape then weakens. If (human) society invites the ‘knowledge’ produced, and even created, by AIs and adapts it into the human way of for example playing a game is it then ‘knowledge’ created by the human as s/he interprets the output created by the AI? Or is it knowledge created by the AI? If the answer is that the AI is granted a contributional status, then there are resemblances to the thoughts presented by Actor Network Theory where things play a participating role in the construction of knowledge.

Conclusion

The stake I set out with this paper was to investigate and clarify if, and how, the demarcation line that Collins drew between human and AI has been moved or maybe even disappeared. The questions asked were whether Collins arguments still are valid and, if not, if his epistemological thesis has been affected. The focus has, in line with Collins arguments, been put on the *cannot* argument stating that a general AI cannot be developed, and this is why it can be empirically tested since Collins hold that an AI will not be able to absorb anything equivalent to human social experience.

To say that an AI will substitute a human being in the future might seem daring. If Collins is right, and if by principle the human level/general/strong AI cannot be constructed, then there is also no cause for concern, or to take any precautions against the replacement of humans by machines, or to (re)direct research in a direction that will safeguard against unfortunate events more than from a functional perspective.

Collins critique against the possibility of a general AI is well formulated and founded in his epistemological theory, and on the surface it appears solid, but there are fatal weaknesses. Collins needs to safeguard against the threat of moving goalposts if he is to withhold his ontological stand. He does this by introducing the minimal embodiment thesis, together with the social embodiment thesis. With these thesis' it will never ever be possible to create a humanlike artificial intelligence, no matter how efficient and advanced any computer or software ever become.

When Collins clutches to the minimal embodiment thesis he takes a step away from social constructivism, upon which his epistemological theory rests, and says that the human uniqueness is founded in the physical constitution of the human body (and brain). He hereby moves his epistemological point of departure from social constructivism to(wards) naturalism. The reason being that Collins holds that for both the thesis' a human body, at least a human brain, is required. Without the physical pre-requirement tacit knowledge, which only humans have (CTK at least) is not possible, and tacit knowledge is the foundation for socialization and all epistemological claims that Collins make in his theory.

It appears clear to me that with the minimal embodiment thesis Collins assumes that human knowledge starts in the physical constitution of the human body and brain. Without the 'flesh and blood', to some extent, the knowledge that Collins hold to be uniquely human will never be possible to (re)construct. With socialization, withheld by Collins as an essential part in his argument against AI, two aspects needs to be settled for the socialization problem to be solved. The first focusing on how humans approach and handle questions and answers, where Collins thinks that humans unfold context,

and in doing so need to be sensitive to what is going on around them which it is handled by e.g. asking and replying to questions. In a rudimentary way that is the same as what digital personal assistants do, and a more advanced example is OpenAI's approach to learning (a) language. Both use questions and answers in order to function well and be of use. The difference being that the AIs do not by themselves define the objective and what is of use and non-use, this is done by the human researchers. Facebook's AIs, which are supposed to have started to invent their own language ('shorthand') in order to be better at their task seems to be different. But even though it appears as if the chat bots did invent something on their own they were designed to do a task, and that is what they did, but since they were not programmed to communicate in English they didn't. (Simonite 2017)

The second aspect put focus on 'the right way to do things', and Collins think that this is only possible to capture through experience, which is something that can vary from country to country and/or cultural contexts. Similar to with the first aspect I hesitate that this is not different to what takes place in certain AI-applications e.g. the development of autonomous cars as well as in the above-mentioned examples with digital personal assistants and OpenAI.

Collins think that language is a vital part of socialization and hesitate that what humans can, and an AI cannot, is to interpret meaning within a natural language. Unlike Collins it seems to me that what artificial intelligence software and applications do, on a very basic level, is similar; they interpret, and translate, something said into some form of action.

Collins believe that physical activity is an important part in acquiring linguistic fluency, and learning 'the right way to do things', and it is only through socialization that his fifth condition for communication can take place; 'The transfer of a string plus a significant flexible and responsive physical change in the entity gives rise to a communication'. It is defined as a flexible ability not possible to be fixed, and has fluency in language as its pre-condition, something that always is changing in relation to changes in society. With the discussed examples of AIs together with the different ways of learning that is used in training AIs it appears to me as if the AIs are socialized into the environment in which they act, but they do not act on their own objective, or if they are interpreted as doing so it is in a very limited environment (as with the Wolfpack and Gathering game).

AIs in 2018 are not general AIs. Some tasks that the AIs do are better and more accurate done by them than by a human, but these are tasks performed in a closed environment, e.g. a virus program aiming at detecting potential data breach. In these enclosed specific areas, with its limited tasks, the AI is more of an 'expert' than a human trying to do the same work.

I do not find Collins demarcation line, or critique solid enough, and it does have implications for his epistemological theory. If Collins demarcation line is resting on the minimal body thesis then this is a

naturalistic foundation, a foundation in the physical constitution of the human brain and body. Human knowledge is then founded in how the cells, neurons, blood vessels, soft tissue, etc. is put together and functions, and social constructionism can be described and defined by the natural sciences. If knowledge is gained, defined and constructed through socialization as Collins withhold, then the physical body and brain are not but 'mere' matter of physical material that, although from what we know, works as merely a possible constitution for knowledge and not necessarily the only one(s).

If I return to the quoted reflection regarding Go and the AlphaGo's way of playing when it won. What happens after the match is that human Go-players are able to reflect and discuss the development of each game, and to put them in a social context. If AlphaGo had lost it also would have learned from it's mistakes but what the Go community reflect upon, is that AlphaGo actually breaks the 'social codex' of the game. AlphaGo does not break the rules of Go but what is socially accepted as an elegant way to play the game. This is from my point of view a creative solution, and method that AlphaGo uses, one where it finds a loophole that human Go-players have been 'taught to ignore' through their socialization into being a GO-player.

The creative way used by AlphaGo resemble Wittgensteins example with a pupil who is presented with the task of continuing a series of numbers. The pupil does not continue the series of numbers as the teacher 'meant', instead the pupil continue as s/he thinks is correct. The reason AlphaGo, just as the student in Wittgensteins example does not go the route the Go-community, or the teacher, 'predicted' is because AlphaGo had not been trained to think in a pre-defined route. To use Wittgensteins example and words; 'These people are so trained that they all take the same step at the same point when they receive the order "+3".' (Wittgenstein 1953, pp. 82) It is not the way the other player (and the Go-community) had been trained (socialized) but it is the way the AI (just as the pupil) found accurate within the game and the prevalent rules of the game. The way humans learn to complete a series of numbers, to be scientist, or a golf player is through 'playing the game', through being part of society. AlphaGo used resources not available to the Go-community due to their social training into being a Go-player. AlphaGo found a solution that did not break the rules, and led to the goal in an unconventional way, a way that required improvisation, something Collins has described as a necessary part of CTK, which he hold only humans have and an AI cannot gain.

Keeping Collins theory in mind the way AlphaGo's style of playing was received by the Go-community, especially the questions raised do imply that the community is influenced, and that it will (or at least can) have impact on the cultural setting of the Go-community, in which it acted. In this way AlphaGo's playing style will (or at least can) be part of the way future Go-players address the game, and maybe even how they are trained and play the game.

Remaining for future investigations are the following questions; Could Collins epistemological theory be developed in order to show and become an argument against the possibility of a general AI? If not what implications does Collins epistemological theory have for at future inhabited by humanoid robots and humans? Since Collins arguments against the possibility of a general AI is not solid enough, is there something that he has missed? Could a solid theory against AI be formulated, one using a constructionist approach?

References

Abbeel, P., Mordatch, I., Lowe, R., Gauthier, J. & Clark, J. (2017). *Learning to Communicate*. [blog post] OpenAI blog. Available at: https://blog.openai.com/learning-to-communicate/ [Accessed 28/5/2018].

Ann Arbor Connected Vehicle Test Environment (2018). [online] The Regents of the University of Michigan. Available at: http://www.aacvte.org [Accessed 28/5/2018].

Baker, R.D. (2017). *Driverless milestone: No Hands Across America.* [online] San Francisco Chronicle. Available at: http://www.sfchronicle.com/business/article/Driverless-milestone-No-Hands-Across-America-11278241.php [Accessed 28/5/2018].

Banini, A., Barry, C. & Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. [online] *Nature*, 557 pp. 429–433. Available at: https://www.nature.com/articles/s41586-018-0102-6 [Accessed 30/5/2018]

Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). The Sociology of Expectations in Science and Technology. [online] *Technology Analysis & Strategic Management*, Vol. 18. Nos. 3/ 4. pp. 285–298. Routledge Taylor & Francis Group. Available at: http://dx.doi.org/10.1080/09537320600777002 [Accessed 28/5/2018].

Bostrom, N. (2014). *Superintelligence Paths, Dangers, Strategies*. New York: Oxford University Press.

Brundage, M., Avin, S. & Clark, J. et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Available at: https://maliciousaireport.com [Accessed 28/5/2018].

Buchanan, G. B. (2005). A (Very) Brief History of Artificial Intelligence. *AI Magazine* Vol. 26 Nr. 4. Association for the Advancement of Artificial Intelligence (AAAI). Available at: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1848/1746 [Accessed 28/5/2018].

Burger, C. (2016). *Google DeepMind's AlphaGo: How it works.* [blog post], Tastehit. Available at: https://www.tastehit.com/blog/google-deepmind-alphago-how-it-works [Accessed 28/5/2018].

Burgess, M. (2017). *DeepMind's AI has learnt to become 'highly aggressive' when it feels like it's going to lose.* [online], The Wired. Available at: http://www.wired.co.uk/article/artificial-intelligence-social-impact-deepmind [Accessed 28/5/2018]

Byford, S. (2016). *Why is Google's Go win such a big deal?* [online], The Verge. Available from: http://www.theverge.com/2016/3/9/11185030/google-deepmind-alphago-go-artificial-intelligence-impact [Accessed 28/5/2018].

Campbell, M., Hoane Jr., A. J., Hsu, F-h. (2002). Deep Blue. *Artificial Intelligence* 134, pp. 57–83. Netherlands: Elsevier Science.

Charles Babbage. (2017) [online] Available at: http://www.charlesbabbage.com [Accessed 29/5/2018].

Chen, X., Shrivastava, A. & Gupta, A. (2013). NEIL: Extracting Visual Knowledge from Web Data. [online] *International Conference on Computer Vision (ICCV)*, December, 2013. Available at: https://www.ri.cmu.edu/pub_files/2013/12/iccv13.pdf [Accessed 29/5/2018].

Collins, H. M. (1990). *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge: The MIT Press.

Collins, H. M. (1992). *Changing Order*. Chicago: The University of Chicago Press.

Collins, H. M. & Yearley, S. (1992). *Epistemological Chicken*. In Pickering, A. Science and practice. Chicago, London: University of Chicago Press. pp. 301–326.

Collins, H. M. & Yearley, S. (1992). *Journey Into Space*. In Pickering, A. Science and practice. Chicago, London: University of Chicago Press. pp. 369–389.

Collins, H. M. (2010). *Tacit and Explicit Knowledge*. Chicago: The University of Chicago Press.

Davies, R. & Hern, A. (2017). *Apple chief: driverless car venture is 'the mother of all AI projects'*. [online], The Guardian. Available at: https://www.theguardian.com/technology/2017/jun/13/apple-self-driving-car-technology-tim-cook#img-1 [Accessed 28/5/2018].

Dreyfus, H. L. (1965). *Alchemy and Artificial Intelligence*. [online] Available at: http://www.rand.org/content/dam/rand/pubs/papers/2006/P3244.pdf [Accessed 28/5/2018].

Dupuy, J-P. (2000). *The Mechanization of Mind: On the Origins of Cognitive Science*. English translation. New Jersey: Princeton University Press.

Encyclopædia Britannica. (2018) [Online]. Available at: https://www.britannica.com [Accessed 28/5/2018].

European Parliament. (2017). A8-0005/2017. [report] Available at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN [Accessed 28/5/2018].

Fairclough, N. (1995). *Critical Discourse Analysis: The Critical Study of Language*. London: Routledge.

Fingas, J. (2017). *Self-driving cars are safer when they talk to each other*. [online], Engadget Available at: https://www.engadget.com/2017/06/21/self-driving-shuttles-university-of-michigan/ [Accessed 28/05/2018].

Griffin, A. (2017). *Facebook's artificial intelligence robots shut down after the start talking to each other in their own language*. [online], The Independent. Available at: https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html [Accessed 28/5/2018].

Hardesty, L. (2017). *Neural networks explained*. [online]. MIT News Office. Available at: http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [Accessed 28/5/2018].

Hofstader, D. R. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid.* New York: Basic Books Inc.

Hof. D. R. (2013). *Deep Learning – With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart*. [online], MIT Technology Review. Available at: https://www.technologyreview.com/s/513696/deep-learning/ [Accessed 28/05/2018].

Huang, B. (2015). *Bayesian Networks*. [online] CS5804 Virginia Tech Introduction to Artificial Intelligence. Available at: https://youtu.be/TuGDMj43ehw [Accessed 28/05/2018].

Jørgensen, M. & Philips, L. (2012). *Discourse Analysis as Theory and Method*. London: SAGE Publications Ltd.

Krishna, S. (2017). *Self-driving cars are safer when they talk to each other*. [online], Engadget. Available at: https://www.engadget.com/2017/06/24/self-driving-cars-mcity-augmented-reality/ [Accessed 28/05/2018].

Latour, B. & Callon, M. (1992). *Don't Throw the Baby Out with the Bath School! A Reply to Collins and Yearley*. In Pickering, A. Science and practice. Chicago, London: University of Chicago Press. pp. 343–368.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J. & Graepel, T. (2017). *Multi-agent Reinforcement Learning in Sequential Social Dilemmas*. [online], Cornell University Library. Available at: https://arxiv.org/abs/1702.03037 [Accessed 28/5/2018].

Linn, A. Divide and conquer: *How Microsoft researchers used AI to master Ms.* Pac-Man. [blog post], Microsoft. Available at: https://blogs.microsoft.com/next/2017/06/14/divide-conquer-microsoft-researchers-used-ai-master-ms-pac-man/#sm.01g84yys11mte3811qf2iaqpup508 [Accessed 28/05/2018].

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 2nd. Edition. Natick: A K Peters Ltd.

Moore, N., C. (2017). *Mcity demos: Self-driving cars can be even safer with connected technology.* [online], Michigan News University of Michigan. Available at: http://ns.umich.edu/new/multimedia/videos/24932-mcity-demos-self-driving-cars-can-be-even-safer-with-connected-technology [Accessed 28/05/2018].

Moral Machine. (2018). [online] Scaleable Cooperation. MIT Media Lab. Available at: http://moralmachine.mit.edu [Accessed 28/05/2018].

Natale. S., & Ballatore. A. (2017). Imagining the thinking machine: Technological myths and the rise of Artificial Intelligence. [online], *Convergence: The International Journal of Research into New Media Technologies* 1–16. Available at: http://journals.sagepub.com/doi/pdf/10.1177/1354856517715164 [Accessed 28/05/2018].

Nilsson, N. J. (2010). *The Quest for Artificial Intelligence.* New York: Cambridge University Press.

RAND Corporation. (2018). [online] Available at: http://www.rand.org [Accessed 31/5/2018].

RoboCup. (2018). [online] Available at: http://www.robocup.org [Accessed 28/05/2018].

Russell, S. J. & Norvig P. (2010). *Artificial Intelligence: A Modern Approach.* 3rd. edition. New Jersey: Upper Saddle River.

SAE International. (2014). *Automated driving*. [online] Available at: http://www.sae.org/misc/pdfs/automated_driving.pdf [Accessed 28/5/2018].

Science Alert. (2017). *Google's New AI Has Learned to Become "Highly Aggressive" in Stressful Situations.* [blog post], Science Alert. Available at: https://www.sciencealert.com/google-s-new-ai-has-learned-to-become-highly-aggressive-in-stressful-situations [Accessed 28/5/2018].

Searle, J. (1980). 'Minds, Brains, and Programs'. [online], *Behavioral and Brain Sciences*. Vol. 3, No. 3, pp. 417-457. Available at: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0140525X00005756

Searle, J. (2015). *Consciousness in Artificial Intelligence*. [online] Talks at Google. URL: https://www.youtube.com/watch?v=rHKwIYsPXLg [Accessed 3175/2018]

van Seijen, H., Fatemi, M., Romoff, J., Laroche, R., Barnes, T., & Tsang, J. (2017). *Hybrid Reward Architecture for Reinforcement Learning*. [online], Cornell University Library. Available at: https://arxiv.org/abs/1706.04208 [Accessed 28/5/2018].

Silver, D. (2016). *Deep Reinforcement Learning*. [blog post], Deep Mind. Available at: https://deepmind.com/blog/deep-reinforcement-learning/ [Accessed 28/5/2018].

Simonite, T. (2017). No, facebook's chatbots will not take over the world. [online] The Wired. Available at: https://www.wired.com/story/facebooks-chatbots-will-not-take-over-the-world/ [Accessed 30/5/2018].

Stilgoe, J. (2017). *What will happen when a self-driving car kills a bystander?* [online], The Guardian. Available at: https://www.theguardian.com/science/political-science/2017/jun/24/what-will-happen-when-a-self-driving-car-kills-a-bystander [Accessed 28/5/2018].

Spice, B. (2007). Meet the "Boss" [online], Mellon University. Available at: https://www.cs.cmu.edu/news/meet-boss [Accessed 28/5/2018].

Techopedia. (2018). [online] Available at: https://www.techopedia.com [Accessed 28/5/2018].

Tesla (2018). [online]. Available at: https://www.tesla.com [Accessed 28/5/2018].

Test Site Sweden. (2018). [online], Test Site Sweden. Available at: https://www.testsitesweden.com/en/projects-1/driveme [Accessed 28/5/2018].

Union of Concerned Scientists. (2018). *Self-Driving Cars Explained.* [online], Cambridge. USA. Available at: http://www.ucsusa.org/clean-vehicles/how-self-driving-cars-work#.WXce-RSPAYw [Accessed 28/5/2018].

Vincent, J. (2016). *What counts as artificially intelligent? AI and deep learning, explained.* [online], The Verge. Available at: https://www.theverge.com/2016/2/29/11133682/deep-learning-ai-explained-machine-learning [Accessed 28/5/2018].

Volvo Car Corporation. (2018). [online] Available at: http://www.volvocars.com [Accessed 28/5/2018].

Wikipedia (2018). *Deep learning* Wikipedia [online] Available at: https://en.wikipedia.org/wiki/Deep_learning [Accessed 28/5/2018].

Winfield, A., Jirotka, M. & Hutton, L. (2017). *Towards an ethical black box.* [online], University of the West of England & University of Oxford. Available at: https://www.cybersecurity.ox.ac.uk/site-resources/uploads/2016/02/Marina-Jirotka-Ethical-Black-Box.pdf [Accessed 28/5/2018].

Wittgenstein, L. (2009). *Philosophical Investigations*. 4th edition. Chichester: Blackwell Publishing

Zhou, N. (2017). *Volvo admits its self-driving cars are confused by kangaroos.* [online], The Guardian. Available at: https://www.theguardian.com/technology/2017/jul/01/volvo-admits-its-self-driving-cars-are-confused-by-kangaroos [Accessed 28/5/2018].