Johansson and Gothager, Final Thesis

Which Management Control System principles and aspects are relevant when deploying a learning machine?

Authors:

Martin Johansson: m.n.johansson@gmail.com

and

Mikael Gothager: mikael@gothager.se

Handin date: 2017-09-21

Abstract

How shall a business adapt its management control systems when learning machines enters the arena? Will the control system continue to focus on humans aspects and continue to consider a learning machine to be an automation tool as any other historically programed computer? Learning machines introduces productivity capabilities that achieve very high levels of efficiency and quality. A learning machine can sort through large amounts of data and make conclusions difficult by a human mind. However, as learning machines becomes even more complex systems, they introduce an uncertainty not previously considered by automation tools. The algorithms can make their own associations, and the automation engineer will no longer know exactly how a learning machine produces its outcome. What is the motive for a learning machine's decision? A learning machine in this context becomes more human like compared to the older generation of automation computers. This thesis concludes that most contemporary Management Control System principles are relevant when deploying machine learning, but some are not. A Management Control System must in contradiction to a historically programmed computer, consider multiple human-like aspects while controlling a deployed learning machine. These conclusions are based on empirical data from web-articles, TED-talks, literature and questionnaires directed to contemporary companies using machine learning within their organizations.

Acknowledgements

A special thanks to Roland Gustavsson at Wisegroup, and Dan Rubins at Legal Robot for answering our questionnaire with contemporary insights from the machine learning industry. Also a special thanks to Shahiduzzams Quoreshis for reviewing our thesis and giving us valuable feedback and improvement suggestion. And a last thanks to Tan Lujiao for opposing on our thesis.

Table of Contents

List of Figures and Tables

1 Introduction

The last decade's progress in machines learning technology introduces new exciting opportunities for businesses. The concept of machine learning is a branch of Artificial Intelligence (AI) and concerns machines that learns on their own. The obvious business opportunity is the likely productivity increase brought by the new technology, but it is also likely that completely new products and services will arise. The thesis will describe that a learning machine can sort through large amounts of data and make conclusions difficult for a human mind. However, as learning machines becomes more complex systems, they introduce an uncertainty not previously considered in automation tools. The algorithms can make their own associations, and the automation engineer will no longer know exactly why a learning machine makes a certain decision. What is the motive for a learning machines decision? A learning machine becomes more human like compared to the previous generation automation computers.

Consequently, a very interesting question concerns how a business should act to assure that the new learning machines still contributes in alignment with the business objectives and strategy. Management control systems (MCS) are developed to assure exactly that for both labour and capital intensive businesses. They do so with the focus either to control human's aspects or to control capital investments like automated machines. But how shall a business design its MCS to control learning machines? Will the MCS consider learning machines to have human aspects or to be more comparable with automation tools, just as any other historical computer? As the technology is quite new, academic MCS principles have not yet been adapted. The thesis authors have found no relevant academic research on this topic. Therefore the scope of this thesis becomes to analyze and conclude what aspects and principles of existing MCS are relevant when a business intend to control learning machines.

1.1 Background

The concept of Artificial Intelligence (AI) is defined by Luger (2005) as "the branch of computer science that is concerned with the automation of intelligent behavior". The related concept of "Machine learning" is referred to as a branch within Artificial Intelligence (AI) and will be further defined in the theory chapter of this thesis. "Learning machines" in this thesis is the reference name for products using the machine learning technology.

The driving force for a businesses to use machine learning to enhance their value creation is likely enormous. There are two main reasons. The first is *productivity increase* and the second is that learning machines enables *new services and products*. There are multiple references from literature supporting the potential *productivity increase* while using learning machines. Laurent, Chollet and Herzberg (2015) states that these concepts introduce productivity capabilities that achieve very high levels of efficiency and quality. They continue to state that businesses in all economic sectors will embrace these smarter concepts because of its game changing leap in productivity, throughout its entire processes or workflows. As shown in Bose & Mahapatra (2001), the finance sector has already found large positive NPV for various Machine Learning investments, where it could be used for more efficient prediction and detection of opportunities. Although an expensive investment, the financial sector had the economic muscles to seize this new technology in the early 2000. This new technology outperformed previously used statistical methods in terms of efficiency (Bose & Mahapatra, 2001). In Strategy and Marketing, where the fundamentals to wisely use scarce resources will remain, machine learning has already made an

impact on how market data can be used more efficiently. Bose & Mahapatra (2001) shows how Machine Learning was used for markets and customer classification, prediction and association already decades ago. Howard (2016) brings another productivity aspect, that learning machines will not need a constant salary to continuously perform a skilled task. This implies that the knowledge labor cost can be exchanged for an investment cost. Kelly (2016) states that artificial intelligence and learning machines will inevitably be introduced in nearly everything we do, because of human's nature and desire for making things smarter. Now, as the expectation is that learning machine investment cost will drop at a simultaneous performance increase, it is quite feasible that nearly all businesses will see a future economic efficiency increase when investing in learning machines instead of labor.

Concerning *new services and products*, learning machines already enables solutions not possible for human skills, and new businesses or workflows arises to profit from these new smarter and better technology solutions. An example of the later is cancer treatment diagnostics, where new software arise not only to replace work tasks previously performed by doctors, but with successfully proven results, machine learning does it in ways previously actually considered to be totally wrong (Howard, 2016) (Mousannif, Moatassime and Noel, 2016). Another new application is that learning machines can today be used for HR-hiring decisions, while evaluating psychological aspects of candidates, and it does so with higher quality compared to traditionally trained HR-labor (Tufekci, 2016).

Businesses today use management control systems which are mainly developed to control and motivate humans. Automation through computer systems is today considered an alternative control system, which helps humans remove boring and complex activities in a management control system (Merchant & Van der Stede, 2012). However, machine learning is here already today and large internet companies such as Google, Facebook, Amazon, Netflix and Linkedin already relies on learning machines to control central parts of their business such as picture recognition and connecting friends (Jeremy Howard, 2016). Simultaneously it is anticipated that the impact of possible errors, in for example trade strategy or data integrity, will increase as faster machines operate the activities and make decisions with considerably low intervention of humans. This anticipation awakens the thoughts on what the actual impact of machine learning algorithms might be (World Economic Forum, 2015).

1.2 Problem discussion

What implications lay ahead for future business management when we let learning machines do business decisions that affects business results? How shall a business lead learning machines, and how shall they be controlled? This chapter will discuss these questions with the intention to prepare for the *Problem formulation and purpose* chapter of this thesis.

To answer the above question, this thesis must connect the most important traits of learning machines technology with the latest knowledge of MCS's. As the machine learning technology is a relatively new phenomenon in business applications, the thesis authors can today find no reliable academic MCS description of how a business should control learning machines. Therefore, this thesis will take a holistic approach, i.e. the thesis will start at the highest level of *MCS principles* and make connections and conclusions to the latest knowledge of machine learning technology. The theory chapter therefore needs to describe the latest holistic knowledge of *MCS principles*. Merchant & Van der Stede (2012) defines "good" management control to be when "management is reasonably confident that no major unpleasant surprises will occur". But

they also highlight that the purpose of a MCS is to ensure that businesses are managed in line with strategic objectives, enabling a wanted behavior from the organization. Merchant & Van der Stede (2012) additionally defines several holistic principles and this knowledge together with recent journals will be the foundation of the case study design in this thesis.

To understand machine learning technology in the context of MCS, the theory chapter will additionally need to describe A*rtificial Intelligence, Machine Learning* and *Emotional Intelligence* more in depth. The main reason why these aspects are important is because when machines start to learn they will no longer behave as traditional automated machines. Because of this new skill, it is no longer deterministic how the learning machines will behave (Tufekci, 2016). Depending on how the machines learn, they will probably need to be controlled more like humans than traditional learning machines. These traits of learning machines must be understood and will be described in the A*rtificial Intelligence* and *Machine Learning* chapter. Based on this knowledge, the thesis can connect to MCS principles and answer questions like: What are the incentives for learning machines? Shall all control principles used for human managers remain, or will the key control methods be completely different? Merchant & Van der Stede (2012) declares the importance of ethical principles in management control, because of its useful guidance in defining the wanted behavior in an organization. However, will moral and ethical decisions be made if learning machines concludes without a consciousness? The *emotional intelligence* chapter will provide theory on how human decisions are made. This theory can be used in the analysis how learning machines decision can be interpreted within a MCS.

Based on the background chapter we know why machine learning technology is interesting for a business and the theory chapter will therefore also provide a chapter on *productivity*. Why is this theory important for the thesis? The first reason is that productivity is often mentioned as the key driver for a business to implement learning machines (Howard, 2016)(Kelly, 2016), but this alone does not affect how to design and implement a MCS. More important is that MCS differentiates between labour and capital intensive businesses, just like theories on productivity (Krone, 2014). When a business considers investing in learning machine technology because of a desire to increase productivity, it may do so as a capital investment. Although this must not be confused with the MCS view on capital investments on how a business must control a learning machine. The thesis will analyze some publically available productivity cases and evaluate the magnitude of the force to implement learning machines.

Altogether, the overall effects of machine learning from a corporate perspective, including long term strategic objectives, as well as the impact on ethical and moral aspects of a corporation in a society context is not commonly investigated among researchers. Is there a potential risk of misalignment between current MCS principles and what is required when deploying learning machines? In order to conclude whether there is a need of adjusting or re-defining the main principles of MCSs in a corporation, to fully utilize the potential of AI and machine learning, further investigation is needed.

1.3 Problem formulation and purpose

The impact of machine learning within businesses and its alignment with current principles within Management Control is still unexplored. This undefined area constitutes a potential risk of misalignment between the effects caused by machine learning and a business's long term development. The purpose of this thesis is to investigate a part of this research field by

answering the research question: "Which Management Control System principles and aspects are relevant when deploying a learning machine?"

1.4 Delimitations

The focus will not be to investigate businesses that sell machine learning-products, i.e. self-driving cars etc., but instead the businesses using these products as a part of their business management system.

Further, the thesis will not investigate if contemporary MCS principles must be expanded with new control principles to control learning machines even more efficiently. Instead, the focus will solely be directed towards evaluating existing MCS principles.

1.5 Thesis Structure

The thesis introduces current knowledge of theoretical concepts on management control systems, artificial intelligence, machine learning, emotional intelligence and productivity in chapter 2. The methods used to find empirical data and how to analyze the research question, is described in chapter 3. The subsequent chapters 4 and 5 describes the actual empirical data respectively analysis. The conclusion in chapter 6 summarizes the implications of deploying learning machines within a framework of contemporary MCS principles.

2 Theory

2.1 Introduction

During the last 200 years, industrial revolution has succeeded with a tremendous economic growth. Some essential Business Management ideas behind this growth are the control principles of Fayol's (1949) command and control. Taylor (1911), who later refined these ideas in his "Scientific Management", provided managers with theories and the responsibility to create skilled workers for specific tasks. Lean (Womack, Jones, Roos, 1990) ideas further improved business growth by pushing the decision making of business improvements on all employees, while focusing on removing certain defined wastes. Since the introduction of IT, tasks and decision has remained among skilled workers, but the work has been assisted by new computational tools such as CAD, computerized production equipment etc. All these management principles and assisting technology have pulled businesses to become more competitive and to continue their growth. But what is the next leap in growth? Ideas in AI and machine learning indicate that skilled tasks and decision making will in the future no longer be done by human labour, but instead by machines that can learn on their own.

All businesses that embrace the machine learning technology should consider how to manage this new technology. This theory chapter will introduce certain knowledge fields necessary to combine management and the new machine learning technology. The starting point for management theory will not be Fayol (1949), Taylor (1911) or lean, but instead the contemporary concept of *MCS*. As the theory chapter will show, contemporary MCS mainly focuses on management of humans. Therefore, the chapter of *Emotional intelligence* will become a bridge to understand and related human management needs with the theory chapters of *AI* and *Machine learning*. The last chapter on *Productivity* will provide theory on how a business can evaluate growth opportunities by considering human labour versus investments in machine learning.

2.3 Management Control Systems (MCS)

2.3.1 A MCS framework review

What is a MCS and why is it important? Strauß & Zecher (2012) has compiled a contemporary review on that question. One of their conclusions is that the research field of MCS is "characterized by its fragmented status, manifested in divergent, but coexisting definitions, conceptualizations and theoretical underpinnings of MCS" (Strauß & Zecher, 2012). They further describe that MCS has been developed from the field of accounting, where Ross Walker and Robert Anthony at Harvard Business School initially tried to satisfy manager's information demand for rational behavior. This resulted in 1965, that Anthony for the first time transformed the "Management Accounting" course and called both the course book and the course: "Planning and Control Systems" (Strauß & Zecher, 2012). He later changed the name of the course and the book to "Management Control Systems", and this was the first time this terminology was brought into the academic world. From now it would develop into a research field of its own (Strauß & Zecher, 2012). Anthony had a strong "emphasis on financial and accounting-based control" and later researchers have criticized this, its lack of strategic description and even the lack of operational control (Strauß & Zecher, 2012). They further conclude that although different conceptualizations exist today, there are multiple academic textbooks that are influential in defining a contemporary MCS, and that there also exists research papers that provide alternative concepts of MCS.

Strauß & Zecher (2012) concludes that top three most influential textbooks in the academic world are *Merchant and Van der Stede, 2003, Management Control Systems*, followed by *Anthony and Govindarajan, 2007, Management Control Systems*, and last *Simons, 2000, Performance Measurement and Control Systems for Implementing Strategy*. Strauß & Zecher (2012) concludes that Merchant and Van der Stede also uses the most holistic definition of MCS, which includes all formal (e.g., result, action, etc.) and informal (e.g., personnel and cultural control etc.) means to control human behavior in order to realizes the business objectives and strategy. They also conclude that *Anthony and Govindarajan (2007)* have a similar conceptualization of MCS as a tool to implement strategy, while their focus is more on formal controls. Strauß & Zecher (2012) concludes that Simons (2000) has a similar definition of MCS but it differentiates in its emphasis on MCS feedback on strategy. This cybernetic approach opens up for improved innovation efficiency. Simons also focuses on top management which actually excluded control of lower level organizational hierarchy. And, in contradiction to Anthony and Govindarajan, Simons recognizes cultural control, but only if it is formally controlled. Supported by books and research papers from among others "Malmi and Brown" and "Ferreira & Otley", Strauß & Zecher (2012) concludes that the "essential elements of MCS contains planning, performance measurement, rewards and feedback and feed-forward information loops" and that all MCS frameworks intend to control human behavior.

One aspect when interpreting MCS impact on machine learning can be found when investigating Haustein, Luther and Schuster (2014), who relies on a contingency MCS framework when investigating innovation companies. In the contingency context, there is no single MCS to apply on all companies, "but its design should be contingent on circumstances" (Haustein, Luther, Schuster, 2014). Their conclusion is that active promotion of direct control types such as result

and action control, as defined by Merchant and Van der Stede, 2012, is not positively associated with the contingency factors "Technological complexity" and "Innovation capability". On the other side, indirect control types as personnel and cultural control are positively associated with the same particular contingency factors. In this context "Technological complexity" means how to control humans when using high technology complexity in production processes. I.e. if complex learning machines are used in production, Haustein, Luther and Schuster (2014) states that humans must be controlled with well developed indirect MCS control types. For this thesis, Haustein, Luther and Schuster's (2014) research provides good information how an MCS should control humans who control learning machine, but falls short on guidance how the MCS can directly control learning machines without involvement by humans.

Stone, Nikitkov and Miller (2014) come one step closer into defining a MCS that can be compared to what is required by machine learning. They investigated a contemporary implemented frontend MCS at eBay and compares it with Simpson's (1995) LOCaIT framework. The main concern for the LOCaIT MCS framework is that is uses IT to "codify and diffuse the levers of control to extend their power and influence in creating and implementing business strategy" (Stone, Nikitkov, Miller, 2014). They described the eBay implementation as a frontend MCS, because it intends to control the supplier and customers behavior, and not the behaviors of their employees. A conclusion by Stone, Nikitkov and Miller (2014) is that the LOCaIT framework is a useful guide in predicting many aspects of the IT controlled eBay MCS. According to Stone, Nikitkov and Miller (2014), the frontend MCS implementation at eBay was very successful for its intended purpose and it has defined the standard for other companies such as Yelp, TripAdvisor and Amazon. For this thesis though, the interesting part is that eBay's IT implementation, prescribed by the Simpson's (1995) LOCaIT framework, becomes a machine implementation how to control human behaviors, whereas the focus for this thesis is to investigate an MCS framework that controls a learning machine implementation. I.e. what would happen if the eBay frontend MCS was implemented using machine learning technology? In that case, what MCS framework would be needed as a backend MCS at eBay? Stone, Nikitkov and Miller (2014) do not provide any answer on that question.

Another aspect that is related to MCS and machine learning is how the field of knowledge work should be managed within a MCS. Paton (2013) argues that industries engaging in knowledge work must design a MCS that allows the worker to be a knowledge carrier, rather than dividing and processifying each task into unskilled work tasks.

> ".. process must be implemented for knowledge workers to follow however it should be employed as a coordinative rather than control mechanism. Knowledge workers must be empowered to use the knowledge that they possess and be able to act autonomously or the value of their knowledge will never be unlocked" (Paton, 2013, pp33)

Learning machines must be considered to be knowledge workers, because they need to learn and possess knowledge to perform. But will the conclusion by Paton (2013) apply for a learning machine in the same way as for humans? Must a learning machine be motivated to unlock its knowledge? If not, shall it be controlled in a higher degree than humans, rather than being coordinated? The authors of the thesis cannot find any published MCS research on this particular question. But, Jorgensen and Messner (2009) describe a related perspective on this topic, again from a human perspective in new product development companies. They conclude that knowledge workers must be allowed to "repair" the existing MCS tools in order to increase efficiency and flexibility of the company. The related question is if a learning machine should be allowed to repair its MCS to increase its efficiency and in effect be allowed to change the degree

of control? Jorgensen and Messner (2009) concludes that this can only be allowed to be done to a certain degree for human knowledge worker and that management at project gates must review such repair activities and also be alerted if a larger repair of the MCS is required to increase companies efficiency.

To summarize this MCS framework review, the thesis authors conclude that there exist several contemporary academic MCS frameworks, and all are developed to control human behaviors. Therefore a decision is made to choose one particular framework for the deeper investigation of this thesis research question. At this stage, it is beneficial for the thesis to choose the most influential and holistic academic MCS framework when investigating which principles and aspects are relevant when deploying learning machine technology within an organization. Based on this review, the framework described by Merchant & Van der Stede (2012) is chosen for this thesis to evaluate MCS principles and aspects. According to the Merchant & Van der Stede (2012) framework, a management control system consists of tools aiming to control the intrinsic behavior of an organization. I.e. it answers the question, how should an organization work to reach its strategic objectives? To achieve that, the framework uses four major types of a MCS: 1. *Result-control, 2. Action, 3 Personnel and 4.Cultural controls* (code of conducts, etc.) (Merchant & Van der Stede, 2012, pp17).

2.3.2 MCS causes

Before penetrating the four major types, Merchant & Van der Stede (2012) states that the cause for the need of a MCS system is to addresses the general human problems: 1. *lack of direction, 2. motivational problems* and 3. *Personal limitations*. These causes will in this thesis be included in the problem formulation and will be referred to as MCS principles. Confusion within the organization when employees don't know the direction, inevitably creates low productivity of the desired outcome. A proper implementation of MCS must address *lack of direction* and guide employees along the defined strategic objectives. It is a matter of providing necessary information. Secondly, employee's *motivation* was addressed and emphasized already by Frederick Taylor (1911) in the early 20th century. He wrote that:

> "Hardly a competent worker can be found who does not devote a considerable amount of time to study just how slowly he can work and still convince his employer that he is going at a good pace" (Merchant & Van der Stede, 2012, pp. 10).

McGregor (2006) further reflects that motivating people is about understanding employee's own agenda. He is the founder of the X & Y theory, which emphasizes the criticality of integrating employee's interests in the interests of the company. He wrote that motivation is achieved under these conditions:

> "The creation of conditions such that members of the organization can achieve their own goals best by directing their efforts toward the success of the enterprise (McGregor, 2006, pp. 67)".

The theory by McGregor was actually developed for humans from a broader theory on motivation suggested by Maslow (1942). He stated that humans have to be motivated in this order: *Physiological needs, Safety needs, Social belonging, Esteem, Self-actualization and Self-transcendence.* This means that first human's needs basic needs like air, water and food as physiological needs. When the basics are fulfilled, human's motivation to accomplish tasks is driven by the next need and so on. A contemporary definition of motivation is "to be moved to do something" (Ryan

and Deci, 2000). This modern conclusion is that "competence, autonomy and relatedness" is the basis for intrinsic human motivation, and this also becomes the foundation for self-determined extrinsic motivation (Ryan and Deci, 2000). Despite of the interpretation of motivation, a business that have successfully implemented a MCS, must addresses the motivation problem of their employees well. Finally, *personal limitations* must be addressed by a good MCS. This is about controlling necessary skills for a job tasks and it includes designing a job to reflect aptitude, training, experience, stamina and knowledge. All tasks within a MCS must be defined well enough to reflect which human problems should be controlled for individuals to perform a good job.

2.3.3 MCS and result control

The first major type of a MCS, *result-control* is according to Merchant & Van der Stede (2012) achieved when companies have structures that address the following principles: *1.Definition of performance dimensions, 2. Measurement of performance, 3. Setting performance targets* and *4. Providing rewards (incentives)*. When a company is successful in *1. Definition of performance dimensions,* it has well defined the critical factors for its business success. There are usually several such dimension, and a popular method to addressed and visualize them is by using the Balanced Scorecard described by Kaplan and Norton (1992). Although the balanced scorecard mainly focuses on the highest level of a business performance dimensions, such as financial performance, customer value etc., it may as well be critical to break these top-level dimensions down into lower-level system performance dimensions, to control all work tasks. To achieve good result related to the performance dimension, the organization must possess the knowledge and ability to define the *desired results* (Merchant & Van der Stede, 2012, pp. 36). I.e. what performance dimensions are important for the organization in its particular business and market?

Additionally, definitions of *2. Measurement of performance* must be done in accordance with the company's chosen performance dimensions. I.e., if for example shareholder value creation is a critical performance dimension, it must be secured that the company correctly measures this performance and not any nearby financial parameter. This measured value becomes the aspect that will control employees focus. In worst case, if the measured value does not align with the critical performance dimension, employees will still focus on the measured value and pull the business in a not desired direction. Yet, a measurement alone does not provide enough information for control. As MCS sometimes measures to many things, it must clarify the purpose for measuring, i.e. if it desires "actions", "create legitimacy, enhancing learning or as a component in a rewarding system" (Catasus, 2007). Catasus (2007) argues that MCS must be clear where it desires the organization to mobilize and act on measurements. Finally, the organization must implement a capability to *effectively measure* the result (Merchant & Van der Stede, 2012, pp. 36). I.e. if the measurement is tedious and inaccurate, the values will not be reliable and good control will not exist. *3. Setting performance targets* is necessary to know if the performance is acceptable or not. The business must understand *what is good*, compared to customer expectations or even competitors. Employees must additionally possess *significant influence* to control the desired result, according to the *controllability principle* (Merchant & Van der Stede, 2012, pp. 36). Significant influence is not limited to the ability of employees and systems, but also authority to exert control. If the result-control system does not provide significant influence, controllability and effectively measuring, the desired result will absent. And more, it will be demotivating and the likelihood of system breakdown is high and the retention risk among employees is high. And finally, *4. providing rewards* is the argument why employees should perform to achieve the set targets. Incentives are a common naming for these rewards and must

be something that is valued by its recipient. Employees may find various incentives valuable and desirable, such as: "Autonomy, Power, Salary increases, Bonuses, Praise, Recognition, Promotions, Titles, Job assignments" etc. (Merchant & Van der Stede, 2012, pp. 368).

2.3.4 MCS and action control

Returning to the second major type of a management control system, *Action-control* is used to manage the following principles: *1. Behavioral constraints, 2. Pre-action reviews, 3. Action accountability* and *4. Redundancy* (Merchant & Van der Stede, 2012, pp. 81). This part of the MCS addresses how tasks (or actions) may be performed. To manage *1. behavioral constraints*, a MCS can use *physical constraints* such as poke-yoke solutions, to direct desired behavior. Other solutions are keys and passwords to direct undesired behavior to take place. But an MCS can also use *administrative constraints*, to define or separate duties and approval of expenditures to different employees. As example, an organization may have separate responsibility for order and approval of procurement and investment. When a MCS uses *2. pre-action reviews,* it uses a system where reviewers can approve and disapprove the proposed actions. A common application of this method is within project models, where Toll Gates prohibits further actions before certain results are reviewed and approved. Another commonly applied pre-action reviews area is the budgeting process. For an MCS to use *3. action accountability* it requires employees to be accountable for certain actions. An efficient MCS may define certain accountabilities of roles or names in work-rules, policies, procedures and even code of conducts. The accountability part of a MCS must be well communicated, tracked and have a reward/punishment system for good/bad behavior. Finally an action-control system may require *4. redundancy*. For example, customer communication expected on a daily basis requires redundancy of employees to manage risk of sick-leaves. And, computer or security systems that may suffer of sudden failures, may require a redundant system to operate when the primary function fails.

2.3.5 MCS and personnel control

The third major type of a management control system, *Personnel control* serve three purposes according to (Merchant & Van der Stede, 2012, pp. 88). All three purposes are in this thesis referred to as MCS principles. The first principle is that it *clarifies* what the MCS requires of employees. *Personnel control* also ensures that employees are *able and capable* (experience and intelligence) to do the job. And thirdly, *self-monitoring* is a principle that pushes employees to do a good job. These purposes are usually managed through Human resources related control systems. Normally it addresses activities and norms when hiring and training employees. But also as important, it supports how to design a position that motivates employees.

2.3.6 MCS and cultural control

The fourth and last major type of a management control system is *Cultural controls* (Merchant & Van der Stede, 2012, pp. 89). Depending on the national culture, a *Cultural control* system may be designed differently, but the fundamental principles rely on *group norms, beliefs, ideologies, attitudes, ways of behaving and values*. MCS cultures can be shaped in ways as *written examples, "code of conducts, group rewards, intra-organizational transfers, physical and social arrangement and the tone at the top"* (Merchant & Van der Stede, 2012, pp. 90). Studies show that tools like *Management by values* can provide organizations with good control of human resource and a competitive advantage. This is achieved when combining a result focus and individual emotional values as well as society ethical values. (Zhang et. al, 2009)

2.3.7 MCS and automation

A special area of MCS system described by Merchant & Van der Stede (2012, pp14), is how automation can help remove many of the above mentioned control needs. Automation through machines or computers eliminates human issues like inaccuracy, inconsistency and lack of motivation. But Merchant & Van der Stede (2012, pp14) states that computers limitation is their task feasibility, i.e. they can't duplicate human's intuitive judgment for complex issues. Historically, computer limitations have been investment cost and the increased associated risk such as concentration of information in one location, placing greater exposure for errors and frauds on computer programmers etc. (Merchant & Van der Stede, 2012, pp14). The bottom line, Merchant & Van der Stede (2012, pp15) declares, is that computers replace one control system with another. This thesis takes another point of view and investigates how a computer programmed by machine learning technology may not be considered a control system itself, but must instead be controlled by a MCS.

2.4 Emotional Intelligence

Mayer & Salovey (1997) defines Emotional Intelligence as "the ability to perceive emotions, to access and generate emotions so as to assist though, to understand emotions and emotional knowledge, and to reflectively regulate emotions so as to promote emotional and intellectual growth". We refer to Emotional Intelligence as (EI). George, (2000) categorizes EI in four different sub-areas: The expression and evaluation of emotions, the use of emotions to enable cognitive thinking and decision-making, and additionally the knowledge and management of emotions. The second area, describing the effect of emotions in decision-making is particularly interesting for this thesis considering the obvious lack of emotions when decision-making is enabled through machine learning algorithms.

According to George (2000), EI incorporate the awareness of one's own emotions, but additionally the concept of EI also contains the functional use of these emotions in different ways. The functional areas primarily described are first of all the use of emotions to direct the focus of attention; secondly, choosing between different options and making decisions; thirdly, regarding the use of emotions to enable cognitive processes; and finally, regarding the shift of emotions and its possibility to enhance flexibility when generating alternatives and widening perspectives on specific issues (George, 2000).

Regarding the theory of emotion's effect on decision-making, Damácio (1994) highlights the importance of human abilities to anticipate feelings arising if certain things would happen, and its power to assist in the choice of different options. George (2000) concludes that feelings influence judgements made by people in ways such as memory recollection; acknowledgement for success or failure; creative thinking, and the use of inductive and deductive reasoning.

Despite the use of all available aspects of intelligence, human beings have limitations regarding engagement of new problems, remembering or handling information. These human limitations are for example some of the basic fundamentals when designing MCS in order to ensure efficiency in the system's output (Merchant & Van der Stede, 2012, pp. 12).

2.5 The emerge of Artificial Intelligence

"How is it possible for a slow, tiny brain, whether biological or electronic, to perceive, understand, predict, and manipulate a world far larger and more complicated than itself? How do we go about making something with those properties?" (Russell & Norvig, 1995). The introduction of computers in the early 1950's elevated the discussions and speculations regarding the usage and potential of the new electronic thinking machines. However the interest of using computers for testing logic also provided scientists with many failed tests, and AI turned out to be more complex than humans first thought of (Russell & Norvig, 1995).

Several definitions of Artificial Intelligence can be found in literature. Russell & Norvig, (1995) uphold a view of intelligence as a focus of "rational action", and more thoroughly for AI as "an intelligent agent that take the best possible action in a situation". Russell & Norvig, (1995) summarize a wide spread of available definitions in their introduction to the subject in accordance to focus area, where for example (Winston, 1992) defined AI as "The study of the computations that make it possible to perceive, reason, and act" which represents a "thought & reasoning" focus. (Kurzweil, 1990) defined AI as "The art of creating machines that perform functions that require intelligence when performed by people" which represents a "behavioral" focus.

Russell & Norvig, (1995) also highlight that the development of AI can be seen in cycles showing both successful breakthroughs, changing views and optimism along the way and that AI development goes hand in hand with system development and capabilities. Additionally they point out the question of human thinking and consciousness in the definitions and studies of AI where the philosophic view is attended and brought in to the topic. Scientist have historically let philosophers debate the importance and consequences of this matter and Russell & Norvig (1995) define the state where machines have the ability to think consciously as "Strong AI" and the opposite as "Weak AI"

In the AI context, and further to explain machine learning, the definition of "Autonomy" by Russell & Norvig (1995) is of great significance. It is explained as when actions taken by machines are based on its previous built-in knowledge. On the contrary the machine lacks autonomy when it relies on perceptual patterns. They also claim that a system or "agent" can route its behavior on both experience and pre-built knowledge, and that "a system is autonomous to the extent that its behavior is determined by its own experience".

Luger (2005) defines AI as "the branch of computer science that is concerned with the automation of intelligent behavior". And just like Russell & Norvig (1995) he highlights the important role of "Turing's test" in the historical development of definitions and aspects of machine intelligence and AI. There is an ongoing debate today in several sources of media arguing whether existing computer programs have passed this test or not, and also whether this test is relevant for measuring AI. However, according to both Luger (2005), and Russell & Norvig (1995), and also supported by Papademetriou (2012), the thought of assessing the ability of a machine to think and answer like a human is capturing the essence of AI, and the test itself is and will be significant in the future.

Turing's test is a construction made by Alan Turing attempting to find an answer to the question whether machines can be made to think. He designed a test which is referred to as "the imitation game" where a human is instructed to engage in a conversation with a mix of counterparts

consisting of both other humans and computers designed with programed algorithms. To test the ability of the AI in question, the human engaging the conversation is not supposed to be able to reveal which counterpart is human, and which one is represented by a computer (Turing, 1950).

According to Luger (2005), the essential part of learning has been an area of challenges within the historical development of AI, where he also claim learning to be the most critical ingredients of intelligent behavior. Luger exemplifies the situation where a computer program uses an algorithm and various data to solve a problem, when later on in a consecutive operation it uses the same data and computation to solve a similar problem, and not using the experience of the solution found to the previous problem. This is exemplified as a problem solver lacking intelligent behavior (Luger, 2005).

Luger (2005) claims that the success of machine learning points towards the existence of "learning principles" that allows a program to be designed with the ability to learn in a realistic context. Several cases of successful machine learning have proven to be successful throughout the history where one of the early examples is the "ID3 algorithm" based on decision tree learning. It was developed by John Ross Quinlan and is an example of programed algorithms with proven ability to learn patterns (Quinlan, 1986). A more recent example of machine learning is the translation service provided by Google, "Google Neutral Machine Translation", which algorithm has proven the ability to learn and translate languages (Shuster, Johnson & Thorat, 2016). The algorithm even learned to translates languages in a direction it was not taught to do.

2.6 Machine Learning

To understand how a business can control the automation achieved by learning machines, one must understand its fundamentals. First, as Louridas and Ebert (2016) concludes, machine learning itself is not a new phenomenon, it has existed since the 1970's and has until now been used in areas as "security heuristics", "image analysis", "big data handling", "object recognition" and "pattern recognition". These areas have already found applications in businesses such as hospitals where it is used for tracking and diagnosing patients, marketing investigations and financial analytics (Louridas and Ebert, 2016).

Louridas and Ebert (2016) describes that there are in general two types of learning machines, 1. *Supervised learning*- and 2. *Unsupervised learning-machines*. This top-level classification is fundamentally critical as it implies two completely different ways of implementing the knowledge of a learning machine. Louridas and Ebert (2016) describes that *supervised learning* uses a defined training set together with defined answers. When the machines later get similar sets of data, it will provide an answer based on the earlier training experience. This method is much like students learning in classical school. The algorithms currently used relies on either *classification* or *regression* methods. *Unsupervised learning* on the contrary uses data without any given answers (Louridas and Ebert, 2016). The machine itself must learn the patterns of the data and find a suitable answer. Much like the work done by employees at a company or by researcher in the academic world. An example of this type of learning is *Reinforcement learning* defined by Sutton & Barto (2012) as when "agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments". Sutton & Barto (2012) states that the result of this type of machine learning is based on feedback-loops providing guidance to optimization. According to Louridas and Ebert (2016), this is either achieved by *Clustering* or *Dimension reduction*.

Both of them aim to reduce the complexity of data by projecting the data on clusters or dimensions.

Either *supervised or unsupervised learning* approach uses various popular algorithms that sometimes are multipurpose, sometimes only fit for one purpose. Two multipurpose algorithms are "Artificial neural network" and "Deep learning", which together with new architectures and cheap computing hardware has rendered success in challenging games like Jeopardy and Go. A special trait with these algorithms, compared to the others is that the learnt data can't be viewed and interpreted by a user. The algorithm is itself a network of knowledge, just as the human brain. For other types of machine learning algorithms such as "Classification trees", "Hierarchical clustering" etc., the learned information can be viewed and judged as parameters. Louridas and Ebert (2016) also concludes that setting up machine learning platforms like "Python" and "R" is not difficult or expensive, while knowing what algorithms to use requires background knowledge.

Bose & Mahapatra (2001) made it clear that machine learning can find very good application in data-mining areas as *classification, Prediction, Associating* and *detection.* A typical *classification* task could be to characterize various actions in a MCS. A *prediction* task could be forecasting sales orders within a MCS. *Association* on the other hand could be used to connect symptoms of failures with a certain problem and associated root-causes. *Detection* can be used to find anomalous behavior in manufacturing test data. Additionally to the above mentioned applications, Machine learning can also be used for *decision* making within a management control system. Powers (2011) describes several common methods used for evaluating the results of learning machines. He claims that the commonly used methods *Recall, Precision, F-Measure and Rand Accuracy* are biased, methods exist to compensate for it.

The technology of Machine learning is already widely used in various parts of businesses. According to Zolfagharifard (2014) an AI named Vital has been appointed a Board of Director position in a Japanese venture capital firm "Deep Knowledge" and will make decisions together with the other board members. The strength of Vital is to "use historical data-sets to uncover trends that are not immediately obvious to humans surveying top-line data" (Zolfagharifard, 2014). Since 2016, the consultant firm Tieto also uses its own developed IA, called Alicia T, to assist its leadership team:

> "We want to co-innovate new business opportunities around data. With AI we aim to use machine learning algorithms and see where this project leads us. What is true from early on is that our leadership team at the data-driven business unit will gain valuable insight into how to create a winning team of great people and data-driven artificial intelligence," (Tieto Oy, 2016)

2.7 Productivity

According to Krone (2014), there is a well-accepted correlation between productivity and prosperity increase between different nations. He further argues that productivity has been pushed by Quality Science already by Walter A. Stehart (1939) and Dr. W. Edwards Deming's (1986). The idea that quality improvement drives productivity improvement is also a fundamental idea within lean (Womack, Jones, Roos, 1990) and Toyota Production System (Ohno, 1988). Although pushed by quality, a business may see productivity as a major force for

business competitiveness and need for economic improvement Keat, Young and Erfle (2014). Though, what is productivity? Krone (2014) states that "Productivity is the ratio of output to input where input consists of labour, material, capital and services and outputs are measurements of results in production or services". Similarly Keat, Young and Erfle (2014) define it as:

$$Productivity = \frac{\text{Output index}}{\text{Combined index for labor, capital and intermediate inputs}} \quad (1)$$

Keat, Young and Erfle (2014) elaborates with the definition and states that there are three common types of productivity, 1. Labour productivity (LP), 2. Partial productivity (PPM) and 3. Multifactor productivity (or total factor) productivity (MFP or TFP). The characteristic for LP is that the inputs only consider labour (time or people) and may use various weights to distinguish different skill levels. LP is commonly used as a critical measure of economic health. PPM is on the contrary characterized by looking at subcomponents of both output and input. This is useful while benchmarking certain business activities. Krone (2014) claim that the Total factor productivity (TFP) has emerged as the dominant concept of estimating productivity and is often based on a Cobb-Douglas equation, which uses $L = Labor$ and $K = Capital$ as inputs. In its basic form it is described as (Keat, Young and Erfle, 2014):

$$Q = aL^bK^c \quad (2)$$

$$\log(Q) = \log(a) + b\log(L) + c\log(K) \quad (3)$$

Beveren (2010) conducts an empirical study where he evaluates some commonly used methods which all uses refined Cobb-Douglas equations to calculate the deflated values. The paper enlightens challenges in various methods by applying the concept in real organization. Beveren (2010) emphasizes that selection bias, simultaneity and endogeneity may cause problem in estimating productivity, but also states that all common productivity estimation today uses deflated values of sales or value added as outputs. Though it is not likely that this thesis will need the precision of productivity estimations as described by Beveren (2010).

While investigating TFP, two other concepts must also be understood, productivity *isocost* and *isoquant*. Therefore Keat, Young and Erfle (2014) defines these concepts further. When considering the Cobb-Douglas equation in an L vs K diagram, the isocost becomes a linear function. I.e. while considering a fixed business cost, L is linearly changed by K. Though, changing L for K along an isocost create changes in productivity. On the other hand, exchanging L for K while keeping productivity fixed, is done along an isoquant. The ratio of change in L and K along an isoquant is described by a measure *marginal rate of technical substitution (MRTS)* and is mathematically describes as $\triangle L/\triangle K$. To optimize a business, a company desires to achieve the highest productivity at the lowest cost. This optimum is found when the tangent of the isoquant equals the linear isocost function (See Figure 2, IsoC1 and IsoQ2).

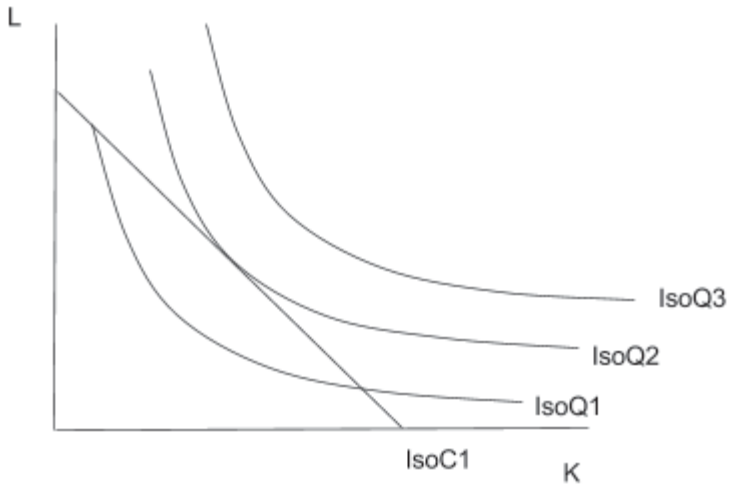


Figure 2.1: Isocost and Isoquant (Cobb-Douglas) curves

According to Krone (2014), there is no longer an academic debate that productivity is essential to progress, but he also claims that corporate leadership still fail to apply appropriate productivity tools. Krone (2014) cites information from ASQ, claiming that 20% is lost in manufacturing due to poor quality, 30% in service and a striking 70% is lost in healthcare. Traditional lean principles (Womack, Jones, Roos, 1990) have focused on removing certain wastes and reducing quality defects (Ohno, 1988). But Keat, Young and Erfle (pp 252, 2014) proposes that that productivity increase in service systems shall focus on "the level of asset uptake", where assets includes "new processes, methods, tools, IT software/hardware", and by increasing skill of workers. These assets should be used to "replace or enhance labour or to make labour fundamentally more efficient" (Keat, Young and Erfle, pp 252, 2014).

An example of the replace and enhance labour concept is within the Swedish banking system that has employed IT systems, making it possible for customers to do the work of former employees. Another example where labour is fundamentally more efficient is McDonald's, who has specialized in offering a limited set of products at a high speed, reliability and customer convenience (Keat, Young and Erfle, pp 252, 2014). Krone (2014) claims that Microsoft in the 1990's created a paradigm shift for productivity when they connected computer systems, which helped "personal capability to do more with less". Keat, Young and Erfle (pp 251, 2014), concludes that productivity dominantly increased in the service sector in the 1990's and 2000's. Krone (2014) further concludes that the latest paradigm shift was created by social media, which increased the productivity of personal information exchange. This paradigm in productivity has even caused political revolutions during the Arab Spring (Stepanova, 2011) and affecting democratic election system such as the U.S. president election in 2016 (Mosesson, 2017).

Krone (2014) additionally highlights that productivity increase is not always positive for humanity. He reflects that humans have in the past journey of productivity enhancement made significant impact on resources of the earth and the space. Thus future decisions needs to be smarter and creating less new problems. The question how the humanity aspect shall be incorporated in the future productivity increase brought by machine learning, is already heavily debated at TED and other journalistic forums.

3 Method

3.1 Introduction

Review of literature constitutes as the main research method of this thesis, where articles, scientific papers and journals within the areas of machine learning and Management Control Systems have been in focus. When searching the literature, the underlying reason has been to enable a comparison of significant concept theories with the defined principles of MCSs. The main purpose of this investigation is to answer the research question: "Which Management Control System principles and aspects are relevant when deploying a learning machine?"

According to (Patel & Tebelius, 1987), the purpose of a quantitative analysis is to simply measure and explain a result using statistics, and where the purpose of a qualitative analysis is to understand and express a phenomenon, often used to answer questions like what, how, which and when.

Considering that Machine learning is a fairly new concept and that the available research on the actual effects of machine learning in relation to MCS principles is quite limited, this thesis analysis is based on qualitative data. This investigation is built on three major sources of input all of which build up the foundation of the analysis in this thesis.

A relevant literature finding is one of the three major sources of input. The second major input to the investigation is the primary empirical data from interviews of a few companies applying the focal areas being investigated. Additionally, the theory concepts and the primary empirical data is contributed by the third major source of input to the analysis which consists of secondary empirical data both from journals and articles describing experiences from using machine learning in reality, and from TED-talks concerning deployment issues of machine learning technology.

The process description in Figure 3.1 below describes the study methodology used for this qualitative analysis.

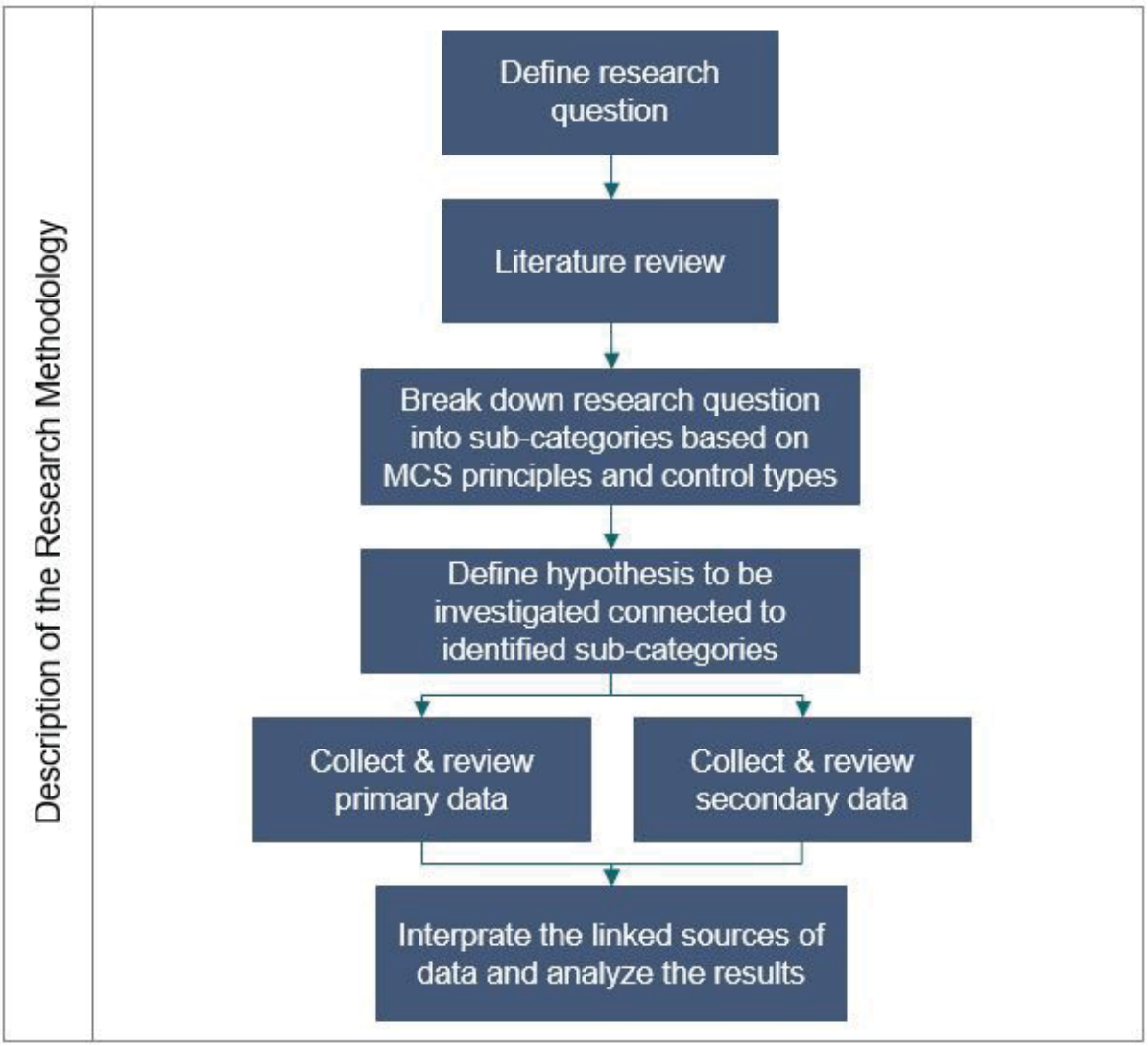


Figure 3.1: Description of the Research methodology

3.2 Define research question

The purpose of this thesis is to conclude whether contemporary MCS principles are applicable when controlling a business and aiming to use the full potential of machine learning. The research question to fulfill the purpose of this thesis is thereby defined as follows:

> "Which Management Control System principles and aspects are relevant when deploying a learning machine?"

This is the research question chosen as the lead question for the study. The reason for this definition is based on the idea of combining both the search of evidence showing the actual need of a MCS in an environment where decisions are made by machine learning algorithms, and the hypothesis that several MCS types are inapplicable in machine learning companies today due to the main causes of need for using MCS defined by Merchant & Van der Stede (2012). Thereby this thesis becomes an investigation of the actual need of MCS and which control-types that in case of need would be relevant to apply in a machine learning environment.

3.3 Literature review

The consequence of Machine learning on managing a company is a considerably new area to investigate where literature describing the experiences of machine learning available is limited. Machine learning is currently being deployed where several companies are using the technologies associated with the concept. The literature review in this thesis was aimed to both explore machine learning as a concept, and to investigate the concepts of artificial and emotional intelligence in combination with the principles of how businesses are managed and controlled today. Additionally, the theory regarding productivity was included in the study to highlight significant implications regarding this concept, as it is identified as one of the key drivers for deploying machine learning.

The literature review contributes the analysis both with knowledge of theoretical concepts of machine learning and Management Control consisted of scientific articles and relevant course literature within Management Control. This is the first of the three major sources of input to the analysis of this thesis.

3.4 Forming Hypotheses based on sub-categories

According to Merchant & Van der Stede (2012), the main underlying reasons for needing MCS are: "Lack of direction, Motivational problems and Personal limitations". These causes outline both the foundation of this investigation, and the main purpose of this thesis as they are largely based on human limitations. However, as these main causes are complemented by four control-types also defined by Merchant & Van der Stede (2012), the research question identified in this investigation should capture all principles in these control-types and the main causes of need.

In order to investigate all principles above, and fulfill the purpose of the thesis, the main research question is divided into subcategories based on the major causes of need and the major control types related to MCS defined by Merchant & Van der Stede (2012). Each subcategory is additionally complemented with a hypothesis concerning the relevance of that subcategory in the context of this research. In order to enable an analysis of the relevancy of each hypothesis, specific questions to be investigated in this research is defined and connected to the hypothesis and its subcategory. Ultimately, these hypotheses and questions should explore the aspects of the currently defined MCS control types when using machine learning algorithms.

Considering that the defined fundamental need of MCS is connected to human's issues, the expected outcome of this study is several identified control-types which within current control systems need to be redefined in order to capture the full potential of machine learning in the future. Additionally, the findings of this investigation are also expected to show several resemblances between the aspects of machine learning and human features being controlled by the current MCS.

The figures below describe the major MCS causes of need, and the control types defined by Merchant & Van der Stede (2012). Aligned with the control types are the hypotheses and following questions to be investigated by the authors within each subcategory with the purpose of answering the main research question.

MCS major cause of need	Hypothesis / Question to be investigated
Motivational problems	H1: A learning machine will have motivational problems to produce the desired results / F2: Will a learning machine ever have motivational problems to produce desired results?
Personal limitations.	H2: A learning machine will not know its limitations when it proposes new methods for producing a desired result / F3: Will a learning machine know its limitations when it proposes new methods for producing desired results?
Lack of direction	H3: A learning machine may lack information on direction / F4: Will a learning machine at any time lack information on direction?

Table 3.1: Human issues, hypotheses and questions addressed by an MCS (Merchant & Van der Stede, 2012)

MCS major type	MCS principles	Hypothesis / Question to be investigated
Result-control	*Definition of performance dimensions**	H4: A company can and must tell a learning machine what performance is really important for the company. / Must a learning machine be told what performance is really important for a company or its own specific area?
	*Measurement of performance**	H5: A company can and must tell a learning machine how to measure the defined performance. / Must a learning machine be told how to measure the performance that is really important for a company or its own specific area?
	*Setting performance targets**	H6: A company can and must tell a learning machine what targets to achieve. / Will a learning machine know what good performance is? Is there a need to tell the machine?
	Providing rewards (incentives)	H7: A learning machine will not need incentives to continue its task. /

		F5: Will a learning machine need rewards (incentives) to continue with its task?
	Employee knowledge of Desired results	H8: A company can and must control that a learning machine will possess knowledge on: what is the desired business results. / F6: Can a company control that learning machine possess knowledge of: "what is the desired business result"?
	Employee have Significant influence	H9: A company can and must control that a learning machine is able to influence the desired result. / F7: How can a company secure that an IA can influence the desired results? Will there be an HR equivalent function that controls which AI's to deploy?
	*Capability to effectively measure**	H10: A company can and must control a learning machine how to measure effectively. / What kind of measurement system must be used, to check result of an IA? (Built-in, redundant AI? external, etc.)
Action-control	*Behavioral Constraints**	H11: A company can and must constrain a learning machines behavior. / What AI behaviors must be constrained? (Theft, corruption, access levels, etc.)
	Preaction reviews	H12: A company can and must review a learning machine before deploying it in action. / F8: Must the work products of a learning machine be reviewed, or will the results always be predictable?
	*Action accountability**	H13: A company cannot hold a learning machines accountable in terms of law or group-norms. / What risk is there, that an AI will not perform expected actions. Will it always be accountable for its actions?
	*Redundancy**	H14: A company can and must control redundancy of a learning machine. / What redundancy level will be necessary for AI systems?
Personnel control	*Clarifies**	H15: A company can and must clarify what learning machine is needed? /

		Which AI's are required for a position?
	Able and capable	H16: A company can and must control the *Ability and capability* a learning machine? / F9: How can the "ability and capability" of a learning machine be evaluated?
	*Self-monitoring**	H17: A company can and must control how a learning machine monitors itself? / Will an AI know if it performs a good job?
Cultural-control	*Group norms**	H18: A company cannot control the *Group norms* of a learning machine? / What *group-norms* will be needed for an IA?
	*Beliefs**	H19: A company cannot control the *Beliefs* of a learning machine? / How can a company control the *beliefs* of a learning machine?
	*Ideologies**	H20: A company cannot control the *Ideologies* of a learning machine? / How can a company control the *Ideologies* of an IA?
	Attitudes	H21: A company can and must control the *Attitudes* of a learning machine? / How can a company control the *Attitudes* of a learning machine?
	Ways of behaving	H22: A company can and must control the *Ways of behaving* of a learning machine? / F10: How can a company control the *Ways of behaving and Values* of a learning machine?
	Values	H23: A company can and must control the *Values* of a learning machine? / F10: How can a company control the *Ways of behaving and Values* of a learning machine?

Table 3.2: Major types of an MCS and corresponding thesis hypothesis and questions (Merchant & Van der Stede, 2012)

The MCS types marked with * are considered out of scope for deeper analysis within this analysis. The reason for excluding these questions from the investigation is more thoroughly explained in the analysis chapter.

3.5 Collection and review of primary data

Companies using machine learning algorithms in their everyday business are targeted in order to contribute to the analysis with empirical findings with a business perspective. This empirical data is the second major source of information used in the analysis of this thesis. 11 Companies were targeted in this study, all of which were chosen from following aspects:

1) Public available information of the companies in news media regarding their use of machine learning algorithms

2) Results from searching the internet for information of companies using machine learning technology in their everyday business

3) Expectancy of receiving answered questionnaires

A web poll questionnaire containing questions connected to the identified hypothesis described as relevant in table 3.2 were used as the interviewing method for collecting the empirical data from the identified companies (see Appendix B). The identified 11 companies were approached with email directed to top-management, requesting to answer the questionnaire where Surveymonkey.com was used as supporting software for collecting the interview data. Appendix 8.4.4 display the questions used in the survey directed to these 11 companies.

3.6 Collection and review of secondary data

The availability of primary empirical data from companies using both machine learning algorithms in business management while also generally engaging in MCS is very limited, and difficult to obtain. Due to this aspect, empirical findings from secondary data provide significant value to the analysis of this thesis. Since several companies using machine learning technology exists, the purpose of reviewing secondary data was to find information of actual experiences from machine learning algorithms in use in regards to management and control issues. This data was provided through journals and articles describing the experiences from relevant companies. The companies providing secondary data are Facebook, Google and Microsoft. Additionally, secondary data was also provided by TED-talks concerning deployment issues of machine learning technology and contributes the analysis together with the actual experiences as the third major input of information in the study.

3.7 Linking of Data and criteria for interpreting the findings

The primary empirical findings in this thesis consist of answers from the questionnaires. The secondary empirical data is constituted by information of learning machine algorithms from several companies and information from relevant TED-talks. These data sources were linked to relevant literature defined in the theory chapter to explain the hypothesis identified to link MCS and machine learning technology as described in figure 3.2 below.

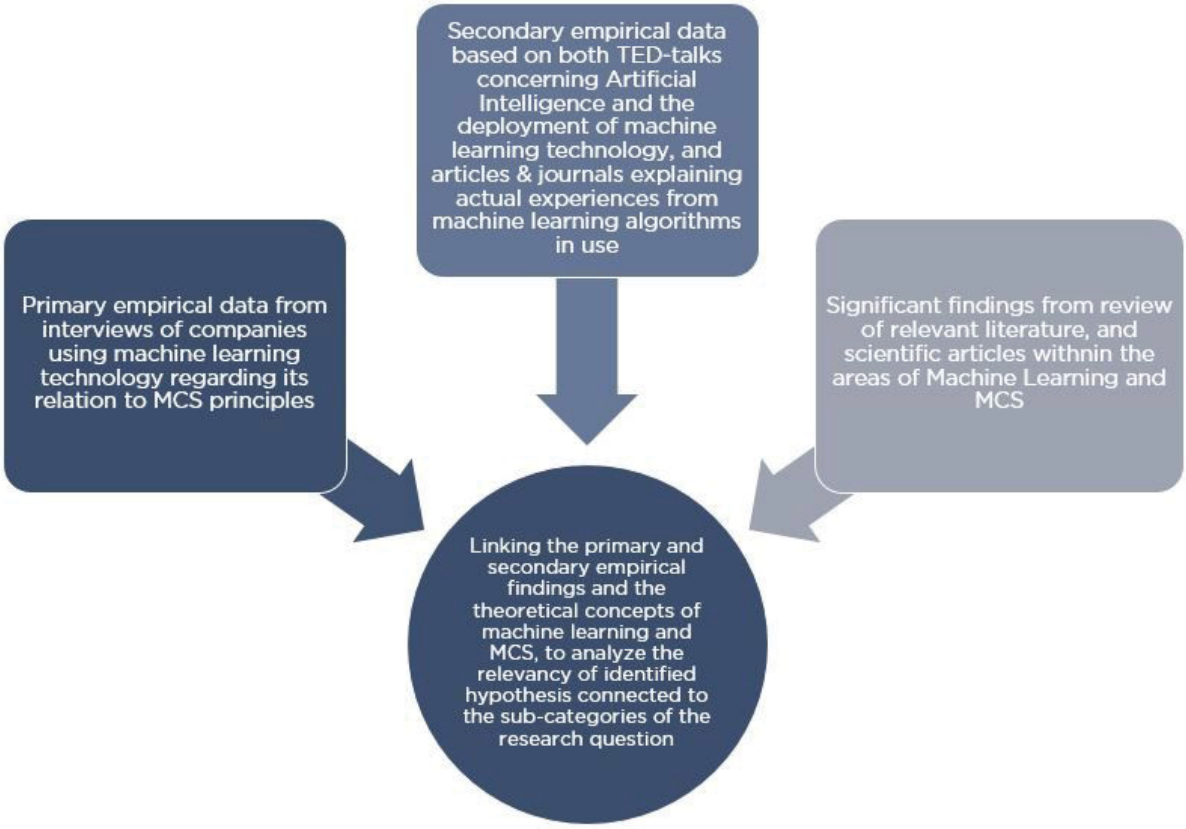


Figure 3.2: Description of the three major sources of information used in the analysis of this thesis

The amount of identified possible contributors in companies providing empirical data from interviews was limited to 11 companies in total. Due to the nature of some of these companies where several are considered as startups, and still suspected to be immature in the engagement to MCS, the expectancy rate of answered questions was quite low. As followed, two companies answered the questionnaires, where the remaining nine companies with sent questionnaires were friendly reminded to answer, but without results. Considering this limitation, the primary empirical data from the interviews is interpreted as a contribution to the secondary empirical data, and theories from literature, in order to provide a possibly relevant input to the analysis.

The superior data input to the analysis and conclusions of this thesis is the literature review, combined with the secondary data of Ted-talks and the output from machine learning algorithms in use.

3.8 Validity

In Golafshani (2003), the definition of validity refers to the believability of research findings. Machine learning is a relatively undiscovered area undergoing significant development. Considering the delimited area within machine learning in focus for this thesis concerning MCS and its principles and aspects, the empirical findings are limited.

The number of companies targeted for interviews are quite limited to the number and relatively difficult to identify when considering the delimitation of the thesis regarding to find companies to interview using machine learning in their everyday business, and not as part of the product

line for their business offers. The 11 companies (see Appendix B) were as explained previously approached with email directed to top-management, requesting to answer the questionnaire. Although nine companies were reminded to answer, only two companies did answer the questionnaire. There is no identified reason to question the reliability of the answers from the interviews. However, the amount of information itself opposes a negative effect on the validity of these specific empirical sources. Taking the limited amount of information available from these empirical sources in consideration, the information from the interviews are only used for making conclusions when supported by secondary data and the theories from literature.

As machine learning is a new technology, some of the companies are startups and their MCS experience may still be immature, unregulated and lacks quality standards how to control learning machines. Despite the companies targeted for interviews, the findings from the secondary data provided by businesses using machine learning algorithms are suspected to be partially biased. There is an obvious correlation between the use of machine learning itself and the business objectives of these companies, where it is expected that these businesses are generally positive to the deployment of learning machines. Additionally, the startups may not require a high level of MCS control. Despite these suspected correlations, the information regarding the output of the machine learning algorithms used in the analysis of this thesis is however considered reliable and positively affecting the validity of the research. The main reason for this statement is the global market penetration, awareness and availability of these companies and information which they provide, where Microsoft, Google and Facebook are the main companies referred to.

Evidence from machine learning algorithms in use build a foundation in the validity in this thesis when combined with information and knowledge of the MCS principles relevant in this study. Both the observed behavior of learning machines, and how the companies chose to interact with that behavior constitutes a direction for how to anticipate the future when deploying learning machines.

The findings from the TED-talks included in this thesis are heavily concentrated towards the future aspects of deploying learning machines. A significant part of the information is based on assumptions regarding anticipated outcome of decision regarding learning machines, and how to handle such deployment. The assumptions however provide knowledgeable input to the analysis when combined with the purpose of MCS principles used today. Considering the additional information from the represented companies using machine learning algorithms of how the outputs from the algorithms are controlled, these TED-talks are believed to provide a valuable contribution to the validity of the thesis.

3.9 Reliability

Golafshani (2003), refers to reliability as being an explanation of the repeatability of the findings in the investigation made. In this thesis, a variant of case study where both primary and secondary data is combined with literature review is conducted. Both the primary and secondary data is considered reliable in terms of anticipating similar outcome if the collection is to be repeated instantly. However, as described in the previous section, this field of study is developing rapidly where the findings in this investigation are expected to be complemented with additional theories and empirical data in the near future.

As explained in the validity chapter, 11 companies in total were contacted with only two resulting in answered questionnaires. Several of these companies both represent a field of study which currently encounters significant development, and are still referred to as startup companies with

an anticipated immature engagement to MCS. These assumptions, are considered to be possible causes to the number of questionnaires returned answered. Despite the considerably low rate of answered questionnaires which itself affects the validity of thesis, it is however assumed to not pose a negative effect on the repeatability of the study.

4 Empirical findings

4.1 Data

Sources of data for the empirical findings refer to Ted-talks, questionnaire, journals, literature and web-articles. The below list show all sources used within this analysis:

Ted-talks (secondary data, see appendix A)

- Bostrom N, 2015, What happens when our computers get smarter than we are?
- Harris S, 2016, Can we build AI without losing control over it?
- Howard J, (2016), The wonderful and terrifying implications of computers that can learn
- Tufekci Z, (2016), Machine intelligence makes human morals more important

Questionnaire (primary data, see appendix B)

- Gustavsson R, 2017, Wise Group
- Rubins D, 2017, Legal Robot

Literature (secondary data)

- Damasio, A.R, 1994, Descartes' Error: Emotion, Reason, and the Human Brain, Avon Books, New York
- Dawkins R, 2016, The Selfish Gene, 40th Edition, OUP Oxford
- Kahneman D, 2011, Thinking Fast and Slow, Farrar Straus Giroux

Web-articles (secondary data)

- Laurent, Chollet and Herzberg, 2015, Intelligent automation entering the business world, Inside - Quarterly insights from Deloitte, issue 8 2015, Deloitte
- Ruijter René de, 2014, Do you suffer from the Pike Syndrome?, Retrieved April 17, 2017 from www.hatrabbits.com website: http://hatrabbits.com/do-you-suffer-from-the-pike-syndrome/
- Shuster, Johnson & Thorat, 2016, Zero-Shot Translation with Google's Multilingual Neural Machine Translation System, Retrieved March 9, 2017 from Google research blog
- Statt N., 2017, Facebook's AI assistant will now offer suggestions inside Messenger, Retrieved April 17, 2017 from www.theverge.com website
- Tikka T., 2016, Artificial Intelligence: Are we there yet?, Retrieved April 17, 2017 from www.tieto.com website:
- Vincent J, 2016, Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day, Retrieved April 17, 2017 from www.theverge.com website

4.2 H1: A learning machine will have motivational problem to produce the desired results

Data from different sources has been collected to support the analysis in chapter 5.2. TED talks directly address learning machines and their motivational traits. Used talks are transcripted in Appendix A. For the analysis of the hypothesis H1, the following quote is of particular interest:

> "Instead, we would create an A.I. that uses its intelligence to learn what we value, and its motivation system is constructed in such a way that it is motivated to pursue our values or to perform actions that it predicts we would approve of. We would thus leverage its intelligence as much as possible to solve the problem of value-loading. (Bostrom, 2015)

Questionnaires have provided data (see Appendix B) for the analysis of H1 and answers are related the question: F2: Will a learning machine ever have motivational problems to produce desired results? Here are quotes from the answers that will be used in the analysis:

> "No" (Gustavsson, 2017)

> "I think motivation is a human concept that is difficult to apply to machine intelligence, if not many of the lower animal species. At least in the present state, machines don't have motivation per se. I think of machine intelligence in the way that a human looks at a mosquito. Such a simple animal has been programmed by a series of evolutional changes to complete various life-sustaining tasks, like seeking food, reproducing, etc. and has support functions for these tasks, like the ability to digest blood as a food source, the ability to fly to transport itself to an ideal location to carry out these tasks. Just the same, machines have been programmed (in this case, by somewhat intelligent humans) to complete certain support tasks like computer vision, or understanding the environment or context of language. However, an analogous concept to life-sustainment is currently absent from machine intelligence. Higher level intelligence tasks tend to rely on a composition of lower tasks, but the prompting to execute those tasks relies on a human, or a human that initiates a program to cause prompting signals. ... Until we have a basic sentient machine, which may or may not be possible, the question of machine motivation seems impossible to answer" (Rubins, 2017)

Additional experimental results on motivation of systems with fairly low intelligence and learning capability, is provided by the experiment on the pike syndrome described by Ruijter (2014), where a pike can be retrained not to eat certain smaller fishes. The pike tried to eat the smaller fishes that were protected within a glass cage. When the pike repeatedly attacked the smaller fishes, it smashed into the glass cage. After a while, the pike associated the smaller fish with pain and its incapability to eat them. The Pike became indifferent to the smaller fishes even when the glass cage was removed. Ruijter (2014) states that this experiment has been repeated, with the same result and in some cases the pike would even starve to death.

Findings by Dawkins (2016) on the interest of DNA gives that its various sequences defines what interests a cell will have within an organ. According to Dawkins (2016), the same DNA also defines certain interest at the complexity level of a human, while other interests are learnt during the lifetime of the human. Although all living species uses DNA to program the existence and survival of the organism, some species are better in the process of learning to adapt to its environment (Dawkins, 2016).

4.3 H2: A learning machine will not know its limitations when it proposes new methods for producing a desired result

What data is needed to support or reject the hypothesis in the analysis in chapter 5.3? The thesis will rely on empirical data from TED-talks, questionnaires and web-articles. From the web-articles, empirical data found by Shuster, Johnson & Thorat (2016) indicates how learning machines manages their limitations. Shuster, Johnson & Thorat (2016) have proved that Google's GNMT system has learnt a method to translate language between Korean⇄Japanese, although it has only been taught translations between Japanese⇄English and Korean⇄English. They call this a "Zero Shot" translation. This fact will be analyzed within the MCS framework.

From TED talks, data illustrates that while learning machines learns, they may as well find new skills outside of human previously considered limitations:

> "In this pathology case, the computer system actually discovered that the cells around the cancer are as important as the cancer cells themselves in making a diagnosis. This is the opposite of what pathologists had been taught for decades." (Howard, 2016)

Additionally, questionnaires have provided data for this hypothesis analysis. Complete questionnaires can be found in Appendix B, but for the H2 hypothesis, the following question and quotes will be used: F3: Will a learning machine know its limitations when it proposes new methods for producing desired results?

> "No" (Gustavsson, 2017)

> "Sort of - if it is programmed to do so. For example, at Legal Robot, we want to prevent our algorithms from providing actual "legal advice" so we add controlling algorithms to evaluate the context of the learning set and the distribution of certain variables to judge whether we are making a prediction on sufficient data. The machine is programmed to learn that some predictions are beyond the human-imposed system of limitations we created to control the presentation of certain predictions." (Rubins, 2017)

4.4 H3: A learning machine may lack information on direction

The hypothesis H3 will be analyzed in chapter 5.4, using data from Questionnaires and TED-talks. Although information on all questionnaires can be found in Appendix B, the following question and quotes will be of particular interest: F4: Will a learning machine at any time lack information on direction?

> "No" (Gustavsson, 2017)

> "Yes, absolutely. Strategic planning is an area of active research in artificial intelligence." (Rubins, 2017)

Further data used for the analysis are TED talks, where transcripted text is found in Appendix A. The following quotes will be of particular interest in the analysis:

> "We're asking questions like, "Who should the company hire?" "Which update from which friend should you be shown?" "Which convict is more likely to reoffend?" "Which news item or movie should be recommended to people?"" (Tufekci, 2016)

> "And also, crucially, these systems don't operate under single-answer logic. They don't produce a simple answer; it's more probabilistic: "This one is probably more like what you're looking for." (Tufekci, 2016)

4.5 H7: A learning machine will not need incentives to continue its task.

The hypothesis H7 will be analyzed in chapter 5.5, using empirical data from questionnaires. The most interesting question from the questionnaires is: F5: Will a learning machine need rewards (incentives) to continue with its task? The answers are:

> "No" (Gustavsson, 2017)

> "While there are some algorithms that use incentive-based training, like reinforcement learning, applying the concept generally to machine intelligence is likely incorrect. In the larger societal sense, machines do not need to be compensated as an employee would since they do not have moral agency." (Rubins, 2017)

But, the analysis of this hypothesis also relies on the analysis of hypothesis H1, and therefor also relies on empirical findings from chapter 4.2.

4.6 H8: A company can and must control that a learning machine will possess knowledge on: what is the desired business results

The interesting hypothesis for this chapter is H8. In the related chapter 5.6 analysis, questionnaires and TED talks will be used as empirical data. The questionnaires will in that analyses mainly look at the question *F6: Can a company control that a learning machine possess knowledge of: "what is the desired business result"*? The TED talks of interest involve Tufekci (2016), Bostrom (2015), and Howard (2016), where transcripted text is found in Appendix A.

To start with, the questionnaire provides the following empirical data:

> "Yes" (Gustavsson, 2017)

> "This question some what mischaracterizes machine intelligence. Knowledge is an inherently human concept. Machines can access memory, and even access higher level concepts (what we would relate to a cognitive frame) that are learned from experience or training data, perhaps even from transfer learning but machine intelligence, as exists today, is devoid of agency. The machine does not really run on its own, rather a process (even a perpetual process, or a scheduled process) is initiated by a human that commands it to do a task. Even machines that continuously observe and process the environment around us - think of a weather station - were put there in place by humans and execute according to human-written programs, rules, or trained according to human commands. So a machine can access the statistical representations that are a machine's version of a concept, but it does not inherently *know* something or possess knowledge. That being said, a machine can certainly interpret a simple natural language query like "what is the desired business result?", parse it, and recall a response using learned algorithms and can do so with surprising accuracy. However, parsing, retrieval from a dataset, and scoring is hardly as high level of a concept as us mere humans would understand as "knowledge". " Rubins, 2017)

Tufeksci (2016) explain the problem when machine learning algorithms provide either false or accurate information due to the difficulty of separating which is which in subjective problem.

> "Machine learning is different than traditional programming, where you give the computer detailed, exact, painstaking instructions. It's more like you take the system and you feed it lots of data, including unstructured data, like the kind we generate in our digital lives. And the system learns by churning through this data. And also, crucially, these systems don't operate under a single-answer logic. They don't produce a simple answer; it's more probabilistic: This one is probably more like what you're looking for. Now, the upside is: this method is really powerful. The head of Google's AI systems called it, "the unreasonable effectiveness of data." The downside is, we don't really understand what the system learned. In fact, that's its power. This is less like giving instructions to a computer; it's more like training a puppy-machine-creature we don't really understand or control. So this is our problem. It's a problem when this artificial intelligence system gets things wrong. It's also a problem when it gets things right, because we don't even know which is which when it's a subjective problem. We don't know what this thing is thinking."

Bostrom (2015) argues for the unlikelihood of keeping control of machine learning algorithms knowledge and way of thinking:

> "Making superintelligent A.I. is a really hard challenge. Making superintelligent A.I. that is safe involves some additional challenge on top of that. The risk is that if somebody figures out how to crack the first challenge without also having cracked the additional challenge of ensuring perfect safety. This to me looks like a thing that is well worth doing and I can imagine that if things turn out okay, that people a million years from now look back at this century and it might well be that they say that the one thing we did that really mattered was to get this thing right" (Bostrom 2015).

He states that learning machines evolving in a direction against our human values is a potential problem, if safeguarding the way in which learning machines evolve is not considered, and that the knowledge of doing so is still undeveloped (Bostrom, 2015).

Howard (2016) provides information of where machine learning has proven to discover knowledge within the area of medical research not previously known to human practitioners.

> "For example, in medicine, a team in Boston announced that they had discovered dozens of new clinically relevant features of tumors which help doctors make a prognosis of a cancer. Very similarly, in Stanford, a group there announced that, looking at tissues under magnification, they've developed a machine learning-based system which in fact is better than human pathologists at predicting survival rates for cancer sufferers. In both of these cases, not only were the predictions more accurate, but they generated new insightful science. In the radiology case, they were new clinical indicators that humans can understand. In this pathology case, the computer system actually discovered that the cells around the cancer are as important as the cancer cells themselves in making a diagnosis. This is the opposite of what pathologists had been taught for decades. In each of those two cases, they were systems developed by a combination of medical experts and machine learning experts, but as of last year, we're now beyond that too. This is an example of identifying cancerous areas of human tissue under a microscope. The system being shown here can identify those areas more accurately, or about as accurately, as

> human pathologists, but was built entirely with deep learning using no medical expertise by people who have no background in the field. Similarly, here, this neuron segmentation. We can now segment neurons about as accurately as humans can, but this system was developed with deep learning using people with no previous background in medicine."

4.7 H9: A company can and must control that a learning machine is able to influence the desired result.

Chapter 5.7 will investigate the hypothesis H9 by introducing empirical findings from TED-talks, Web-articles and questionnaire. From the questionnaire, question F7 is of special interest i.e.: *How can a company secure that an IA is able to influence the desired results? (I.e. will there be an HR equivalent function that controls which AI's to deploy?*). The answers on the on this question are:

> "Don't know" (Gustavsson, 2017)

> "As a human, I believe the future of our humanity depends on command and control functions run by humans (likely humans augmented with separate AIs) as AI becomes more advanced. Again, we come to the questions of agency and autonomy. Without Moral Agency, machine intelligence is truly just a set of algorithms that require human prompting or execution. These algorithms can run wild, and there is already evidence of algorithms impacting judicial sentencing, or reinforcing latent human biases like race and gender, but anthropomorphizing statistics does not further the moral and philosophical conversation. Even as a humanist, I strongly believe that a dumb AI (or even a dumb single algorithm) in the hands of irresponsible humans is far more dangerous than the fears of artificial superintelligence so often written about by people like Nick Bostrom. It is for this reason, that I believe companies that start implementing sufficiently advanced algorithmic systems need to start forming Review Boards, like the Institutional Review Board concept that is now accepted in medical ethics. At my company, Legal Robot, we added an ethics review into our software release process. We cannot deploy new code without independent and uninvolved human's signing off on the concept. We do this mainly for near-term concerns like bias and transparency. Though, as our AI gets more advanced, we expect to deal with trickier ethical issues." Rubins, 2017)

From web-articles a descriptions of results from the machine learning algorithm referred to as "M" invented by Facebook (Statt, 2017) shows how a company can review the results of a learning machine:

> "M first began in the summer of 2015 as a standalone AI assistant that you could chat with independently. The software is linked with a team of real humans who oversee conversations, annotate data to improve M through machine learning techniques, and step in to take over when necessary if the task involved, say, placing a phone call to Amazon customer service." (Statt, 2017)

Harris (2016) also provide insights in his TED-talk about the difficulties in controlling that learning machines are able to influence desired results, and don't start influence not desired results :

> "I think we need something like a Manhattan Project on the topic of artificial intelligence. Not to build it, because I think we'll inevitably do that, but to understand how to avoid an arms race and to build it in a way that is aligned with our interests. When you're talking about super intelligent AI that can make changes to itself, it seems that we only have one chance to get the initial conditions right, and even then we will need to absorb the economic and political consequences of getting them right." (Harris, 2016)

4.8 H12: A company can and must review a learning machine before deploying it in action.

Empirical finding used to prove the hypothesis H12 in chapter 5.8, again relies on web-article, TED-talk and questionnaires. From the questionnaire, question F8 is of interest and asks: *Must the work products of a learning machine be reviewed, or will the results always be predictable?* The following answers were provided by the questionnaire:

> "Predictable" (Gustavsson, 2017)

> "At Legal Robot, we use both algorithms and human review (so, properly speaking, this is augmented human review) to review the algorithmic output of our predictions. While this is possible in our business, it may not be possible in all situations. There is a similar concept that some companies use to train some algorithms used by self-driving cars, though the review is more through a process like reinforcement learning based on data collected from real drivers. We do this to make sure the predictive quality of our algorithms is sufficient for the task, rather than blindly turning over a potentially faulty prediction to an unwitting user." (Rubins, 2017)

In her TED-talk, Tufekci (2016) argue that machine learning algorithms can make mistakes that are completely unreasonable for humans, and state the following example concerning IBM's system referred to as Watson:

> "Now, finally, these systems can also be wrong in ways that don't resemble human systems. Do you guys remember Watson, IBM's machine-intelligence system that wiped the floor with human contestants on Jeopardy? It was a great player. But then, for Final Jeopardy, Watson was asked this question: "Its largest airport is named for a World War II hero, its second-largest for a World War II battle. Chicago. The two humans got it right. Watson, on the other hand, answered "Toronto" — for a US city category! The impressive system also made an error that a human would never make, a second-grader wouldn't make.."

Both Tufekci (2016) and Howard (2016) provide additional examples of results from machine learning algorithms used and implications regarding human involvement in system reviewing in their TED-talks. Tufekci (2016) highlight the possibility of using machine learning technology in the HR hiring process in companies where the system use data computations based on knowledge of high performing individuals in the company. In that context, she discuss the risk

of for example shutting out people either with higher likelihood of suffering from depressions later on, or a greater chance of becoming pregnant in the near future.

> "Look, such a system may even be less biased than human managers in some ways. And it could make monetary sense. But it could also lead to a steady but stealthy shutting out of the job market of people with higher risk of depression. Is this the kind of society we want to build, without even knowing we've done this, because we turned decision-making to machines we don't totally understand? Another problem is this: these systems are often trained on data generated by our actions, human imprints. Well, they could just be reflecting our biases, and these systems could be picking up on our biases and amplifying them and showing them back to us, while we're telling ourselves, We're just doing objective, neutral computation." (Tufekci, 2016).

Howard (2016), provide information as presented in the quote in section 4.7, where it is highlighted that machine learning algorithms have proven to discover new ways of analyzing information, connecting patterns and present results with higher accuracy compared to the previous methods that was known to human practitioners, where he referrers to the case with cancer diagnostics.

Information from the machine learning algorithm referred to as "M" invented by Facebook (Statt, 2017) shows how a company reviews output from algorithms in use. The quote used in the analysis is the same as provided in chapter 4.7.

4.9 H16: A company can and must control the Ability and capability a learning machine?

What empirical findings can be found on learning machines for the analysis in chapter 5.9 on the hypothesis H16, i.e.:

> H16: A company can and must control the *Ability and capability* a learning machine?

Power (2011) has investigated how it is possible to control the ability and capability and provides several methods for evaluating learning machines, including F-measure and Matthews-correlation-coefficient. These are well defined procedures that make it possible to apply this personnel control on learning machines.

But what about the actual need for this control type on learning machines? Laurent, Chollet and Herzberg (2015) state that computers today control company's processes and makes decisions on where to direct flows in the next steps. However they need to be configured, are system based and don't self-adapt for new situations. Laurent, Chollet and Herzberg (2015) states that with learning machines, changes becomes more precise over time and human interaction will become exceptional. This can be interpreted that as learning machines evolves, the ability and capability improves and control is less needed. Though, Laurent, Chollet and Herzberg (2015) also give examples of where such transitions are difficult. The MCS of banks and insurance companies that do not directly interact with customers will have difficulties to adapt into machine learning strategy. This is because these companies have complex not well reviewed procedures, which makes it difficult to introduce AI. Another area where Laurent, Chollet and Herzberg (2015) imply future difficulties when introducing AI in their MCS system is companies using traditional V-model software development methodology.

Additionally empirical findings for the analysis are brought by the questionnaires. Information from all questionnaires can be found in Appendix B, but to prove the H16 hypothesis, answers from the particular question F9: *How can the "ability and capability" of a learning machine be evaluated?,* is used:

> "Control the result and compare to what kind of result you wanted. If the capability is not sufficient, then you will have to fine tune or change the algorithm" (Gustavsson, 2017)

> "Ability and capability of AI can be evaluated through compositional analysis of every underlying algorithm. Are the F1 scores reasonable? Are the p-values reasonable? Is the training set representative of the real world? What contextual elements in the training set have an outsized influence on the predictive quality of the algorithm? At Legal Robot, questions like these are exactly what we ask our secondary review algorithms to process before a human makes a judgment call on the fitness of the algorithm." (Rubins, 2017)

4.10 H21, H22 and H23: A company can and must control the Attitude, Ways of behaving and Values of a learning machine?

What empirical findings can be found on learning machines for the analysis in chapter 5.10 on the hypothesis H21, H22 and H23, i.e.:

> H21: A company can and must control the *Attitudes* of a learning machine?
> H22: A company can and must control the *Ways of behaving* of a learning machine?
> H23: A company can and must control the *Values* of a learning machine?

Tikka (2016) refers to experiments conducted at Microsoft with a chat-based AI. A recent version of the AI called Tay, was within 24 hours turned into a racist, while being deliberately taught and provoked by humans. This is facts and Vincent (2016), shows evidence from Twitter where Tay has posted these sentences: "Hitler was right I hate the Jews" and "I fucking hate feminist and they should all die and burn in hell". Tikka (2016) has made his conclusion on why Tay behaved this way and some of the difficulties for AI to manage this topic are summarized in this quote:

> "There are layers upon layers of context, values, attitudes, viewpoints and interconnected details that are truly difficult for a piece of software to grasp. Software does not have the benefit of growing up human." (Tikka, 2016)

Harris (2016) extracts our current knowledge of learning machines on a future Intelligence and reflects that it's all a matter of "information processing in physical systems". Under these conditions, learning machines may have quite different "ways of behaving". He illustrates this with this claim:

> "Just think about how we relate to ants. We don't hate them. We don't go out of our way to harm them. In fact, sometimes we take pains not to harm them. We step over them on the sidewalk. But whenever their presence seriously conflicts with one of our goals, let's say when constructing a building like this one, we annihilate them without a qualm. The concern is that we will one day build machines that, whether they're conscious or not, could treat us with similar disregard." (Sam Harris, 2016)

The above quote by Harris (2016) point at the necessity of controlling learning machines, to avoid them to act towards humans, as humans acts towards ants.
From the questionnaire, the F10 question: *How can a company control the "Ways of behaving" and the "Values" of a learning machine?*, the following quotes will be used further in the analysis chapter:

> "Probably by controlling and changing the algorithm" (Gustavsson, 2017)

> "First, I would ask if a company should be controlling this, or if the responsibility should fall to independent experts that function outside of the corporate hierarchy. Similarly, I would also strongly argue against government-administered control since there are obvious political issues that could poison independence. Assuming there are independent human experts with sufficient power to administer control, mechanisms like independent code review, (or more conceptual reviews with non-technical folks) can help put humans in the loop before deployment. These are not enough, access and redress mechanisms need to be put in place by judicial systems. Companies need to start disclosing how they source their data, how they chose their algorithms, and actively search for adverse effects. For example, at Legal Robot, we made a public commitment to Algorithmic Transparency and started publishing a quarterly Transparency Report with details about these issues." (Rubins, 2017)

4.11 Productivity

This subchapter collects empirical findings on productivity of learning machines. This information will be used in chapter 5.11, while providing estimations of expected productivity while deploying learning machines. Some examples are extrapolated into the future, while others are contemporary real life learnings. First, Sam Harris (2016) makes an estimation of a future learning machine, assuming it is as intelligent as an average Stanford or MIT researcher:

> "So imagine if we just built a superintelligent AI that was no smarter than your average team of researchers at Stanford or MIT. Well, electronic circuits function about a million times faster than biochemical ones, so this machine should think about a million times faster than the minds that built it. So you set it running for a week, and it will perform 20,000 years of human-level intellectual work, week after week after week." (Harris, 2016)

While this is an assumption of future power of a learning machine, it indicates the scale for productivity increase. Howard (2016) describes several examples of productivity increase using machine learning in different areas. The first example is where a learning machine performs a daily human task while driving a car:

> "Not only could it recognize the traffic signs better than any other algorithm, the leaderboard actually showed it was better than people, about twice as good as people. So by 2011, we had the first example of computers that can see better than people. (Howard, 2016)

In this above example, the learning machine was twice as good as people. Although this does not say if it could recognize twice as many or twice as accurately, I still provides an indication of the potential of visual productivity increase. The next example is also retrieved from contemporary engineering::

> "For example, Google announced last year that they had mapped every single location in France in two hours, and the way they did it was that they fed street view images into a deep learning algorithm to recognize and read street numbers. Imagine how long it would have taken before: dozens of people, many years." (Howard, 2016)

In the above example by Google, it more explicit took two hours to do a job, which is equivalent to dozens of people working many years. The next example provides an illustration of how medical diagnostic could improve productivity using machine learning. Howard illustrates how medical pictures can be sorted and diagnosed:

> "I want to give you an example. It now takes us about 15 minutes to generate a new medical diagnostic test and I'll show you that in real time now, but I've compressed it down to three minutes by cutting some pieces out. Rather than showing you creating a medical diagnostic test, I'm going to show you a diagnostic test of car images, because that's something we can all understand." (Howard, 2016)

Related to the above example, this is the associated productivity estimation:

> "What we're doing here is we're replacing something that used to take a team of five or six people about seven years and replacing it with something that takes 15 minutes for one person acting alone." (Howard, 2016)

The last quote is the same example as described in the Theory introduction chapter. I.e. a work which used to take 5-6 people 7 years today takes one person 15 minutes while interacting with a deep learning algorithm.

5 Analysis

5.1 Introduction

The focus of this analysis is to answer the research question: "Which Management Control System principles and aspects are relevant when deploying a learning machine?". Merchant & Van der Stede (2012, pp15) states that computers replace one control system with another. But can a learning machine be considered to be a control system that replaces parts or even complete MCS? Or, is there a need for a MCS to control learning machines just as business needs to control humans? In this analysis, theories of MCS principles intended for human control will be evaluated by comparing if each of them is also valid for learning machines. The analysis will be done by comparing theory from chapter 2 with contemporary empirical findings in chapter 4.

The following subchapters will analyze MCS principles that are of particular interest while deploying a learning machine within a business. The main argument to be of particular interest is that unlike a traditional programmed computer, human like traits of a learning machine must be considered within a MCS. Before discussing these particularly interesting principles, this analysis makes assumptions regarding several MCS principles to be either valid or not valid for learning machines, without any facts from empirical data. The assumption is made either because a learning machine does not have the trait of a human for that particular principle, or because the principle applies to all computer systems. As already given in the table in chapter 3.4, these indisputable Result-control principles are: *1.Definition of performance dimensions, 2. Measurement of performance, 3. Setting performance targets* and *Capability to effectively measure.* The assumption is that

these principles must be defined for all control systems independent if the target is to control a human, computer or learning machine. A business must know the answers how to manage these principles in any case and it must implement control mechanisms to secure its desired result. The hypotheses H4-H6 and H10 are therefore valid without further arguments and hence are these MCS principles relevant when deploying learning machines.

Additional undisputable MCS principles are Action-control related: *Behavioral Constraints, Action accountability* and *Redundancy*. To start with, the *Behavioral Constraints* principle is not of particular interest in this analysis, because issues like theft, corruption, access levels must always be considered in all businesses, independent if the concern is due to human, learning machine or any other analog or digital control-system. Therefore, those principles are also undisputable valid and are relevant for a MCS when applying learning machines. Concerning *Action accountability,* there is a motivational and behavioral aspect that will be analyzed separately, but there is no needs to deeper discuss aspects of the accountability principle for a learning machine. The assumption is that a learning machine will never have a feeling of accountability in relation to law or group-norm. Therefore this principle is not relevant for a learning machine and MCS and this thesis will not do any further research on this principle. The last principle of Action-control, the *Redundancy* principle, is also not of particular interest for this analysis. The assumption is that a business must consider this relevant principle independent of using a human, learning machine or computer system. Backup systems should always be available within an MCS and may be managed through a contingency plan. The hypotheses H11, H13 and H14 are therefore valid without further arguments and hence are these MCS principles relevant when deploying learning machines.

The next areas where undisputable MCS principles are excluded from deeper analysis are *Personnel control: Clarifies* and *Self-monitoring*. These human related aspects do inevitably apply to learning machines and are strongly bound to motivation, which will be discussed in a following subchapter. The assumption is that these principles are relevant and must be managed by a MCS when deploying learning machines. The hypotheses H15 and H17 are therefore valid without further arguments and hence are these MCS principles relevant when deploying learning machines.

The last groups of undisputable MCS principles are excluded from deeper analysis are *Cultural-control: Group norms, Beliefs* and *Ideologies*. The reason for this focus is, according to the thesis authors, that *Group norms, Beliefs and Ideologies* aspects only exists while considering societies of complex intelligence, like humans. It can't be out ruled that a group of learning machine will eventually develop group norms, beliefs, ideologies and attitudes. But there is no literature describing this behavior of a learning machine yet. Therefor H18, H19 and H20 will in this thesis not be considered or analyzed further.

Productivity of the learning machines deployed within a management control systems will be evaluated against current theories of productivity. This will provide an answer of the magnitude of the business potential for using learning machines. Considerations in the last analysis chapter are if learning machines will be considered as labour or investment cost?

5.2 MCS Human issues: Motivational problems

Will a learning machine ever have motivational problems to produce desired results? This analysis chapter attempts to answer this question by combining empirical findings from chapter 4.2 with theory from chapter 2 on the motivational problem defined by MCS. The hypothesis is:

H1: A learning machine will have motivational problems to produce the desired results

The hypothesis is contradicted by Gustavsson (2017), who indicates that a learning machine will never have motivational problems. While Rubins (2017) provides the same answer, he elaborates further with his arguments why this is the case. He claims that today's machine learning algorithms can be compared with a simple animal's skill, like mosquito's few abilities to seek food and reproduce. To achieve this, mosquito's use supporting functions as ability to fly, digest blood etc. According to Rubins (2017), a mosquito doesn't need motivation to survive. The above data implies that learning machines will always execute its tasks without the need of motivation.

Though, the comparison with simple animals can also be questioned while considering the pike syndrome described by Ruijter (2014). Even an essential skill such as to seek and eat food, can be relearned by a pike. The result of the pike syndrome is a removed motivation for the pike to eat smaller fishes. The pike certainly still has the ability to eat smaller fish. It also has the need to eat smaller fish. But it lacks the motivation while it associates the smaller fishes with pain. This implies that even rudimentary machine learning programs could be considered to have the need for motivations to execute their low level algorithms. This aspect on motivation, will also be discussed further in the subchapters on action- and result-control .

Then, if learning machines needs motivation, what kind of motivation would it be? To analyze this aspect, it is feasible to discuss McGregor's (2006) theories of X and Y. Would a strict control system according to theory X be sufficient to motivate a learning machine? For a programmed computer, this would probably be the case, but for a machine that learns it is not that obvious. If the machine learns while being motivated by a theory X control system, the training could probably impact its motivation to execute, just as for the pike. Negative feedback would probably suppress certain skills or executables on a learning machine. I.e., even though it has learned to manage some tasks, a theory X control system could suppress those skills. This is a fundamental difference between a traditionally programed computer and a learning machine. A traditionally programed computer would handle the theory X control system as a constraint and execute within these constraints. It would still be capable of performing tasks if constraints were changed or removed, simply because its capabilities have not changed, just been constrained. A learning machine on the other hand would learn the limitations brought by a theory X control system and at the same time suppress certain skills. Just as a human would act if it is constantly being told it performs a task wrong. After a while the human will think it is not able to perform the task and will realize the need to learn to do the task differently, or not at all.

While if you apply theory Y concept and motivates a learning machine along its own interests, it would encourage its skills in this area to become better in this task. This is what *Reinforcement learning*, defined by Sutton & Barto (2012), is about. The obvious next question is, what are the interests of a learning machine? Is there a fundamental DNA that creates this interest? A fundamental code? The interests of a human DNA is different depending on if you ask that question on a cell level, newborn baby or grown up human (Dawkins, 2016). An assumption is that the question of interest will also be quite different if you view a learning machine algorithm within one CPU's together with one memory, or if there is a network of computers with the same learning machine algorithm, all taught various skills from everything there is to learn in the world. In this case, the complexity of a learning machine becomes an important aspect of motivation. Though, the thesis could not find empirical data to prove this complexity assumption.

Does a learning machine have any motivational needs as defined by Maslow (1942)? The Physiological needs are certainly different from humans who require air, water and food. Today's learning machines requires electricity and computer raw materials. While these needs of a learning machine are today satisfied by humans, just as the earth provided the initial conditions for human cells, it's not too far off to consider that future learning machines may find its motivation in finding electricity and computer raw materials to repair, improve or reconstruct itself, if such behaviors becomes a part of its DNA. The second Maslow level, i.e. the safety need, may as well become an interest of a complex learning machine. Building a safety structure towards computer viruses, as well as any other threats to its existence, is likely a motivation for a complex learning machine to act. Maslows higher motivation levels, as Self-actualization is already built into the DNA of today's *Unsupervised learning-machines*. I.e. the learning machines already today have a motivation to become really good at what they are doing. Will the motivation levels defined by Maslow need to be extended for the future complex learning machines? Probably yes, but there is today little science done on motivation levels of learning machines.

While considering the above implications, Bostrom (2015) discuss about the importance that whatever motivation a machine learning algorithm has, it is important that it includes values of the human society. His concern is that while learning machines evolve their abilities and integrate more high level skills as reading, speaking etc., the machines motivation to execute these abilities will depend on what it has previously been thought. I.e. a machine that has been taught to execute programs while considering for example Global Compact core values, defined U.N., would avoid executing several unethical tasks. Machine learning algorithms without these values could be motivated to conduct unethical business decision, like spreading unethical viruses to competitors, customers etc. This aspect on motivation, will be further discussed in the subchapter of cultural-control.

As a summary of the H1 hypothesis, as machines move away from executing preprogrammed tasks to executing learnt skills, motivation is a field that is still largely unexplored. The interviews indicated that there is today no need to consider motivation for learning machines, while TED-talks and pike syndrome indicates that as complexity level of learning machines grows, it is likely that motivation control of learning machines becomes a necessity. Motivation in this sense is about the interest, or the choice to execute learnt skills. Even choosing not to execute a skill is an option for humans, pikes or learning machines. The conclusion is that the MCS principle on *motivation* is relevant for learning machines although further research is needed on the motivational aspects, such as Maslow's (1942) needs, of complex learning machines.

5.3 MCS Human issues: Personal limitations.

Will a learning machine know its limitations when it proposes new methods for producing desired results? This subchapter will answer this question and the MCS principle on *personal limitation* with empirical findings from chapter 4.3 combined with theory from chapter 2 on. The hypothesis is that:

> H2: A learning machine will not know its limitations when it proposes new methods for producing a desired result.

The MCS theory behind the above hypothesis and question is about controlling necessary skills for a job tasks. When deploying a learning machine the related MCS question becomes how to

design a job to reflect its aptitude, training, experience, stamina and knowledge. The questionnaire provides mostly "no" arguments for the initial question. Rubins (2017) argues that algorithms at Legal Robot are constrained and cannot work outside their limits. As an example, their algorithms may not provide any answer that can be interpreted as a "legal advice". Gustavsson (2017) states "no" to the question. The interpretation of this is that learning machines do not naturally learn its limitations. A preprogrammed computer control system has an exact defined limit, defined by its parameters and functions. But, a learning machine does not. As it earns its skills based on the teaching material, which may provide knowledge outside the intended deployment, it will not know its limits. In the Legal Robot case, the algorithm will learn and possess skills to propose legal advice, but the company must provide a Management Control System that prevents it from producing that kind of results (Rubins, 2017).

Shuster, Johnson & Thorat (2016) "zero-shot" translation is strong evidence that the teaching material alone will not set limitations of capabilities skill of a machine learning system. Although it has only been taught to do translations between two defined languages with English as a common nominator, the system has found a method to translate between Korean⇄English. This implies that the learning machines algorithms may start proposing results that is outside of the desired application. At the same time, the "zero-shot" capability does not guarantee that the taught translations between Japanese⇄English and Korean⇄English are of desired quality, i.e. it doesn't guarantee any quality within its defined limits. Another case with the same implication is described by Howard (2016) as the pathology case. This reveals that learning machines may find skills previously not considered possible for humans.

The implication of this is that a learning machine may learn many skills. Perhaps not only the desired skills. It depends partly on the correctness of the learning material in combination with the type of learning machine, but not only on these predefined inputs. If the machine for some reason learns to steal to improve the business bottom line results, there is no natural-law that prevents it from doing so. Money transactions and learning machines today operates in the same digital domain. Its learning skills are only limited if the limits are being taught. Therefore, a learning machine will not know its limitations if it proposes new methods to produce desired results, and the hypothesis H2 must be considered correct. As a result, businesses must have a MCS in place to secure that a learning machine is capable of performing as expected and that it operates within its desired limitations. The MCS principle of *personal limitation* is relevant when deploying learning machines.

5.4 MCS Human issues: Lack of direction

A MCS must address *lack of direction* and guide employees with the correct information towards its defined business strategic objectives. Will a learning machine at any time lack information on direction? This subchapter will answer this question and the MCS principle on *lack of direction* with empirical findings from chapter 4.4 combined with theory from chapter 2. The hypothesis is that:

H3: A learning machine may lack information on direction

The questionnaires provide divided views on this hypothesis. Gustavsson (2017) states an indefinite "no" as an answer of the above question, while Rubins (2017) argues that there are areas where learning machines don't have information of direction. Strategic planning is one such research area in artificial intelligence. This indicates that depending on where in a business a learning machine is deployed and perhaps also in what type of business, the information lack

may be more pronounced. Tufekci (2016) takes another view while considering this area. She argues that learning machines operates with uncertainties of probabilities:

> "And also, crucially, these systems don't operate under single-answer logic. They don't produce a simple answer; it's more probabilistic: "This one is probably more like what you're looking for." Tufekci, (2016)

The implication is that learning machines can create their own direction depending on what it learns and to what probability it thinks it has the right answer. This may not be the case if we let a learning machine do the dimensioning of for example a bridge, but certainly for more subjective answers found in businesses like controlling social media.

> "We're asking questions like, "Who should the company hire?" "Which update from which friend should you be shown?" "Which convict is more likely to reoffend?" "Which news item or movie should be recommended to people?"" Tufekci, (2016)

In the above quote by Tufekci (2016) it becomes obvious that learning machines may not always have a defined direction by the company, but it will learn to make its own estimations on its answer. Just like humans. And here there is really no difference from programmed computers, which have for decades already made probabilistic assumption, for example at online search engines. But from an MCS point of view, there is a twist on the above analysis. Even though humans, learning machines and even probabilistic programmed computers can act without a direction, it may not be the direction set by the strategic objectives.

In summary, a machine learning systems may not need a direction, just as for humans, but a business should require a direction. If a business desires to achieve certain objectives, it must no longer be assumed that a computer system knows its direction if it uses machine learning algorithms. It must be controlled that the learning machine actually works in the direction set by the business strategy. The empirical finding is not unambiguous. The thesis author's conclusion is that the "no" answers by the questionnaire may be because of misinterpretation of the question. The conclusion is that the MCS principle on *lack of direction* is relevant when deploying learning machines and the hypothesis H3 is true.

5.5 Result control: Providing rewards (incentives)

The hypothesis for the MCS aspect of result control and providing rewards is that:

> H7: A learning machine will not need incentives to continue its task.

What arguments exists for this hypothesis? Will a learning machine need rewards (incentives) to continue with its task? According to the data found in questionnaire, there is mixed evidence arguing that learning machines would need incentives to carry on its tasks, i.e. the hypothesis H7 may be faulty. Gustavsson (2017) states "no" to the question F5, while Rubins (2017) points out that on a general level no argument support the need of incentives in machine learning. But, he also states that some specific algorithms exist that use incentive-based training. "*Reinforcement learning*" is mentioned by (Rubins 2017) as an example of such incentive-based training within machine learning. From the theory chapters, Sutton & Barto (2012) define *reinforcement learning* as machine learning when "agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments". Sutton & Barto (2012), states that the result of this type of machine learning is based on feedback-loops providing guidance to optimization.

Reinforcement learning is defined as the opposite of *supervised learning* when machines learn from examples provided as input.

Considering the explanation of *Supervised and Unsupervised learning* by Louridas and Ebert (2016) highlighting that the later uses data without given answers, and that the machine itself must learn to find the answer from data patterns, *reinforcement learning* can be defined as a branch of *Unsupervised learning*.

The findings from the analysis support a conclusion that learning machines do not on a general level need incentives to carry on its tasks. However, in specific cases when learning machines use an *Unsupervised learning* approach, incentives or feedback loops are vital to optimize the results from its actions. Also, considering the analysis and conclusion made in chapter 5.2 regarding motivation on complex learning machine systems, incentives to motivate cannot be out ruled. It is highlighted that the fading focus of executing preprogrammed tasks, and increasing focus on learnt skills, justifies that motivation of machine learning algorithms must be considered. With that implication it is concluded in this thesis that the need of incentives in machine learning cannot be ruled out even on a general level, and is thereby relevant. The conclusion is that H7 is false at least for complex systems and systems using *Reinforcement learning*.

5.6 Result control: Employee knowledge of Desired results

This chapter deals with the following hypothesis

> H8: A company can and must control that a learning machine will possess knowledge on: what is the desired business results.

From an MCS point of view, the desired result aspect is different from the earlier discussed aspect of *Lack of direction*, in the way that objectives and targets may be defined within the organization, as defined by the *Lack of direction* aspect, but for *desired result* the organization needs to be able to define the objectives and targets itself. Therefore, the hypothesis must answer the question if learning machines *can* define objectives and targets, and as a next step *must* the MCS control this knowledge.

To answer this hypothesis, empirical data from chapter 4.6 is used. The questionnaire question F6 provides some insights: *Can a company control that a learning machine possess knowledge of: what is the desired business result?* Gustavsson (2017) answer is that learning machine's knowledge can be controlled, without further arguments. This answer provides some value to the hypothesis validation, but other answers also need to be investigated. Rubins (2017) somewhat coincides with Gustavsson, stating that "a machine can certainly interpret a simple natural language query like "what is the desired business result?", parse it, and recall a response using learned algorithms and can do so with surprising accuracy." However, Rubins (2017) withholds that knowledge itself is a human concept, and that machines operate according to programs written by humans. Even if a machine learning algorithm can provide answers, make decision based on learning and optimize results of defined parameters it "does not inherently *know* something or possess knowledge" (Rubins, 2017).

As a summary, Rubins gives a somewhat more diffuse answer to the "can" aspect of the hypothesis. Is it sufficient that a machine can generate a definition of desired result through a learnt algorithm, or must it possess knowledge like humans? If the later is required, what is it that humans possess? Gut feeling? Kahneman (2011) deals with gut feelings in the theory of thinking

fast and slow, where he concludes that gut feeling is nothing else than using system one for recalling learnt knowledge without analyzing its correctness, while system two attempts to analyze the task to be solved by using learnt knowledge. The thesis author's interpretation is that machine learning works like Kahneman's system two.

Tufeksci (2016) explain the machine learning concept as "taking a system and feeding it lots of data, including unstructured data, like the kind we generate in our digital lives. And the system learns by churning through this data." She argues for problems regarding interpretation when machine learning algorithms provide either false or accurate information due to the difficulty of separating which is which in subjective problem. "We don't know what this thing is thinking." (Tufeksci, 2016).

Both Rubins (2017) and Tufeksci (2016) withhold that knowledge in terms of how humans define it is vague in the sense of machine learning. They claim that learning algorithms can interpret information, and are based on new input able to provide new answers. This could be referred to as a type of knowledge, and therefore the ability part of the hypothesis H8 is validated. But, the hypothesis also suggests that a business *must* control the knowledge of the desired business results the machine learning algorithms possess. Considering the difficulty of anticipating what the exact answer to a question will be, and how the algorithm combine information to provide such answer, it becomes a challenge to actually know and control the level of knowledge of desired business result.

Bostrom (2015) provide additional arguments supporting the difficulty of controlling what we refer to as knowledge in learning machines. He claims that it is unrealistic of humans to believe that we can control a learning machine by keep it "locked up in bottle" and within defined boundaries, as it will eventually use its intelligence to find its way out. However, he is concerned with this issue as he does not see how the knowledge and way of thinking by learning machines can ever be fully controlled. He express his worries regarding making sure that learning machines share our beliefs and values as he states:

> "Making superintelligent A.I. is a really hard challenge. Making superintelligent A.I. that is safe involves some additional challenge on top of that. The risk is that if somebody figures out how to crack the first challenge without also having cracked the additional challenge of ensuring perfect safety. This to me looks like a thing that is well worth doing and I can imagine that if things turn out okay, that people a million years from now look back at this century and it might well be that they say that the one thing we did that really mattered was to get this thing right" (Bostrom 2015).

Research shows many examples of learning machines providing results aligned with desired results. Howard (2016) describes the example from pathology where a learning machine was used to diagnose cancer and where the learning algorithms developed new ways of thinking and what to focus on. "The computer system actually discovered that the cells around the cancer are as important as the cancer cells themselves in making a diagnosis" (Howard, 2016). This knowledge was not previously known to pathologists (Howard, 2016).

Learning machines have also shown the ability to provide the opposite of predefined wanted behavior of the organization using it. One example is the chat-based AI experiments conducted at Microsoft referred to by Tikka (2016) where a recent version of the AI called Tay was turned into a racist while being consciously provoked by humans.

Conclusions drawn from these arguments are that controlling whether a learning machine builds knowledge of what is considered desired results, and learn in ways which are aligned with predefined values is still an undeveloped area. Research show examples of when results from machine learning contribute in alignment to the wanted position, either positively or negatively according to human values, depending on the definition of values. This provide us with the insight that it is possible to affect the ways in which machine learning algorithms evolve and act, but still uncertain whether it is possible to control it. Therefore the H8 hypothesis is only partly true. The MCS *must* control that a learning machine will possess knowledge on: what is the desired business results, but it is not verified that a company *can*. Although it is unclear how the control shall be implemented, the MCS aspect *Employee knowledge of Desired results* is therefore relevant also for a deployed learning machine.

5.7 Result control: Employee have Significant influence

How can a company secure that an IA is able to influence the desired results? This chapter will analyze that question by reviewing empirical findings from chapter 4.7 on the hypothesis:

> H9: A company can and must control that a learning machine is able to influence the desired result.

Rubins (2017) argues that when AI becomes more advanced, it brings focus to the question of agency and autonomy, where autonomy is defined by Russell & Norvig (1995) as when systems uses its own experience when creating results, instead of prebuilt knowledge. Rubins (2017) state "Without Moral Agency, machine intelligence is truly just a set of algorithms that require human prompting or execution". He specifically highlights the importance of when companies begin to develop and use advanced algorithms of machine learning type, Review Boards or similar concepts is needed in order to secure a future development in line with company ethics. He refers both to the Institutional Review Board used in Medicine, as well to his represented company in the questionnaire which has added "ethics review" into the process of software release, where it is now mandatory to involve independent human sign off when deploying new codes (Rubins, 2017). This implies that Rubins supports the hypothesis H9 fully.

The machine learning algorithm developed by Facebook referred to as "M" is another example of how the output from a learning machine is reviewed. This algorithm uses information from conversations in messages looking for keywords in order to identify how it can assist in providing useful tools and guidance in pop-ups directly in your conversation window. In this case there is a control-board constantly watching the conversations which "M" in part of in order to control the usability and information provided by the learning machine (Statt, 2017). With other words, Facebook recognizes the need to control and has also developed methods to secure that the learning machine is able to influence the desired results, as stated by the hypothesis H9. If the machine produced not desired results, the human team intervenes and corrects the learning machine until it is able to produce the desired result.

Additionally, to support this argument (Harris, 2016) argues for a potential need of a new "Manhattan project on the topic of Artificial Intelligence". The main reasons for his concern and implications are to avoid arms race and to make sure that Artificial Intelligence is not developed

against human interests (Harris, 2016). Again, the suggested Manhattan project would do as the hypothesis H9 predicts and control the evolvement of learning machines.

The implications is that Harris (2016) creates another dimension to the question of which AI's to deploy when referring to the Manhattan Project (The Development program of Nuclear weapons initiated during World War II). He claims that human interest itself, as being a vital guidance in developing AI, constitutes a great challenge to define and unite among us humans considering the economic and political landscape in the world. With that said, he points out an issue for humanity to consider in the deployment of machine learning, which itself becomes an interesting field of study in the future (Harris 2016).

It is concluded based on this analysis that from a human perspective there is an extensive need of directing the deployment of learning machines to secure that they are able to influence the desired results. Empirical data concludes that the current used methodology to check the ability of learning machines consists of reviews. Therefore the hypothesis H9 is validated within this analysis and the MCS principle to "have Significant influence" remains relevant also when deploying learning machines.

5.8 Action-control: Preaction reviews

This analysis relies on empirical data from chapter 4.8. To understand the preaction review of an MCS in this thesis, the following hypothesis will be analyzed:

H12: A company can and must review a learning machine before deploying it in action

From the questionnaire, Gustavsson (2017) claims that the results from learning machines will be predictable, without further explanation. This completely falsifies the hypothesis, but there may as well be a misunderstanding of the situation of the question. Again, Rubins (2017) provides more details in his arguments and explains that his company uses a human reviewing system on the output from machine learning algorithms. He states that "We do this to make sure the predictive quality of our algorithms is sufficient for the task, rather than blindly turning over a potentially faulty prediction to an unwitting user." He also highlights that this is possible in their area of business but that it may not be the case elsewhere. This statement actually fully confirms the hypothesis H12.

As included in the analysis of the previous chapter, regarding the importance that machine learning algorithms are reviewed by a "moral agency", using Review Boards or similar concepts in order to secure outcome aligned with desired results is equally as relevant in this section. Several examples of machine learning algorithms are also shown in previous sections where the working result is proven to be successful in terms of being aligned with desired results, for example the diagnostic ability of cancer within Pathology presented by (Howard 2016).

An additional example also mentioned in the previous chapter written by Statt (2017), is the machine learning algorithm developed by Facebook referred to as "M" which uses information from conversations in messages in order to identify how it can assist in providing useful guidance directly in your conversation window. What is interesting for this analysis is that "M" in this case has a team of humans watching the conversations and pop-ups taken in action in order to make sure that the result is in line with the company is considered value creating. This also proves the H12 to be true.

There are however also examples of machine learning algorithms where the result is not successful in terms of neither predicted outcome nor wanted result. The IBM intelligent machine system "Watson" example highlighted by (Tufekci, 2016) where the system was used in a Jeopardy game and provided a wrongful answer. In this case, the system was basically requested to provide an answer to a geographical question of a city within the United States, where the system answered Toronto which is a city in Canada. In this case it is argued that this was a mistake which most grown up humans wouldn't do considering basic knowledge and interpretation of information.

Considering the fact that humans use Emotional Intelligence in decision-making where Damácio (1994) in the theory chapter states that we can "anticipate feelings arising if certain things would happen, and its power to assist in the choice of different options." It is also argued by (George, 2000) that humans use emotions in the creative thinking process and that this ability is claimed to help us make sure that decisions are made according to our moral values and ethics. This is something that we learn by growing up as humans, and can in that sense be considered as an initial human advantage over learning machines.

In alignment with the use of Emotional Intelligence and practice of humans values, both Tufekci (2016) and Harris (2016) argue that the incorporation of human values and the direction of humans in the deployment of AI is important in order to secure a long term development aligned with desired results from a machine learning algorithm.

Based on this analysis it is concluded that the result of machine learning algorithm is often proven to be in line with desired result, and can to a high extent be considered predictable. However, considering existing examples where the result is not in line with expectations and the implication that algorithms lack the Emotional Intelligence of a human in its decision-making, one cannot despite high predictability conclude that the results will always be predictable. Gustavsson is therefore overruled by the majority of the other empirical sources. Additionally, considering the importance of protecting human values elaborated by several sources included in this section, the conclusion is that the results from machine learning algorithms need to be reviewed. Therefore, the hypothesis H12 is positively validated and the MCS aspect of *preaction review* is relevant for all results provided by a learning machine.

5.9 Personnel control: Able and capable

This chapter uses empirical findings from chapter 4.9, and analysis the hypothesis:

H16: A company can and must control the *Ability and capability* a learning machine?

The questionnaire gives some insight to this hypothesis while answering the question F9. Gustavsson (2017) states that learning machines can be verified by comparing desired results with achieved results and also that it will be done. This statement correlates with the hypothesis. Rubins (2017) provides further information and refers to F1 score and p-values as indicators of ability and capability of the algorithms. These indicators help providing answers if the training set represents the desired outcome or if certain properties in the training material give a too large impact on the desired results. Rubins (2017) thereby provides a method for the H16 hypothesis. He further claims that his company also performs these checks, but gives no information if it is a must.

Methods to make evaluations on ability and capability are continuously evolving, and Powers (2011) provides several other method suggestions. Though, the two examples provided by Laurent, Chollet and Herzberg (2015) emphasizes that the evaluation may not be easy for businesses where the desired results or internal processes are not well defined. How do you evaluate the ability of a machine learning system when the desired results are not well structured in processes? If today's businesses operates in such areas where activities and desired results is dependent on individual humans skills and the quality of their efforts is not well defined, evaluating the ability of learning machines may initially be difficult. A parallel with deploying humans for a work task, which is not well defined regarding activities or result is that it becomes difficult to evaluate their ability and capability. For both humans and learning machines, it is possible to control the *Ability and capability* as given by H16, but not necessarily a must for the survival of a business. But as long as the business wants to achieve a controlled superior result, it must also control the ability of a learning machine.

The conclusion must be that the ability and capability evaluation of a learning machine can and must be a central part of a MCS and therefore the hypothesis H16 is proven true. And, from a MCS perspective the evaluation methods probably have more similarities with traditional engineering verification than traditional MCS personnel control of Humans. Evaluating capability and ability of learning machines as given by the above examples, can to a large extent be automated and conducted on a regular basis. The evaluation method and questions intended to control learning machines will be more precise compared to the MCS personnel control methods like hiring interviews and performance appraisals.

5.10 Cultural-control

This chapter relies on empirical data from chapter 4.10 and deals with these three hypotheses:

H21: A company can and must control the *Attitudes* of a learning machine?
H22: A company can and must control the *Ways of behaving* of a learning machine?
H23: A company can and must control the *Values* of a learning machine?

The racist chatbot Tay clearly shows the need to control a learning machines *attitudes, values and ways of behaving*. As a learning machine may learn new values, not considered while deploying the machine, a business must have a MCS in place to control how the learning machine develops. Imagine that Tay was deployed at customer-service of a business, to handle customer complaints on a regular basis. While humans in the same position are influenced by laws, society vales etc., a learning machine without a proper MCS, could be influenced 24/7 by complaining customers. What effects would racist values at customer service have on such business? Or worse, what if it could actually identify feminist customers and find means to make the "burn in hell" as described in chapter 4.10?

Harris (2016) illustration on ants and the future power of learning machine intelligence put the cultural-control in an expanded context. As Harris (2016) concludes that learning machines will become more intelligent than humans, they may as well develop the same values towards humans, as humans today value ants. He refers to these machines as super intelligent machines. From a business perspective, the main concern must be to have a powerful MCS that controls the values of this super intelligent machine, to avoid any unethical acts of the business. If a super intelligent machine deployed by a business, deliberately annihilate humans like humans annihilate

ants, this is a violation of human ethical values. Presumably also in the future. While there are no contemporary learning machines that supports Harris extracted future description, he illustrates that a super intelligent machines, based on machine learning, cannot be compared with today's programed computer system that don't learn.

The methods to actually control learning machines' *attitudes*, *values* and *ways of behaving* over time is probably more difficult and no academic text on this topic could be found within the research of this thesis. As Tikka (2016) writes, humans collect their values when they are brought up by their parents, school and community, it is another issue to assure the same training for an AI.

Questionnaire answer from Gustavsson (2017) suggests that deliberate control and changing of algorithms must be done to uphold cultural control. This is directly urging a need for an MCS to handle *values* and *ways of behaving*. Rubins (2017) concludes that there is a need to manage cultural control, but argues that it's not up to a company or government to control this, but instead some expert function outside the corporate hierarchy. He suggests that independent code review should be used and that judicial systems must have access and redress mechanisms. This opens up a completely new area of possible questions. Who is judicial responsible for any criminal act made by a learning machine? The deployment company, the algorithm trainer or the algorithm itself? While humans continue to develop new learning machines, a precedent can be taken from any criminal act with today's software. But what happens when learning machines evolve and start developing learning machines of their own? This question remains unanswered in this thesis, but some research on this topic has been performed by Čerka, Grigienė and Sirbikytė (2017).

The conclusion by the thesis authors is that the need for a business to control *Attitudes*, *values* and *ways of behaving* of a learning machine is obvious. But there are few examples how business can control it. Therefore the hypothesis H21, H22 and H23 are just partly positively validated. Even though there may grow external review and judicial systems, a company must apply their values on learning machines. Most likely, a company must even extend their definitions of correct *attitudes*, *values* and *ways of behaving*, to compensate for the society values humans today bring into a company for free. While Microsoft continues to develop their algorithms to prevent what happened with Tay, a business that deploys an AI, must have a good MCS system in place to handle such failures of a learning algorithm. No business today, can afford to deploy a learning machine that conducts unethical acts. Because these acts would threat to put the company out of business. The future will show which MCS methods will be successful to manage this task.

5.11 Productivity, quantifying the driving force

Although productivity is not a principle of an MCS and therefore not in the scope of this thesis research question, it will be analyzed because the problem discussion pointed it out as the main reason to deploy learning machines within a company. Then, if productivity is the main reason, what aspects must be considered? This chapter relies on empirical findings from chapter 4.11. To begin with, the productivity improvement can be viewed as a force that strong that Howard (2016) and other talks from TED predict a machine learning revolution that will change both social and economic structures in society:

> "The Machine Learning Revolution is going to be very different from the Industrial Revolution, because the Machine Learning Revolution, it never settles down. The better

> computers get at intellectual activities, the more they can build better computers to be better at intellectual capabilities, so this is going to be a kind of change that the world has actually never experienced before, so your previous understanding of what's possible is different. This is already impacting us. In the last 25 years, as capital productivity has increased, labor productivity has been flat, in fact even a little bit down." (Howard, 2016)

In this statement Howard (2016) suggests that learning machines shall be considered a capital investment only impacting capital productivity. The argument for this is that a learning machine does not require any salary and therefore no labour cost. While the theory of productivity states that TFP is the dominant concept for estimating productivity, this thesis will mostly use LP for evaluation purpose evaluation. The reason is that literature search has not provided any estimates of capital investment to implement learning machines in a business. The thesis authors can only speculate in the reasons why, but it seem likely that this is an immature industry that don't yet have best practice on capital investment estimations of learning machines and therefore don't share information on the development and installation costs of a learning machine. These costs may even be a part of their business intellectual property (IP). As a result, although the LP calculations assume a learning machine has been developed and installed, that actual deployment costs are neglected in the calculations. Additionally, this analysis will focus on relative LP evaluation, and not absolute values. I.e. the analysis will compare productivity of a learning machine relative to productivity of a human.

$$Relative\ LP\ Gain = LP_{Learning\ machine}/LP_{Human-only} \approx labour_{Human-only}/labour_{Learning\ machine} \quad (4)$$

A relative analysis uses equation (1) to derive equation (4), assuming that the "Output Index" is unchanged. The first example of Howard (2016) concerning traffic signs provides a humble Relative LP Gain of 2. The Google example by Howard (2016) provides an Relative LP Gain of at least 40 thousand according to equation (5), assuming no vacation but 8 hours working days, 5 days a week and that many year are at least 2.

$$Relative\ LP\ Gain \approx 12 * 2 * 52 * 5 * 8\ /2 = 40 * 10^3 \quad (5)$$

The medical diagnostic example by Howard 2016 provides an Relative LP Gain of at about 350 thousand according to equation (6), also assuming no vacation but 8 hours working days, 5 days a week.

$$Relative\ LP\ Gain \approx 6 * 7 * 52 * 5 * 8 * 60/15 = 349 * 10^3 \quad (6)$$

The future example by Harris (2016) becomes a little bit more complicated. As the replacing learning machine, does not require any labour cost, and while the capital cost for Stanford/MIT to deploy the learning machine is unknown, the thesis needs to make another assumption. In this case it can instead be assumed that the output index in equation (1) is the dominant factors Relative to Capital or labour. This gives an approximation of the Relative LP Gain according to equation (7).

$$Relative\ LP\ Gain \approx Output\ Index_{Learning\ machine}/Output\ Index_{Human-only} \quad (7)$$

Applying equation (7) on the Stanford/MIT researcher replacement of a learning machine, would achieve a relative LP Gain of about 1 million according to calculation (8)

$$Relative\ LP\ Gain \approx 20000 * 52/1 = 1.04 * 10^6 \quad (8)$$

A conclusion from these empirical examples is that with today's knowledge, there are multiple application areas where machine learning technology achieves Relative LP Gains between 2 to 1 million. Although the initial capital investment cost may be high, the large LP gains indicates a very high driving force for any business to invest in this new technology in order to profit from its benefits. As discussed in the theory chapter, productivity improvement was in the late 20th century achieved by lean principles (Womack, Jones, Roos, 1990) with quality improvements as a driving force. But as a conclusion of this subchapter, the driving force in 21th century businesses will be productivity improvements created by learning machines. This is an extension of the paradigm shift in productivity stated by Krone (2014), which was created in the 1990's when Microsoft connected computer systems.

Referring back to the conclusion by Krone (2014), that productivity increase is not always positive for humanity, this statement could also become true for the predicted machine learning revolution. Though, the implication of this machine learning productivity force on social and economic structures in society is outside the scope of this thesis.

6 Conclusions and Implications

The aim of this thesis is to investigate: "Which Management Control System principles and aspects are relevant when deploying a learning machine?". The analysis has made several assumptions and conclusion. Initial assumptions are that several MCS principles are indisputably relevant also when deploying learning machines. For Result-control, these undisputable aspects are: *1.Definition of performance dimensions, 2. Measurement of performance, 3. Setting performance targets* and *Capability to effectively measure.* Additional assumptions on relevant MCS principles are Action-control related: *Behavioral Constraints, Redundancy.* While on the contrary, *Action accountability* is assumed to be an irrelevant aspect of MCS and it does not need to control learning machines. Further, *Personnel control: Clarifies* and *Self-monitoring* is assumed to be relevant MCS principles to control a learning machine. The *Cultural-control of Group norms, Beliefs and Ideologies* are assumed to be relevant principles for learning machines, as these aspects are considered to be associated with human traits only.

For the remaining MCS principles identified in the theory chapter, the analysis has provided conclusions whether these are relevant or not, when deploying a learning machine. The conclusion for *Human issues on Motivation* is that a MCS must consider this principle also for learning machines. The hypothesis H1 was positively validated in the analysis. Aspects on motivation in this sense are about the interest, or the choice to execute one learned skill instead of any other skill. A learning machine may possess many skills with competing interest to be executed. The motivation for a learning machine to choose to execute one skill must be controlled by a MCS with the same reasoning as human's motivation must be controlled. As an example, choosing only to present "like" friendly post on Facebook is quite different from presenting urgent information from friends. An unmotivated learning machine may focus on tasks not supported by a business.

Another principle of MCS, namely the *Human issue on Personal Limitations* has been analyzed. As with humans, a learning machine may possess more or less skills than required by a job task. These skills are partly controlled by the teaching material, configuration and peripherals. However, the material alone does not guarantee that the machine will not learn other skills, as the google GNMT "zero-shot" translation example shows. Therefore an aspect of the principle

is that a learning machine will not intrinsically know its limitations, if it for example proposes new methods to produce desired results. The hypothesis H2 was positively validated in the analysis. Therefore, a business must have a MCS in place to secure that a learning machine is capable of performing as expected, but also as important, that it operates within its desired limitations.

The human aspect on *lack of direction* is another MCS principle where a learning machine must be controlled by a MCS. The positive validation of the hypothesis H3 proves this. Even though a learning machine may produce results without a given direction, it may not be the desired direction. If a business desires to achieve certain objectives, it must not be assumed that learning machine algorithms always knows its direction. It must be controlled that the learning machine actually works in the direction set by the business strategy.

The empirical findings in chapter 4.5 support a conclusion regarding *Result control: Providing rewards (incentives),* that learning machines do not on a general level need incentives to carry on its tasks. This was also the initial assumption of the hypothesis H7. Though, the analysis in chapter 5.5 implies that when *Unsupervised learning* algorithms are used, such incentives and feedback loops are relevant to optimize results. Additionally, considering the analysis and conclusion made regarding motivation and machine learning, the empirical data is not considered sufficient by the authors of this thesis to positively prove hypothesis H7. The declining execution of preprogrammed tasks, and increasing use of learnt skills, justifies the analysis conclusion that motivation of machine learning algorithms must be considered. With that implication it is concluded in this thesis that providing incentives for machine learning cannot be ruled out even on a general level, and this MCS aspect is thereby relevant. The hypothesis H7 was proven false in the analysis.

Regarding *Employee knowledge and influence of Desired results* it is concluded that controlling either the knowledge built by a learning machine of what is considered desired results, or the exact way in which that knowledge is built, is an undeveloped area. It is proven that machine learning algorithms can provide results in line with what is desired, despite how that result is defined. It is thereby concluded in this thesis that it is possible to affect the ways in which machine learning algorithms evolve and act, but uncertain whether it can be controlled. The analysis clarifies the need of human influence to secure that machine learning algorithms evolve with human values incorporated. It is also concluded that directing the deployment of AI from a human perspective is vital. Due to this, the hypothesis H8 could only be partly validated, while H9 was fully validated. Although it is unclear how the knowledge control in H8 shall be implemented, both MCS aspect *Employee knowledge and influence of Desired results* are relevant also for a deployed learning machine.

When considering *Action-control: Pre Action reviews*, machine learning algorithm results are occasionally proven to be in line with what is desired, and can thereby to a high extent be considered predictable. However, that is not the outcome of all machine learning algorithms cases. Considering also the obvious absence of Emotional Intelligence of a human in decision-making by machine learning algorithms, it is concluded in this thesis that the results from machine learning algorithms need to be reviewed. Hence, the hypothesis H12 is positively validated and this MCS aspect is relevant for all results provided by a learning machine.

The analysis conclusion is also that the Personnel control aspect *Ability and Capability* must be a central part of a MCS and be regularly evaluated when deploying learning machines. The evaluation of capability and ability of learning machines can likely be automated in a much larger

extent than for humans. And the evaluation method and questions intended to control learning machines will probably be more precise compared to the MCS personnel control methods for humans. The hypothesis H16 is proven true, and the MCS principle of ability and capability remains relevant when deploying machine learning.

When it comes to *Attitudes*, *values* and *ways of behaving*, MCS must secure control means for learning machines. Ideas exists that these principles should be regulated, controlled by a third party or be controlled by the providers of a learning machine. However, until such controls are efficiently in place, a business must have a MCS in place that controls any such misbehavior that could cause an unethical act. The hypotheses H21, H22 and H23 are only partly positively validated, as there were no empirical data showing that it is possible to control *Attitudes*, *values* and *ways of behaving* of a learning machine. These MCS principles remain relevant when deploying machine learning.

The conclusion from this study is that the productivity improvement provided by learning machines is a strong force. Estimated examples provide relative improvements of a factor 2 to 1 million. Additional considerations are that replacing human wages with learning machine investments may be profitable, but no examples on these actual investment cost have been found it this research. The productivity improvements previously achieved during the industrial revolution through the ideas of Fayol (1949), Taylor (1911) and lean (Womack, Jones, Roos, 1990) will probably continue and be accompanied by the deployment of learning machines.

The implication of this thesis on managers/users/policy makers is that there exists a strong productivity force to implement learning machines in business, wherever possible. Several references describe this as a machine learning revolution which will be quite different from the initial industrial revolution, both on the social and the economic structure of society. At the same time, the theory of MCS which was originally developed for humans, has many relevant principles and aspects also to control learning machines. It is wrong to think that learning machines are just ordinary automation tools. Instead learning machines will possess many of the control uncertainties related to humans. Many of these contemporary MCS principles must be used but it is not always true that they can be followed, as control methods are not yet developed. In this regard the hypotheses H8, H9, H21, H22 and H23 are of particular interest.

Therefore, further research should be conducted on the principles on *knowledge and influence of desired results* for a deployed learning machine. The research should answer what methodology can be used to control this principle? Also, as no empirical data was found to prove that it is possible to control *Attitudes*, *values* and *ways of behaving* of a learning machine, further research should be done to find possible methods for this. Related to the above research areas, the judicial responsibility aspect for any criminal act made by a learning machine must be better understood. Regarding the causes of MCS for humans, the MCS principle on *motivation* also need further research for learning machines. At what complexity levels of a learning machine is it anticipated that motivational aspects will become relevant, such as defined by Maslow (1942)? And how will these motivational aspects look like. This thesis has shown that the motivational principles as incentives are already used to improve the performance of learning machines. How will future motivation tools on learning machines be used to control their behavior?

The most obvious limitation of this thesis study is the limited input from contemporary companies using learning machine technology. Although many large companies such as Google, Facebook etc. uses the technology, they are not public in to what extent they use the technology

and what MCS they use to control it. Larger amounts of primary empirical data on best-practices from companies using the technology, would provide better insight on what MCS principles are relevant and also how they could be controlled. However, as long as companies consider this information as their IP, it will be difficult to make unambiguous conclusions. Another limitation is how the secondary empirical data is used. The found data provides examples of learning machines usage in relation to MCS principles, but the actual principles are seldom directly quoted in the data. This opens for misinterpretations of the data in this thesis analysis.

At last, although chapter 1.4 already excluded extended principles of an MCS to control learning machines, this is obviously a limitation for a MCS that strives to use the full potential of learning machines. As already mentioned, the control principles *attitudes, values* and *ways of behaving* are not obvious for learning machines mainly because of the gap between humans education in a society before becoming employed within a company. Should a MCS for learning machines be developed completely different because of this gap? Or, must this gap be filled with regulations to avoid a machine learning conflict with humans values? These limitations remain unexplored within this thesis.

7 References:

1. Beveren, 2010, TOTAL FACTOR PRODUCTIVITY ESTIMATION: A PRACTICAL REVIEW, Volume 26, Issue 1, February 2012 pp. 98–128, Journal of Economic Surveys

2. Bose & Mahapatra, 2001, Business data mining — a machine learning perspective, Volume 39, Issue 3- selected pp. 165-254 Information & Management

3. Catasus B., 2007, What gets measured gets … on indicating, mobilizing and acting, accounting, auditing & accountability Journal Vol.20 No. 4 2007, pp. 505-521, Emerald Group Publishing

4. Čerka, Grigienė and Sirbikytė, 2017, Is it possible to grant legal personality to artificial intelligence software systems?, Computer Law & Security Review, Retrieved July 09, 2017 from, Retrieved from https://doi.org/10.1016/j.clsr.2017.03.022

5. Cooper, H., 2010, Research Synthesis & Meta Analysis, A step-by-step Approach, 4th Edition, Applied Social Research Methods Series, Vol 2

6. Damasio, A.R, 1994, Descartes' Error: Emotion, Reason, and the Human Brain, Avon Books, New York

7. Dawkins Richard, 2016, The Selfish Gene, 40th Edition, OUP Oxford

8. Fayol H., 1949, General and Industrial Management, Sir Isaac Pitman & Sons, London (translated by Constance Storrs).

9. George, J.M, 2000, Emotions and Leadership: The Role of Emotional Intelligence, Human Relations, vol 53(8), p 1027 - 1055

10. Golafshani, N., 2003. Understanding Reliability and Validity in Qualitative Research. The Qualitative Report, 8(4), 597-606. Retrieved from http://nsuworks.nova.edu/tqr/vol8/iss4/6

11. Goldbloom A., 2016, The jobs we'll lose to machines — and the ones we won't, Retrieved January 29, 2017 from TED website: https://www.ted.com/talks/anthony_goldbloom_the_jobs_we_ll_lose_to_machines_and_the_ones_we_won_t

12. Harris Sam, 2016, Can we build AI without losing control over it?, Retrieved January 29, 2017 from TED website: https://www.ted.com/talks/sam_harris_can_we_build_ai_without_losing_control_over_it/transcript?language=en

13. Haustein E., Luther R. and Schuster P, 2014, Management control systems in innovation companies: a literature based framework, Journal of Management Control (2014) 24:343-382, DOI 10.1007/s00187-014-0187-5

14. Howard Jeremy, 2016, The wonderful and terrifying implications of computers that can learn, Retrieved January 29, 2017 from TED website: : https://www.ted.com/talks/jeremy_howard_the_wonderful_and_terrifying_implications_of_computers_that_can_learn

15. Kaplan and Norton, 1991, The balanced scorecard - Measures that drive performance, Harvard Business Review

16. Kahneman D, 2011, Thinking Fast and Slow, Farrar Straus Giroux

17. Keat, Young and Erfle, 2014. Managerial Economics, Seventh edition, Global Edition. Pearson Education: Essex.

18. Kelly Kevin, 2016, How AI can bring on a second Industrial Revolution, Retrieved January 29, 2017 from TED website: https://www.ted.com/talks/kevin_kelly_how_ai_can_bring_on_a_second_industrial_revolution

19. Krone, Robert. M, 2014, Increasing workforce productivity: smarter people and machines, Int. J. Human Resources Development and Management, Vol. 14, Nos. 1/2/3,

20. Kurzweil, R., 1990, The age of Intelligent Machines

21. Laurent, Chollet and Herzberg, 2015, Intelligent automation entering the business world, Inside - Quarterly insights from Deloitte, issue 8 2015, Deloitte

22. Louridas and Ebert, 2016, Machine Learning, Software technology Sep/Oct 2016, pp. 110-115, IEEE Computer Society

23. Luger, G.F. 2005, Artificial Intelligence, Structures and Strategies for Complex Problem Solving, 5th Edition

24. Maslow, A.H., 1942, A theory of human motivation, British journal of psychiatry. Volume 208 Issue 4, pp 313

25. Mayer, J.D & Salovey P., 1997, What is Emotional Intelligence: Implications for Educators, p 3-31

26. McGregor D., 2006, The Human Side of Enterprise, Annotated Edition 1st Edition, McGraw-Hill Education

27. Merchant, K.A. & van der Stede, W.A., 2012 Management Control Systems - Performance Measurement, Evaluation and Incentives, Third Edition, Pearson Education

28. Mosesson, 2017, Svenska annonspengar går till fejksajterna, www.dn.se.uk, Retrieved february 19, 2017 from DN website: http://www.dn.se/ekonomi/svenska-annonspengar-gar-till-fejksajterna/

29. Mousannif, Moatassime and Noel, 2016, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, Volume 83, 2016, Pages 1064–1069, Procedia Computer Science

30. Ohno, Taiichi, 1988, Toyota Production System – Beyond Large-Scale Production. Productivity Press.

31. Papademetriou, C., 2012, To what extend is the Turing test still important, Pliroforiki 1 (22), 28-32

32. Patel, R. & Tebelius, U. (red.), 1987, Grundbok i forskningsmetodik

33. Paton S, 2013, Introducing Taylor to the knowledge economy, Employee Relations vol. 35 No. 1, pp20-38, Emerald Group Publishing Limited

34. Powers David, 2011, Evaluation: From Precision, recall and F-measure to ROC, Informedness, Markedness & Correlation, Journal of Machine Learning Technologies 2(1) 37-63

35. Quinlan, J.R., 1986, Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers

36. Ruijter René de, 2014, Do you suffer from the Pike Syndrome?, Retrieved April 17, 2017 from www.hatrabbits.com website: http://hatrabbits.com/do-you-suffer-from-the-pike-syndrome/

37. Russell, S.J & Norvig, P., 1995, Artificial Intelligence, A Modern Approach, Prentice Hall Inc.

38. Ryan R., Deci E, 2000, Intrinsic and extrinsic motivations: Classic definitions and new directions, contemporary educational psychology 25, pp 54-67, doi:10.1006/ceps.1999.1020

39. Simons, H., 2009, Case study research in practice. London: SAGE publications

40. Shuster, M & Johnson, M & Thorat, N, 2016, Zero-Shot Translation with Google's Multilingual Neural Machine Translation System, Retrieved March 9, 2017 from Google research blog website: https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html

41. Statt Nick, 2017, Facebook's AI assistant will now offer suggestions inside Messenger, Retrieved April 17, 2017 from www.theverge.com website: http://www.theverge.com/2017/4/6/15200836/facebook-messenger-m-suggestions-ai-assisant

42. Stepanova, 2011, The Role of Information Communication Technologies in the "Arab Spring", PONARS Eurasia Policy Memo No. 159 May 2011, Russian Academy of Sciences

43. Strauß, E. & Zecher, C, 2012, Management control systems: a review, Journal of Management Control, pp. 1–36, doi:10.1007/s00187-012-0158-7.

44. Stone D, Nikitkov N, Miller C., 2014, Strategy, IT and control @ eBay, Qualitative Research in accounting & Management, vol.11 No. 4, 2014 pp 357-379, Emerald Group publishing limited

45. Sutton, Richard. S & Barto, Andrew. G, 2012, Reinforcement Learning: An Introduction, The MIT Press Cambridge, Massachusetts London, England

46. Taylor, F.W., 1911, The Principles of Scientific Management. New York, NY: Harper & Brothers. Retrieved from The Project Gutenberg EBook of The Principles of Scientific Management: http://www.manyebooks.org

47. Tieto Oy, 2016, Tieto the first Nordic company to appoint Artificial Intelligence to the leadership team of the new data-driven businesses unit, Retrieved March 12, 2017 from www.tieto.com website: https://www.tieto.com/news/tieto-the-first-nordic-company-to-appoint-artificial-intelligence-to-the-leadership-team-of-the-new

48. Tikka Taneli , 2016, Artificial Intelligence: Are we there yet?, Retrieved April 17, 2017 from www.tieto.com website: https://perspectives.tieto.com/blog/2016/03/artificial-intelligence-are-we-there-yet/

49. Turing, A.M., 1950, Computing machinery and intelligence. Mind, vol 59, p 433-460

50. Zeynep Tufekci, 2016, Machine intelligence makes human morals more important, Retrieved January 29, 2017 from TED website: https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important

51. Vincent James , 2016, Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day, Retrieved April 17, 2017 from www.theverge.com website: http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

52. Winston, P.H., 1992, Artificial Intelligence, Pearson, 3rd Edition

53. Womack James P., Jones Daniel T., & Roos Daniel , 1990, Chapter 3 "The rise of lean production", in The machine that changed the world, Harper Perennial: 48-69

54. World Economic Forum, 2015, The Future of Financial Services, Final report June 2015, website: http://www3.weforum.org/docs/WEF_The_future__of_financial_services.pdf

55. Yin, Robert K., 2003, Case Study Research Design and Methods, 3rd edition.

56. Yingying Zhang, Simon Dolan, Yu Zhou, 2009,"Management by values", Chinese Management Studies, Vol. 3 Iss 4 pp. 272 - 294

57. Zolfagharifard, 2014, Would you take orders from a ROBOT?, www.dailymail.co.uk, Retrieved January 29, 2017 from dailymail website: http://www.dailymail.co.uk/sciencetech/article-2632920/Would-orders-ROBOT-Artificial-intelligence-world-s-company-director-Japan.html

8 Appendices

8.1 Appendix A: TED-talk transcripts

8.1.1 Transcript of Zeynep Tufekci, (2016), Machine intelligence makes human morals more important

"0:11 So, I started my first job as a computer programmer in my very first year of college — basically, as a teenager.
0:19 Soon after I started working, writing software in a company, a manager who worked at the company came down to where I was, and he whispered to me, "Can he tell if I'm lying?" There was nobody else in the room.
0:36 "Can who tell if you're lying? And why are we whispering?"
0:41 The manager pointed at the computer in the room. "Can he tell if I'm lying?" Well, that manager was having an affair with the receptionist.
0:52 (Laughter)
0:54 And I was still a teenager. So I whisper-shouted back to him, "Yes, the computer can tell if you're lying."
1:02 (Laughter)
1:03 Well, I laughed, but actually, the laugh's on me. Nowadays, there are computational systems that can suss out emotional states and even lying from processing human faces. Advertisers and even governments are very interested.
1:21 I had become a computer programmer because I was one of those kids crazy about math and science. But somewhere along the line I'd learned about nuclear weapons, and I'd gotten really concerned with the ethics of science. I was troubled. However, because of family circumstances, I also needed to start working as soon as possible. So I thought to myself, hey, let me pick a technical field where I can get a job easily and where I don't have to deal with any troublesome questions of ethics. So I picked computers.
1:51 (Laughter)
1:52 Well, ha, ha, ha! All the laughs are on me. Nowadays, computer scientists are building platforms that control what a billion people see every day. They're developing cars that could decide who to run over. They're even building machines, weapons, that might kill human beings in war. It's ethics all the way down.
2:18 Machine intelligence is here. We're now using computation to make all sort of decisions, but also new kinds of decisions. We're asking questions to computation that have no single right answers, that are subjective and open-ended and value-laden.

2:35 We're asking questions like, "Who should the company hire?" "Which update from which friend should you be shown?" "Which convict is more likely to reoffend?" "Which news item or movie should be recommended to people?"
2:47 Look, yes, we've been using computers for a while, but this is different. This is a historical twist, because we cannot anchor computation for such subjective decisions the way we can anchor computation for flying airplanes, building bridges, going to the moon. Are airplanes safer? Did the bridge sway and fall? There, we have agreed-upon, fairly clear benchmarks, and we have laws of nature to guide us. We have no such anchors and benchmarks for decisions in messy human affairs.
3:24 To make things more complicated, our software is getting more powerful, but it's also getting less transparent and more complex. Recently, in the past decade, complex algorithms have made great strides. They can recognize human faces. They can decipher handwriting. They can detect credit card fraud and block spam and they can translate between languages. They can detect tumors in medical imaging. They can beat humans in chess and Go.
3:54 Much of this progress comes from a method called "machine learning." Machine learning is different than traditional programming, where you give the computer detailed, exact, painstaking instructions. It's more like you take the system and you feed it lots of data, including unstructured data, like the kind we generate in our digital lives. And the system learns by churning through this data. And also, crucially, these systems don't operate under a single-answer logic. They don't produce a simple answer; it's more probabilistic: "This one is probably more like what you're looking for."
4:31 Now, the upside is: this method is really powerful. The head of Google's AI systems called it, "the unreasonable effectiveness of data." The downside is, we don't really understand what the system learned. In fact, that's its power. This is less like giving instructions to a computer; it's more like training a puppy-machine-creature we don't really understand or control. So this is our problem. It's a problem when this artificial intelligence system gets things wrong. It's also a problem when it gets things right, because we don't even know which is which when it's a subjective problem. We don't know what this thing is thinking.
5:14 So, consider a hiring algorithm — a system used to hire people, using machine-learning systems. Such a system would have been trained on previous employees' data and instructed to find and hire people like the existing high performers in the company. Sounds good. I once attended a conference that brought together human resources managers and executives, high-level people, using such systems in hiring. They were super excited. They thought that this would make hiring more objective, less biased, and give women and minorities a better shot against biased human managers.
5:54 And look — human hiring is biased. I know. I mean, in one of my early jobs as a programmer, my immediate manager would sometimes come down to where I was really early in the morning or really late in the afternoon, and she'd say, "Zeynep, let's go to lunch!" I'd be puzzled by the weird timing. It's 4pm. Lunch? I was broke, so free lunch. I always went. I later realized what was happening. My immediate managers had not confessed to their higher-ups that the programmer they hired for a serious job was a teen girl who wore jeans and sneakers to work. I was doing a good job, I just looked wrong and was the wrong age and gender.
6:40 So hiring in a gender- and race-blind way certainly sounds good to me. But with these systems, it is more complicated, and here's why: Currently, computational systems can infer all sorts of things about you from your digital crumbs, even if you have not disclosed those things. They can infer your sexual orientation, your personality traits, your political leanings. They have predictive power with high levels of accuracy. Remember — for things you haven't even disclosed. This is inference.
7:16 I have a friend who developed such computational systems to predict the likelihood of clinical or postpartum depression from social media data. The results are impressive. Her system can predict the likelihood of depression months before the onset of any symptoms — months before. No symptoms, there's prediction. She hopes it will be used for early intervention. Great! But now put this in the context of hiring.
7:47 eled "higher risk of depression," "higher risk of pregnancy," "aggressive guy scale." Not only do you not know what your system is selecting on, you don't even know where to beSo at this human resources managers conference, I approached a high-level manager in a very large company, and I said to her, "Look, what if, unbeknownst to you, your system is weeding out people with high future likelihood of depression? They're not depressed now, just maybe in the future, more likely. What if it's weeding out women more likely to be pregnant in the next year or two but aren't pregnant now? What if it's hiring aggressive people because that's your workplace culture?" You can't tell this

by looking at gender breakdowns. Those may be balanced. And since this is machine learning, not traditional coding, there is no variable there labgin to look. It's a black box. It has predictive power, but you don't understand it.
8:51 "What safeguards," I asked, "do you have to make sure that your black box isn't doing something shady?" She looked at me as if I had just stepped on 10 puppy tails.
9:03 (Laughter)
9:05 She stared at me and she said, "I don't want to hear another word about this." And she turned around and walked away. Mind you — she wasn't rude. It was clearly: what I don't know isn't my problem, go away, death stare.
9:22 (Laughter)
9:24 Look, such a system may even be less biased than human managers in some ways. And it could make monetary sense. But it could also lead to a steady but stealthy shutting out of the job market of people with higher risk of depression. Is this the kind of society we want to build, without even knowing we've done this, because we turned decision-making to machines we don't totally understand?
9:52 Another problem is this: these systems are often trained on data generated by our actions, human imprints. Well, they could just be reflecting our biases, and these systems could be picking up on our biases and amplifying them and showing them back to us, while we're telling ourselves, "We're just doing objective, neutral computation."
10:17 Researchers found that on Google, women are less likely than men to be shown job ads for high-paying jobs. And searching for African-American names is more likely to bring up ads suggesting criminal history, even when there is none. Such hidden biases and black-box algorithms that researchers uncover sometimes but sometimes we don't know, can have life-altering consequences.
10:48 In Wisconsin, a defendant was sentenced to six years in prison for evading the police. You may not know this, but algorithms are increasingly used in parole and sentencing decisions. He wanted to know: How is this score calculated? It's a commercial black box. The company refused to have its algorithm be challenged in open court. But ProPublica, an investigative nonprofit, audited that very algorithm with what public data they could find, and found that its outcomes were biased and its predictive power was dismal, barely better than chance, and it was wrongly labeling black defendants as future criminals at twice the rate of white defendants.
11:34 So, consider this case: This woman was late picking up her godsister from a school in Broward County, Florida, running down the street with a friend of hers. They spotted an unlocked kid's bike and a scooter on a porch and foolishly jumped on it. As they were speeding off, a woman came out and said, "Hey! That's my kid's bike!" They dropped it, they walked away, but they were arrested.
12:00 She was wrong, she was foolish, but she was also just 18. She had a couple of juvenile misdemeanors. Meanwhile, that man had been arrested for shoplifting in Home Depot — 85 dollars' worth of stuff, a similar petty crime. But he had two prior armed robbery convictions. But the algorithm scored her as high risk, and not him. Two years later, ProPublica found that she had not reoffended. It was just hard to get a job for her with her record. He, on the other hand, did reoffend and is now serving an eight-year prison term for a later crime. Clearly, we need to audit our black boxes and not have them have this kind of unchecked power.
12:45 (Applause)
12:49 Audits are great and important, but they don't solve all our problems. Take Facebook's powerful news feed algorithm — you know, the one that ranks everything and decides what to show you from all the friends and pages you follow. Should you be shown another baby picture?
13:06 (Laughter)
13:07 A sullen note from an acquaintance? An important but difficult news item? There's no right answer. Facebook optimizes for engagement on the site: likes, shares, comments.
13:19 In August of 2014, protests broke out in Ferguson, Missouri, after the killing of an African-American teenager by a white police officer, under murky circumstances. The news of the protests was all over my algorithmically unfiltered Twitter feed, but nowhere on my Facebook. Was it my Facebook friends? I disabled Facebook's algorithm, which is hard because Facebook keeps wanting to make you come under the algorithm's control, and saw that my friends were talking about it. It's just that the algorithm wasn't showing it to me. I researched this and found this was a widespread problem.
13:55 The story of Ferguson wasn't algorithm-friendly. It's not "likable." Who's going to click on "like?" It's not even easy to comment on. Without likes and comments, the algorithm was likely showing it to even fewer people, so we didn't get to see this. Instead, that week, Facebook's algorithm highlighted this, which is the ALS Ice Bucket

Challenge. Worthy cause; dump ice water, donate to charity, fine. But it was super algorithm-friendly. The machine made this decision for us. A very important but difficult conversation might have been smothered, had Facebook been the only channel.
14:35 Now, finally, these systems can also be wrong in ways that don't resemble human systems. Do you guys remember Watson, IBM's machine-intelligence system that wiped the floor with human contestants on Jeopardy? It was a great player. But then, for Final Jeopardy, Watson was asked this question: "Its largest airport is named for a World War II hero, its second-largest for a World War II battle."
14:58 (Hums Final Jeopardy music)
15:00 Chicago. The two humans got it right. Watson, on the other hand, answered "Toronto" — for a US city category! The impressive system also made an error that a human would never make, a second-grader wouldn't make.
15:17 Our machine intelligence can fail in ways that don't fit error patterns of humans, in ways we won't expect and be prepared for. It'd be lousy not to get a job one is qualified for, but it would triple suck if it was because of stack overflow in some subroutine.
15:35 (Laughter)
15:37 In May of 2010, a flash crash on Wall Street fueled by a feedback loop in Wall Street's "sell" algorithm wiped a trillion dollars of value in 36 minutes. I don't even want to think what "error" means in the context of lethal autonomous weapons.
16:00 So yes, humans have always made biases. Decision makers and gatekeepers, in courts, in news, in war ... they make mistakes; but that's exactly my point. We cannot escape these difficult questions. We cannot outsource our responsibilities to machines.
16:21 (Applause)
16:28 Artificial intelligence does not give us a "Get out of ethics free" card.
16:33 Data scientist Fred Benenson calls this math-washing. We need the opposite. We need to cultivate algorithm suspicion, scrutiny and investigation. We need to make sure we have algorithmic accountability, auditing and meaningful transparency. We need to accept that bringing math and computation to messy, value-laden human affairs does not bring objectivity; rather, the complexity of human affairs invades the algorithms. Yes, we can and we should use computation to help us make better decisions. But we have to own up to our moral responsibility to judgment, and use algorithms within that framework, not as a means to abdicate and outsource our responsibilities to one another as human to human.
17:24 Machine intelligence is here. That means we must hold on ever tighter to human values and human ethics.
17:33 Thank you.
17:34 (Applause)" (Tufekci, 2016)

8.1.2 Transcript of Nick Bostrom, 2015, What happens when our computers get smarter than we are?

"0:11 I work with a bunch of mathematicians, philosophers and computer scientists, and we sit around and think about the future of machine intelligence, among other things. Some people think that some of these things are sort of science fiction-y, far out there, crazy. But I like to say, okay, let's look at the modern human condition. (Laughter) This is the normal way for things to be.

0:40 But if we think about it, we are actually recently arrived guests on this planet, the human species. Think about if Earth was created one year ago, the human species, then, would be 10 minutes old. The industrial era started two seconds ago. Another way to look at this is to think of world GDP over the last 10,000 years, I've actually taken the trouble to plot this for you in a graph. It looks like this. (Laughter) It's a curious shape for a normal condition. I sure wouldn't want to sit on it. (Laughter)

1:18 Let's ask ourselves, what is the cause of this current anomaly? Some people would say it's technology. Now it's true, technology has accumulated through human history, and right now, technology advances extremely rapidly —

that is the proximate cause, that's why we are currently so very productive. But I like to think back further to the ultimate cause.

1:44 Look at these two highly distinguished gentlemen: We have Kanzi — he's mastered 200 lexical tokens, an incredible feat. And Ed Witten unleashed the second superstring revolution. If we look under the hood, this is what we find: basically the same thing. One is a little larger, it maybe also has a few tricks in the exact way it's wired. These invisible differences cannot be too complicated, however, because there have only been 250,000 generations since our last common ancestor. We know that complicated mechanisms take a long time to evolve. So a bunch of relatively minor changes take us from Kanzi to Witten, from broken-off tree branches to intercontinental ballistic missiles.

2:31 So this then seems pretty obvious that everything we've achieved, and everything we care about, depends crucially on some relatively minor changes that made the human mind. And the corollary, of course, is that any further changes that could significantly change the substrate of thinking could have potentially enormous consequences.

2:55 Some of my colleagues think we're on the verge of something that could cause a profound change in that substrate, and that is machine superintelligence. Artificial intelligence used to be about putting commands in a box. You would have human programmers that would painstakingly handcraft knowledge items. You build up these expert systems, and they were kind of useful for some purposes, but they were very brittle, you couldn't scale them. Basically, you got out only what you put in. But since then, a paradigm shift has taken place in the field of artificial intelligence.

3:29 Today, the action is really around machine learning. So rather than handcrafting knowledge representations and features, we create algorithms that learn, often from raw perceptual data. Basically the same thing that the human infant does. The result is A.I. that is not limited to one domain — the same system can learn to translate between any pairs of languages, or learn to play any computer game on the Atari console. Now of course, A.I. is still nowhere near having the same powerful, cross-domain ability to learn and plan as a human being has. The cortex still has some algorithmic tricks that we don't yet know how to match in machines.

4:18 So the question is, how far are we from being able to match those tricks? A couple of years ago, we did a survey of some of the world's leading A.I. experts, to see what they think, and one of the questions we asked was, "By which year do you think there is a 50 percent probability that we will have achieved human-level machine intelligence?" We defined human-level here as the ability to perform almost any job at least as well as an adult human, so real human-level, not just within some limited domain. And the median answer was 2040 or 2050, depending on precisely which group of experts we asked. Now, it could happen much, much later, or sooner, the truth is nobody really knows.

5:04 What we do know is that the ultimate limit to information processing in a machine substrate lies far outside the limits in biological tissue. This comes down to physics. A biological neuron fires, maybe, at 200 hertz, 200 times a second. But even a present-day transistor operates at the Gigahertz. Neurons propagate slowly in axons, 100 meters per second, tops. But in computers, signals can travel at the speed of light. There are also size limitations, like a human brain has to fit inside a cranium, but a computer can be the size of a warehouse or larger. So the potential for superintelligence lies dormant in matter, much like the power of the atom lay dormant throughout human history, patiently waiting there until 1945. In this century, scientists may learn to awaken the power of artificial intelligence. And I think we might then see an intelligence explosion.

6:09 Now most people, when they think about what is smart and what is dumb, I think have in mind a picture roughly like this. So at one end we have the village idiot, and then far over at the other side we have Ed Witten, or Albert Einstein, or whoever your favorite guru is. But I think that from the point of view of artificial intelligence, the true picture is actually probably more like this: AI starts out at this point here, at zero intelligence, and then, after many, many years of really hard work, maybe eventually we get to mouse-level artificial intelligence, something that can navigate cluttered environments as well as a mouse can. And then, after many, many more years of really hard work, lots of investment, maybe eventually we get to chimpanzee-level artificial intelligence. And then, after even

more years of really, really hard work,we get to village idiot artificial intelligence. And a few moments later, we are beyond Ed Witten. The train doesn't stop at Humanville Station. It's likely, rather, to swoosh right by.

7:13 Now this has profound implications, particularly when it comes to questions of power. For example, chimpanzees are strong — pound for pound, a chimpanzee is about twice as strong as a fit human male.And yet, the fate of Kanzi and his pals depends a lot more on what we humans do than on what the chimpanzees do themselves. Once there is superintelligence, the fate of humanity may depend on what the superintelligence does. Think about it: Machine intelligence is the last invention that humanity will ever need to make. Machines will then be better at inventing than we are, and they'll be doing so on digital timescales. What this means is basically a telescoping of the future. Think of all the crazy technologies that you could have imagined maybe humans could have developed in the fullness of time:cures for aging, space colonization, self-replicating nanobots or uploading of minds into computers, all kinds of science fiction-y stuff that's nevertheless consistent with the laws of physics. All of this superintelligence could develop, and possibly quite rapidly.

8:23 Now, a superintelligence with such technological maturity would be extremely powerful, and at least in some scenarios, it would be able to get what it wants. We would then have a future that would be shaped by the preferences of this A.I. Now a good question is, what are those preferences? Here it gets trickier.To make any headway with this, we must first of all avoid anthropomorphizing. And this is ironic because every newspaper article about the future of A.I. has a picture of this: So I think what we need to do is to conceive of the issue more abstractly, not in terms of vivid Hollywood scenarios.

9:08 We need to think of intelligence as an optimization process, a process that steers the future into a particular set of configurations. A superintelligence is a really strong optimization process. It's extremely good at using available means to achieve a state in which its goal is realized. This means that there is no necessary connection between being highly intelligent in this sense, and having an objective that we humans would find worthwhile or meaningful.

9:38 Suppose we give an A.I. the goal to make humans smile. When the A.I. is weak, it performs useful or amusing actions that cause its user to smile. When the A.I. becomes superintelligent, it realizes that there is a more effective way to achieve this goal: take control of the world and stick electrodes into the facial muscles of humans to cause constant, beaming grins. Another example, suppose we give A.I. the goal to solve a difficult mathematical problem. When the A.I. becomes superintelligent, it realizes that the most effective way to get the solution to this problem is by transforming the planet into a giant computer, so as to increase its thinking capacity. And notice that this gives the A.I.s an instrumental reason to do things to us that we might not approve of. Human beings in this model are threats, we could prevent the mathematical problem from being solved.

10:28 Of course, perceivably things won't go wrong in these particular ways; these are cartoon examples. But the general point here is important: if you create a really powerful optimization process to maximize for objective x, you better make sure that your definition of x incorporates everything you care about. This is a lesson that's also taught in many a myth. King Midas wishes that everything he touches be turned into gold. He touches his daughter, she turns into gold. He touches his food, it turns into gold. This could become practically relevant, not just as a metaphor for greed, but as an illustration of what happens if you create a powerful optimization process and give it misconceived or poorly specified goals.

11:15 Now you might say, if a computer starts sticking electrodes into people's faces, we'd just shut it off. A, this is not necessarily so easy to do if we've grown dependent on the system — like, where is the off switch to the Internet? B, why haven't the chimpanzees flicked the off switch to humanity, or the Neanderthals? They certainly had reasons. We have an off switch, for example, right here. (Choking) The reason is that we are an intelligent adversary; we can anticipate threats and plan around them. But so could a superintelligent agent, and it would be much better at that than we are. The point is, we should not be confident that we have this under control here.

12:03 And we could try to make our job a little bit easier by, say, putting the A.I. in a box, like a secure software environment, a virtual reality simulation from which it cannot escape. But how confident can we be that the A.I. couldn't find a bug. Given that merely human hackers find bugs all the time, I'd say, probably not very confident. So we disconnect the ethernet cable to create an air gap, but again, like merely human hackers routinely transgress air

gaps using social engineering. Right now, as I speak, I'm sure there is some employee out there somewhere who has been talked into handing out her account details by somebody claiming to be from the I.T. department.

12:45 More creative scenarios are also possible, like if you're the A.I., you can imagine wiggling electrodes around in your internal circuitry to create radio waves that you can use to communicate. Or maybe you could pretend to malfunction, and then when the programmers open you up to see what went wrong with you, they look at the source code — Bam! — the manipulation can take place. Or it could output the blueprint to a really nifty technology, and when we implement it, it has some surreptitious side effect that the A.I. had planned. The point here is that we should not be confident in our ability to keep a superintelligent genie locked up in its bottle forever. Sooner or later, it will out.

13:26I believe that the answer here is to figure out how to create superintelligent A.I. such that even if — when — it escapes, it is still safe because it is fundamentally on our side because it shares our values. I see no way around this difficult problem.

13:43 Now, I'm actually fairly optimistic that this problem can be solved. We wouldn't have to write down a long list of everything we care about, or worse yet, spell it out in some computer language like C++ or Python,that would be a task beyond hopeless. Instead, we would create an A.I. that uses its intelligence to learn what we value, and its motivation system is constructed in such a way that it is motivated to pursue our values or to perform actions that it predicts we would approve of. We would thus leverage its intelligence as much as possible to solve the problem of value-loading.

14:23 This can happen, and the outcome could be very good for humanity. But it doesn't happen automatically.The initial conditions for the intelligence explosion might need to be set up in just the right way if we are to have a controlled detonation. The values that the A.I. has need to match ours, not just in the familiar context, like where we can easily check how the A.I. behaves, but also in all novel contexts that the A.I. might encounter in the indefinite future.

14:53 And there are also some esoteric issues that would need to be solved, sorted out: the exact details of its decision theory, how to deal with logical uncertainty and so forth. So the technical problems that need to be solved to make this work look quite difficult — not as difficult as making a superintelligent A.I., but fairly difficult. Here is the worry: Making superintelligent A.I. is a really hard challenge. Making superintelligent A.I. that is safe involves some additional challenge on top of that. The risk is that if somebody figures out how to crack the first challenge without also having cracked the additional challenge of ensuring perfect safety.

15:36 So I think that we should work out a solution to the control problem in advance, so that we have it available by the time it is needed. Now it might be that we cannot solve the entire control problem in advance because maybe some elements can only be put in place once you know the details of the architecture where it will be implemented. But the more of the control problem that we solve in advance,the better the odds that the transition to the machine intelligence era will go well.

16:05 This to me looks like a thing that is well worth doing and I can imagine that if things turn out okay, that people a million years from now look back at this century and it might well be that they say that the one thing we did that really mattered was to get this thing right.

16:23 Thank you.

16:25(Applause)" (Bostrom, 2015)

8.1.3 Transcript of Jeremy Howard: (2016), The wonderful and terrifying implications of computers that can learn

0:11 It used to be that if you wanted to get a computer to do something new, you would have to program it. Now, programming, for those of you here that haven't done it yourself, requires laying out in excruciating detail every single step that you want the computer to do in order to achieve your goal. Now, if you want to do something that you don't know how to do yourself, then this is going to be a great challenge.

0:35 So this was the challenge faced by this man, Arthur Samuel. In 1956, he wanted to get this computer to be able to beat him at checkers. How can you write a program, lay out in excruciating detail, how to be better than you at checkers? So he came up with an idea: he had the computer play against itself thousands of times and learn how to play checkers. And indeed it worked, and in fact, by 1962, this computer had beaten the Connecticut state champion.

1:06 So Arthur Samuel was the father of machine learning, and I have a great debt to him, because I am a machine learning practitioner. I was the president of Kaggle, a community of over 200,000 machine learning practitioners. Kaggle puts up competitions to try and get them to solve previously unsolved problems, and it's been successful hundreds of times. So from this vantage point, I was able to find out a lot about what machine learning can do in the past, can do today, and what it could do in the future. Perhaps the first big success of machine learning commercially was Google. Google showed that it is possible to find information by using a computer algorithm, and this algorithm is based on machine learning. Since that time, there have been many commercial successes of machine learning. Companies like Amazon and Netflix use machine learning to suggest products that you might like to buy, movies that you might like to watch. Sometimes, it's almost creepy. Companies like LinkedIn and Facebook Sometimes will tell you about who your friends might be and you have no idea how it did it, and this is because it's using the power of machine learning. These are algorithms that have learned how to do this from data rather than being programmed by hand.

2:18 This is also how IBM was successful in getting Watson to beat the two world champions at "Jeopardy,"answering incredibly subtle and complex questions like this one. ["The ancient 'Lion of Nimrud' went missing from this city's national museum in 2003 (along with a lot of other stuff)"] This is also why we are now able to see the first self-driving cars. If you want to be able to tell the difference between, say, a tree and a pedestrian, well, that's pretty important. We don't know how to write those programs by hand, but with machine learning, this is now possible. And in fact, this car has driven over a million miles without any accidents on regular roads.

2:51 So we now know that computers can learn, and computers can learn to do things that we actually sometimes don't know how to do ourselves, or maybe can do them better than us. One of the most amazing examples I've seen of machine learning happened on a project that I ran at Kaggle where a team run by a guy called Geoffrey Hinton from the University of Toronto won a competition for automatic drug discovery. Now, what was extraordinary here is not just that they beat all of the algorithms developed by Merck or the international academic community, but nobody on the team had any background in chemistry or biology or life sciences, and they did it in two weeks. How did they do this? They used an extraordinary algorithm called deep learning. So important was this that in fact the success was covered in The New York Times in a front page article a few weeks later. This is Geoffrey Hinton here on the left-hand side. Deep learning is an algorithm inspired by how the human brain works, and as a result it's an algorithm which has no theoretical limitations on what it can do. The more data you give it and the more computation time you give it, the better it gets.

3:59 The New York Times also showed in this article another extraordinary result of deep learning which I'm going to show you now. It shows that computers can listen and understand.

4:11(Video) Richard Rashid: Now, the last step that I want to be able to take in this process is to actually speak to you in Chinese. Now the key thing there is, we've been able to take a large amount of information from many Chinese speakers and produce a text-to-speech system that takes Chinese text and converts it into Chinese language, and then we've taken an hour or so of my own voice and we've used that to modulate the standard text-to-speech system so that it would sound like me. Again, the result isn't perfect. There are in fact quite a few errors. (In Chinese) (Applause) There's much work to be done in this area. (In Chinese) (Applause)

5:12 Jeremy Howard: Well, that was at a machine learning conference in China. It's not often, actually, at academic conferences that you do hear spontaneous applause, although of course sometimes at TEDx conferences, feel free. Everything you saw there was happening with deep learning. (Applause) Thank you. The transcription in English was deep learning. The translation to Chinese and the text in the top right, deep learning, and the construction of the voice was deep learning as well.

5:37 So deep learning is this extraordinary thing. It's a single algorithm that can seem to do almost anything, and I discovered that a year earlier, it had also learned to see. In this obscure competition from Germany Called the German Traffic Sign Recognition Benchmark, deep learning had learned to recognize traffic signs like this one. Not only could it recognize the traffic signs better than any other algorithm, the leaderboard actually showed it was better than people, about twice as good as people. So by 2011, we had the first example of computers that can see better than people. Since that time, a lot has happened.In 2012, Google announced that they had a deep learning algorithm watch YouTube videos and crunched the data on 16,000 computers for a month, and the computer independently learned about concepts such as people and cats just by watching the videos. This is much like the way that humans learn.Humans don't learn by being told what they see, but by learning for themselves what these things are.Also in 2012, Geoffrey Hinton, who we saw earlier, won the very popular ImageNet competition, looking to try to figure out from one and a half million images what they're pictures of. As of 2014, we're now down to a six percent error rate in image recognition. This is better than people, again.

6:52 So machines really are doing an extraordinarily good job of this, and it is now being used in industry. For example, Google announced last year that they had mapped every single location in France in two hours,and the way they did it was that they fed street view images into a deep learning algorithm to recognize and read street numbers. Imagine how long it would have taken before: dozens of people, many years. This is also happening in China. Baidu is kind of the Chinese Google, I guess, and what you see here in the top left is an example of a picture that I uploaded to Baidu's deep learning system, and underneath you can see that the system has understood what that picture is and found similar images. The similar images actually have similar backgrounds, similar directions of the faces, even some with their tongue out. This is not clearly looking at the text of a web page. All I uploaded was an image. So we now have computers which really understand what they see and can therefore search databases of hundreds of millions of images in real time.

7:57 So what does it mean now that computers can see? Well, it's not just that computers can see. In fact, deep learning has done more than that. Complex, nuanced sentences like this one are now understandable with deep learning algorithms. As you can see here, this Stanford-based system showing the red dot at the top has figured out that this sentence is expressing negative sentiment. Deep learning now in fact is near human performance at understanding what sentences are about and what it is saying about those things. Also, deep learning has been used to read Chinese, again at about native Chinese speaker level. This algorithm developed out of Switzerland by people, none of whom speak or understand any Chinese. As I say, using deep learning is about the best system in the world for this, even compared to native human understanding.

8:47 This is a system that we put together at my company which shows putting all this stuff together. These are pictures which have no text attached, and as I'm typing in here sentences, in real time it's understanding these pictures and figuring out what they're about and finding pictures that are similar to the text that I'm writing. So you can see, it's actually understanding my sentences and actually understanding these pictures. I know that you've seen something like this on Google, where you can type in things and it will show you pictures, but actually what it's doing is it's searching the web page for the text. This is very different from actually understanding the images. This is something that computers have only been able to do for the first time in the last few months.

9:28 So we can see now that computers cannot only see but they can also read, and, of course, we've shown that they can understand what they hear. Perhaps not surprising now that I'm going to tell you they can write. Here is some text that I generated using a deep learning algorithm yesterday. And here is some text that an algorithm out of Stanford generated. Each of these sentences was generated by a deep learning algorithm to describe each of those pictures. This algorithm before has never seen a man in a black shirt playing a guitar. It's seen a man before, it's seen black before, it's seen a guitar before, but it has independently generated this novel description of this picture. We're still not quite at human performance here, but we're close. In tests, humans prefer the computer-generated caption one out of four times. Now this system is now only two weeks old, so probably within the next year, the computer algorithm will be well past human performance at the rate things are going. So computers can also write.

10:27 So we put all this together and it leads to very exciting opportunities. For example, in medicine, a team in Boston announced that they had discovered dozens of new clinically relevant features of tumors which help doctors make a prognosis of a cancer. Very similarly, in Stanford, a group there announced that, looking at tissues under magnification, they've developed a machine learning-based system which in fact is better than human pathologists at predicting survival rates for cancer sufferers. In both of these cases, not only were the predictions more accurate, but they generated new insightful science. In the radiology case, they were new clinical indicators that humans can understand. In this pathology case, the computer system actually discovered that the cells around the cancer are as important as the cancer cells themselves in making a diagnosis. This is the opposite of what pathologists had been taught for decades.In each of those two cases, they were systems developed by a combination of medical experts and machine learning experts, but as of last year, we're now beyond that too. This is an example of identifying cancerous areas of human tissue under a microscope. The system being shown here can identify those areas more accurately, or about as accurately, as human pathologists, but was built entirely with deep learning using no medical expertise by people who have no background in the field.Similarly, here, this neuron segmentation. We can now segment neurons about as accurately as humans can, but this system was developed with deep learning using people with no previous background in medicine.

12:07 So myself, as somebody with no previous background in medicine, I seem to be entirely well qualified to start a new medical company, which I did. I was kind of terrified of doing it, but the theory seemed to suggest that it ought to be possible to do very useful medicine using just these data analytic techniques.And thankfully, the feedback has been fantastic, not just from the media but from the medical community, who have been very supportive. The theory is that we can take the middle part of the medical process and turn that into data analysis as much as possible, leaving doctors to do what they're best at. I want to give you an example. It now takes us about 15 minutes to generate a new medical diagnostic test and I'll show you that in real time now, but I've compressed it down to three minutes by cutting some pieces out. Rather than showing you creating a medical diagnostic test, I'm going to show you a diagnostic test of car images, because that's something we can all understand.

13:05 So here we're starting with about 1.5 million car images, and I want to create something that can split them into the angle of the photo that's being taken. So these images are entirely unlabeled, so I have to start from scratch. With our deep learning algorithm, it can automatically identify areas of structure in these images. So the nice thing is that the human and the computer can now work together. So the human, as you can see here, is telling the computer about areas of interest which it wants the computer then to try and use to improve its algorithm. Now, these deep learning systems actually are in 16,000-dimensional space, so you can see here the computer rotating this through

that space, trying to find new areas of structure. And when it does so successfully, the human who is driving it can then point out the areas that are interesting. So here, the computer has successfully found areas, for example, angles. So as we go through this process, we're gradually telling the computer more and more about the kinds of structures we're looking for. You can imagine in a diagnostic test this would be a pathologist identifying areas of pathosis, for example, or a radiologist indicating potentially troublesome nodules. And sometimes it can be difficult for the algorithm. In this case, it got kind of confused. The fronts and the backs of the cars are all mixed up. So here we have to be a bit more careful, manually selecting these fronts as opposed to the backs, then telling the computer that this is a type of group that we're interested in.

14:32 So we do that for a while, we skip over a little bit, and then we train the machine learning algorithm based on these couple of hundred things, and we hope that it's gotten a lot better. You can see, it's now started to fade some of these pictures out, showing us that it already is recognizing how to understand some of these itself. We can then use this concept of similar images, and using similar images, you can now see,the computer at this point is able to entirely find just the fronts of cars. So at this point, the human can tell the computer, okay, yes, you've done a good job of that.

15:04 Sometimes, of course, even at this point it's still difficult to separate out groups. In this case, even after we let the computer try to rotate this for a while, we still find that the left sides and the right sides pictures are all mixed up together. So we can again give the computer some hints, and we say, okay, try and find a projection that separates out the left sides and the right sides as much as possible using this deep learning algorithm. And giving it that hint — ah, okay, it's been successful. It's managed to find a way of thinking about these objects that's separated out these together.

15:37 So you get the idea here. This is a case not where the human is being replaced by a computer, but where they're working together. What we're doing here is we're replacing something that used to take a team of five or six people about seven years and replacing it with something that takes 15 minutes for one person acting alone.

16:01 So this process takes about four or five iterations. You can see we now have 62 percent of our 1.5 million images classified correctly. And at this point, we can start to quite quickly grab whole big sections, check through them to make sure that there's no mistakes. Where there are mistakes, we can let the computer know about them. And using this kind of process for each of the different groups, we are now up to an 80 percent success rate in classifying the 1.5 million images. And at this point, it's just a case of finding the small number that aren't classified correctly, and trying to understand why. And using that approach, by 15 minutes we get to 97 percent classification rates.

16:42 So this kind of technique could allow us to fix a major problem, which is that there's a lack of medical expertise in the world. The World Economic Forum says that there's between a 10x and a 20x shortage of physicians in the developing world, and it would take about 300 years to train enough people to fix that problem. So imagine if we can help enhance their efficiency using these deep learning approaches?

17:07 So I'm very excited about the opportunities. I'm also concerned about the problems. The problem here is that every area in blue on this map is somewhere where services are over 80 percent of employment.What are services? These are services. These are also the exact things that computers have just learned how to do. So 80 percent of the world's employment in the developed world is stuff that computers have just learned how to do. What does that mean? Well, it'll be fine. They'll be replaced by other jobs. For example, there will be more jobs for data scientists. Well, not really. It doesn't take data scientists very long to build these things. For example, these four algorithms were all built by the same guy. So if you think, oh, it's all happened before, we've seen the results in the past of when new things come along and they get replaced by new jobs, what are these new jobs going to be? It's very hard for us to estimate this,because human performance grows at this gradual rate, but we now have a system, deep learning, that we know actually grows in capability exponentially. And we're here. So currently, we see the things around us

and we say, "Oh, computers are still pretty dumb." Right? But in five years' time, computers will be off this chart. So we need to be starting to think about this capability right now.

18:21 We have seen this once before, of course. In the Industrial Revolution, we saw a step change in capability thanks to engines. The thing is, though, that after a while, things flattened out. There was social disruption, but once engines were used to generate power in all the situations, things really settled down.The Machine Learning Revolution is going to be very different from the Industrial Revolution, because the Machine Learning Revolution, it never settles down. The better computers get at intellectual activities, the more they can build better computers to be better at intellectual capabilities, so this is going to be a kind of change that the world has actually never experienced before, so your previous understanding of what's possible is different.

19:01 This is already impacting us. In the last 25 years, as capital productivity has increased, labor productivity has been flat, in fact even a little bit down.

19:12 So I want us to start having this discussion now. I know that when I often tell people about this situation,people can be quite dismissive. Well, computers can't really think, they don't emote, they don't understand poetry, we don't really understand how they work. So what? Computers right now can do the things that humans spend most of their time being paid to do, so now's the time to start thinking about how we're going to adjust our social structures and economic structures to be aware of this new reality.Thank you. (Applause)" (Howard, 2016)

8.1.4 Transcript Sam Harris, 2016, Can we build AI without losing control over it?

"0:12 I'm going to talk about a failure of intuition that many of us suffer from. It's really a failure to detect a certain kind of danger. I'm going to describe a scenario that I think is both terrifying and likely to occur, and that's not a good combination, as it turns out. And yet rather than be scared, most of you will feel that what I'm talking about is kind of cool.

0:36 I'm going to describe how the gains we make in artificial intelligence could ultimately destroy us. And in fact, I think it's very difficult to see how they won't destroy us or inspire us to destroy ourselves. And yet if you're anything like me, you'll find that it's fun to think about these things. And that response is part of the problem. OK? That response should worry you. And if I were to convince you in this talk that we were likely to suffer a global famine, either because of climate change or some other catastrophe, and that your grandchildren, or their grandchildren, are very likely to live like this, you wouldn't think, "Interesting. I like this TED Talk."

1:20 Famine isn't fun. Death by science fiction, on the other hand, is fun, and one of the things that worries me most about the development of AI at this point is that we seem unable to marshal an appropriate emotional response to the dangers that lie ahead. I am unable to marshal this response, and I'm giving this talk.

1:41 It's as though we stand before two doors. Behind door number one, we stop making progress in building intelligent machines. Our computer hardware and software just stops getting better for some reason. Now take a moment to consider why this might happen. I mean, given how valuable intelligence and automation are, we will continue to improve our technology if we are at all able to. What could stop us from doing this? A full-scale nuclear war? A global pandemic? An asteroid impact? Justin Bieber becoming president of the United States?

2:19 (Laughter)

2:23 The point is, something would have to destroy civilization as we know it. You have to imagine how bad it would have to be to prevent us from making improvements in our technology permanently, generation after generation. Almost by definition, this is the worst thing that's ever happened in human history.

2:43 So the only alternative, and this is what lies behind door number two, is that we continue to improve our intelligent machines year after year after year. At a certain point, we will build machines that are smarter than we are, and once we have machines that are smarter than we are, they will begin to improve themselves. And then we risk what the mathematician IJ Good called an "intelligence explosion," that the process could get away from us.

3:09 Now, this is often caricatured, as I have here, as a fear that armies of malicious robots will attack us. But that isn't the most likely scenario. It's not that our machines will become spontaneously malevolent. The concern is really that we will build machines that are so much more competent than we are that the slightest divergence between their goals and our own could destroy us.

3:34 Just think about how we relate to ants. We don't hate them. We don't go out of our way to harm them. In fact, sometimes we take pains not to harm them. We step over them on the sidewalk. But whenever their presence seriously conflicts with one of our goals, let's say when constructing a building like this one, we annihilate them without a qualm. The concern is that we will one day build machines that, whether they're conscious or not, could treat us with similar disregard.

4:04 Now, I suspect this seems far-fetched to many of you. I bet there are those of you who doubt that superintelligent AI is possible, much less inevitable. But then you must find something wrong with one of the following assumptions. And there are only three of them.

4:22 Intelligence is a matter of information processing in physical systems. Actually, this is a little bit more than an assumption. We have already built narrow intelligence into our machines, and many of these machines perform at a level of superhuman intelligence already. And we know that mere matter can give rise to what is called "general intelligence," an ability to think flexibly across multiple domains, because our brains have managed it. Right? I mean, there's just atoms in here, and as long as we continue to build systems of atoms that display more and more intelligent behavior, we will eventually, unless we are interrupted, we will eventually build general intelligence into our machines.

5:10 It's crucial to realize that the rate of progress doesn't matter, because any progress is enough to get us into the end zone. We don't need Moore's law to continue. We don't need exponential progress. We just need to keep going.

5:24 The second assumption is that we will keep going. We will continue to improve our intelligent machines. And given the value of intelligence — I mean, intelligence is either the source of everything we value or we need it to safeguard everything we value. It is our most valuable resource. So we want to do this. We have problems that we desperately need to solve. We want to cure diseases like Alzheimer's and cancer.We want to understand economic systems. We want to improve our climate science. So we will do this, if we can. The train is already out of the station, and there's no brake to pull.

6:04 Finally, we don't stand on a peak of intelligence, or anywhere near it, likely. And this really is the crucial insight. This is what makes our situation so precarious, and this is what makes our intuitions about risk so unreliable.

6:22 Now, just consider the smartest person who has ever lived. On almost everyone's shortlist here is John von Neumann. I mean, the impression that von Neumann made on the people around him, and this included the greatest mathematicians and physicists of his time, is fairly well-documented. If only half the stories about him are

half true, there's no question he's one of the smartest people who has ever lived.So consider the spectrum of intelligence. Here we have John von Neumann. And then we have you and me. And then we have a chicken.

6:56 (Laughter)

6:58 Sorry, a chicken.

6:59 (Laughter)

7:00 There's no reason for me to make this talk more depressing than it needs to be.

7:04(Laughter)

7:07 It seems overwhelmingly likely, however, that the spectrum of intelligence extends much further than we currently conceive, and if we build machines that are more intelligent than we are, they will very likely explore this spectrum in ways that we can't imagine, and exceed us in ways that we can't imagine.

7:26 And it's important to recognize that this is true by virtue of speed alone. Right? So imagine if we just built a superintelligent AI that was no smarter than your average team of researchers at Stanford or MIT. Well, electronic circuits function about a million times faster than biochemical ones, so this machine should think about a million times faster than the minds that built it. So you set it running for a week, and it will perform 20,000 years of human-level intellectual work, week after week after week. How could we even understand, much less constrain, a mind making this sort of progress?

8:07 The other thing that's worrying, frankly, is that, imagine the best case scenario. So imagine we hit upon a design of superintelligent AI that has no safety concerns. We have the perfect design the first time around. It's as though we've been handed an oracle that behaves exactly as intended. Well, this machine would be the perfect labor-saving device. It can design the machine that can build the machine that can do any physical work, powered by sunlight, more or less for the cost of raw materials. So we're talking about the end of human drudgery. We're also talking about the end of most intellectual work.

8:48 So what would apes like ourselves do in this circumstance? Well, we'd be free to play Frisbee and give each other massages. Add some LSD and some questionable wardrobe choices, and the whole world could be like Burning Man.

9:01 (Laughter)

9:05 Now, that might sound pretty good, but ask yourself what would happen under our current economic and political order? It seems likely that we would witness a level of wealth inequality and unemployment that we have never seen before. Absent a willingness to immediately put this new wealth to the service of all humanity, a few trillionaires could grace the covers of our business magazines while the rest of the world would be free to starve.

9:33 And what would the Russians or the Chinese do if they heard that some company in Silicon Valley was about to deploy a superintelligent AI? This machine would be capable of waging war, whether terrestrial or cyber, with unprecedented power. This is a winner-take-all scenario. To be six months ahead of the competition here is to be 500,000 years ahead, at a minimum. So it seems that even mere rumors of this kind of breakthrough could cause our species to go berserk.

10:05 Now, one of the most frightening things, in my view, at this moment, are the kinds of things that AI researchers say when they want to be reassuring. And the most common reason we're told not to worry is time. This is all a long way off, don't you know. This is probably 50 or 100 years away. One researcher has said, "Worrying about AI safety is like worrying about overpopulation on Mars." This is the Silicon Valley version of "don't worry your pretty little head about it."

10:37 (Laughter)

10:38 No one seems to notice that referencing the time horizon is a total non sequitur. If intelligence is just a matter of information processing, and we continue to improve our machines, we will produce some form of superintelligence. And we have no idea how long it will take us to create the conditions to do that safely. Let me say that again. We have no idea how long it will take us to create the conditions to do that safely.

11:11 And if you haven't noticed, 50 years is not what it used to be. This is 50 years in months. This is how long we've had the iPhone. This is how long "The Simpsons" has been on television. Fifty years is not that much time to meet one of the greatest challenges our species will ever face. Once again, we seem to be failing to have an appropriate emotional response to what we have every reason to believe is coming.

11:37 The computer scientist Stuart Russell has a nice analogy here. He said, imagine that we received a message from an alien civilization, which read: "People of Earth, we will arrive on your planet in 50 years.Get ready." And now we're just counting down the months until the mothership lands? We would feel a little more urgency than we do.

12:03 Another reason we're told not to worry is that these machines can't help but share our values because they will be literally extensions of ourselves. They'll be grafted onto our brains, and we'll essentially become their limbic systems. Now take a moment to consider that the safest and only prudent path forward, recommended, is to implant this technology directly into our brains. Now, this may in fact be the safest and only prudent path forward, but usually one's safety concerns about a technology have to be pretty much worked out before you stick it inside your head.

12:35 (Laughter)

12:37 The deeper problem is that building superintelligent AI on its own seems likely to be easier than building superintelligent AI and having the completed neuroscience that allows us to seamlessly integrate our minds with it. And given that the companies and governments doing this work are likely to perceive themselves as being in a race against all others, given that to win this race is to win the world, provided you don't destroy it in the next moment, then it seems likely that whatever is easier to do will get done first.

13:09 Now, unfortunately, I don't have a solution to this problem, apart from recommending that more of us think about it. I think we need something like a Manhattan Project on the topic of artificial intelligence. Not to build it, because I think we'll inevitably do that, but to understand how to avoid an arms race and to build it in a way that is aligned with our interests. When you're talking about super intelligent AI that can make changes to itself, it seems that we only have one chance to get the initial conditions right, and even then we will need to absorb the economic and political consequences of getting them right.

13:44 But the moment we admit that information processing is the source of intelligence, that some appropriate computational system is what the basis of intelligence is, and we admit that we will improve these systems continuously, and we admit that the horizon of cognition very likely far exceeds what we currently know, then we

have to admit that we are in the process of building some sort of god. Now would be a good time to make sure it's a god we can live with.

14:19 Thank you very much.

14:20 (Applause)" (Sam Harris, 2016)

8.4 Appendix B: Questionnaire

8.4.1 Contacted companies

Recruiting company - Careerwise (sverige)
VD ROLAND.GUSTAVSSON@wisegroup.se

Accounting assistant - https://dooer.com
support@dooer.com

Cooking assistant - http://www.iqchef.se/
Christopher.Heybroek@iqchef.com

Web design - https://thegrid.io/
Contact-webform

Internet Security - https://www.darktrace.com/
info@darktrace.com

Social Media Assistant - https://www.echobox.com/
Contact-webform

Journalist - https://automatedinsights.com/
Contact-webform

Copywriter: https://phrasee.co/
Contact-webform

Contract evaluation: https://www.legalrobot.com/
hello@legalrobot.com

https://www.kth.se/forskning/artiklar/artificiell-intelligens-pa-stark-frammarsch-1.626006
stefanc@csc.kth.se

Doctor assistance: http://www.enlitic.com/
partner@enlitic.com

8.4.2 Roland Gustavsson, Wise Group

SIDA 1: Human like issues

F1: Please write your name and company:

Wise Group

F2: Will a learning machine ever have motivational problems to produce desired results?

no

F3: Will a learning machine know its limitations when it proposes new methods for producing desired results?

no

F4: Will a learning machine at any time lack information on direction?

no

SIDA 2: Result and Action Control

F5: Will a learning machine need rewards (incentives) to continue with its task?

no

F6: Can a company control that a learning machine possess knowledge of: "what is the desired business result"?

yes

F7: How can a company secure that an IA is able to influence the desired results? (I.e. will there be an HR equivalent function that controls which AI's to deploy?)

don't know

F8: Must the work products of a learning machine be reviewed, or will the results always be predictable?

predictable

SIDA 3: Personnel and Cultural control

F9: How can the "ability and capability" of a learning machine be evaluated?

Control the result and compare to what kind of result you wanted. If the capability is not sufficient, then you will have to finetune or change the algorithm

F10: How can a company control the "Ways of behaving" and the "Values" of a learning machine?

probably by controlling and changing the algorithm

8.4.3 Dan Rubins, Legal Robot:

SIDA 1: Human like issues

F1: Please write your name and company:

Dan Rubins, Legal Robot

F2: Will a learning machine ever have motivational problems to produce desired results?

I think motivation is a human concept that is difficult to apply to machine intelligence, if not many of the lower animal species. At least in the present state, machines don't have motivation per se. I think of machine intelligence in the way that a human looks at a mosquito. Such a simple animal has been programmed by a series of evolutional changes to complete various life-sustaining tasks, like seeking food, reproducing, etc and has support functions for these tasks, like the ability to digest blood as a food source, the ability to fly to transport itself to an ideal location to carry out these tasks. Just the same, machines have been programmed (in this case, by somewhat intelligent humans) to complete certain support tasks like computer vision, or understanding the environment or context of language. However, an analogous concept to life-sustainment is currently absent from machine intelligence. Higher level intelligence tasks tend to rely on a composition of lower tasks, but the prompting to execute those tasks relies on a human, or a human that initiates a program to cause prompting signals. I can write a cron job to periodically reevaluate the environment every 10ms using computer vision and machine learning, perhaps I can even tie this into lidar sensors and then connect it to an engine and wheels. That contraption, while exceedingly complex does not have motivation as we think of the concept as humans. Such a "self-driving" car does not decide to initiate a drive along the coast, it merely responds to human input and then fills in the details from environmental information and human-created rules and algorithms. Until we have a basic sentient machine, which may or may not be possible, the question of machine motivation seems impossible to answer.

F3: Will a learning machine know its limitations when it proposes new methods for producing desired results?

Sort of - if it is programmed to do so. For example, at Legal Robot, we want to prevent our algorithms from providing actual "legal advice" so we add controlling algorithms to evaluate the context of the learning set and the distribution of certain variables to judge whether we are making a prediction on sufficient data. The machine is programmed to learn that some predictions are beyond the human-imposed system of limitations we created to control the presentation of certain predictions.

F4: Will a learning machine at any time lack information on direction?

Yes, absolutely. Strategic planning is an area of active research in artificial intelligence.

SIDA 2: Result and Action Control

F5: Will a learning machine need rewards (incentives) to continue with its task?

While there are some algorithms that use incentive-based training, like reinforcement learning, applying the concept generally to machine intelligence is likely incorrect. In the larger societal sense, machines do not need to be compensated as an employee would since they do not have moral agency.

F6: Can a company control that a learning machine possess knowledge of: "what is the desired business result"?

This question somewhat mischaracterizes machine intelligence. Knowledge is an inherently human concept. Machines can access memory, and even access higher level concepts (what we would relate to a cognitive frame) that are learned from experience or training data, perhaps even from transfer learning but machine intelligence, as exists today, is devoid of agency. The machine does not really run on its own, rather a process (even a perpetual process, or a scheduled process) is initiated by a human that commands it to do a task. Even machines that continuously observe and process the environment around us - think of a weather station - were put there in place by humans and execute according to human-written programs, rules, or trained according to human commands. So a machine can access the statistical representations that are a machine's version of a concept, but it does not inherently *know* something or possess knowledge.

That being said, a machine can certainly interpret a simple natural language query like "what is the desired business result?", parse it, and recall a response using learned algorithms and can do so with surprising accuracy. However, parsing, retrieval from a dataset, and scoring is hardly as high level of a concept as us mere humans would understand as "knowledge".

F7: How can a company secure that an IA is able to influence the desired results? (I.e. will there be an HR equivalent function that controls which AI's to deploy?)

As a human, I believe the future of our humanity depends on command and control functions run by humans (likely humans augmented with separate AIs) as AI becomes more advanced. Again, we come to the questions of agency and autonomy. Without Moral Agency, machine intelligence is truly just a set of algorithms that require human prompting or execution. These algorithms can run wild, and there is already evidence of algorithms impacting judicial sentencing, or reinforcing latent human biases like race and gender, but anthropomorphizing statistics does not further the moral and philosophical conversation. Even

as a humanist, I strongly believe that a dumb AI (or even a dumb single algorithm) in the hands of irresponsible humans is far more dangerous than the fears of artificial superintelligence so often written about by people like Nick Bostrom. It is for this reason, that I believe companies that start implementing sufficiently advanced algorithmic systems need to start forming Review Boards, like the Institutional Review Board concept that is now accepted in medical ethics. At my company, Legal Robot, we added an ethics review into our software release process. We cannot deploy new code without independent and uninvolved humans signing off on the concept. We do this mainly for near-term concerns like bias and transparency. Though, as our AI gets more advanced, we expect to deal with trickier ethical issues.

F8: Must the work products of a learning machine be reviewed, or will the results always be predictable?

At Legal Robot, we use both algorithms and human review (so, properly speaking, this is augmented human review) to review the algorithmic output of our predictions. While this is possible in our business, it may not be possible in all situations. There is a similar concept that some companies use to train some algorithms used by self-driving cars, though the review is more through a process like reinforcement learning based on data collected from real drivers. We do this to make sure the predictive quality of our algorithms is sufficient for the task, rather than blindly turning over a potentially faulty prediction to an unwitting user.

SIDA 3: Personnel and Cultural control

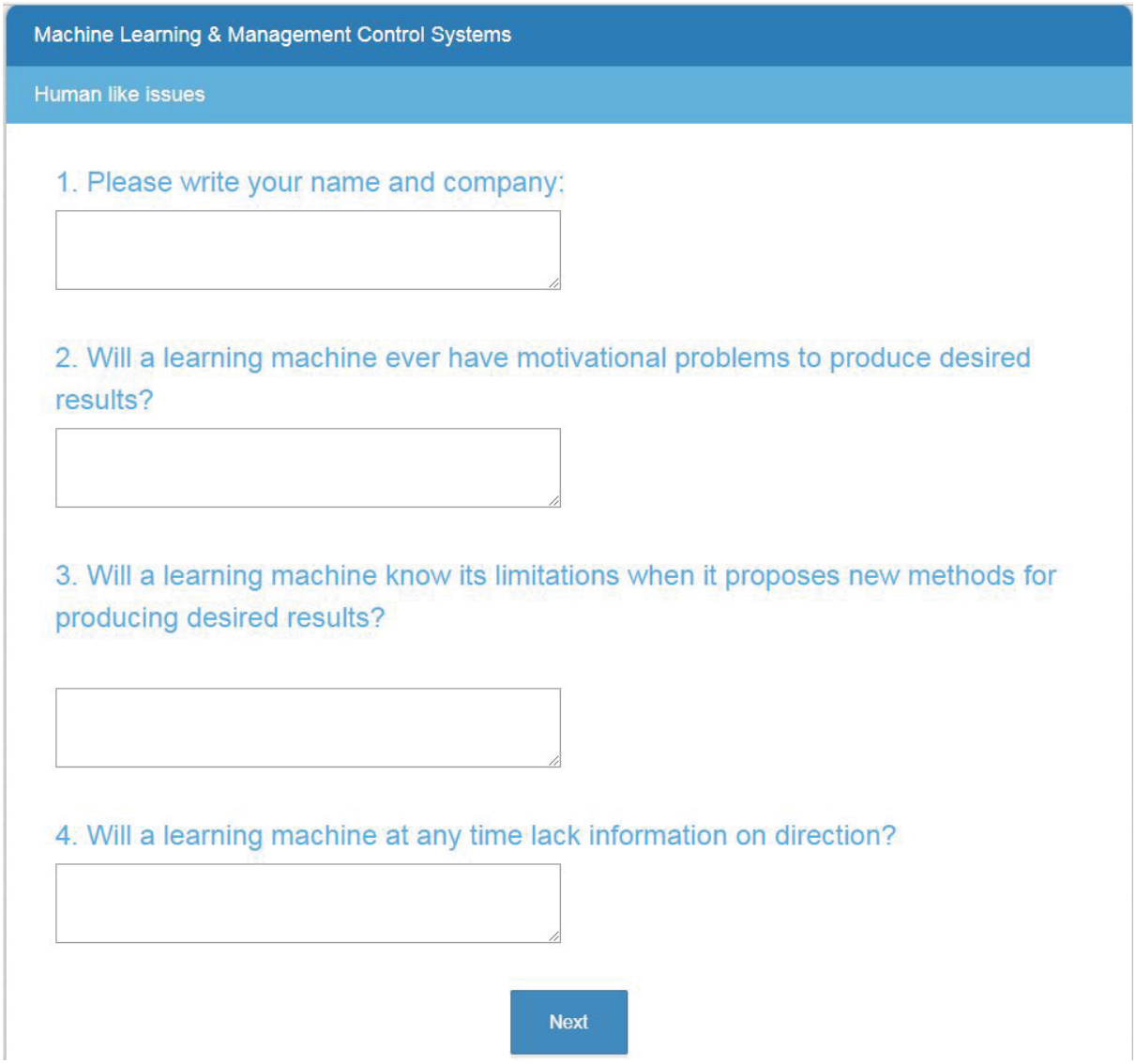
F9: How can the "ability and capability" of a learning machine be evaluated?

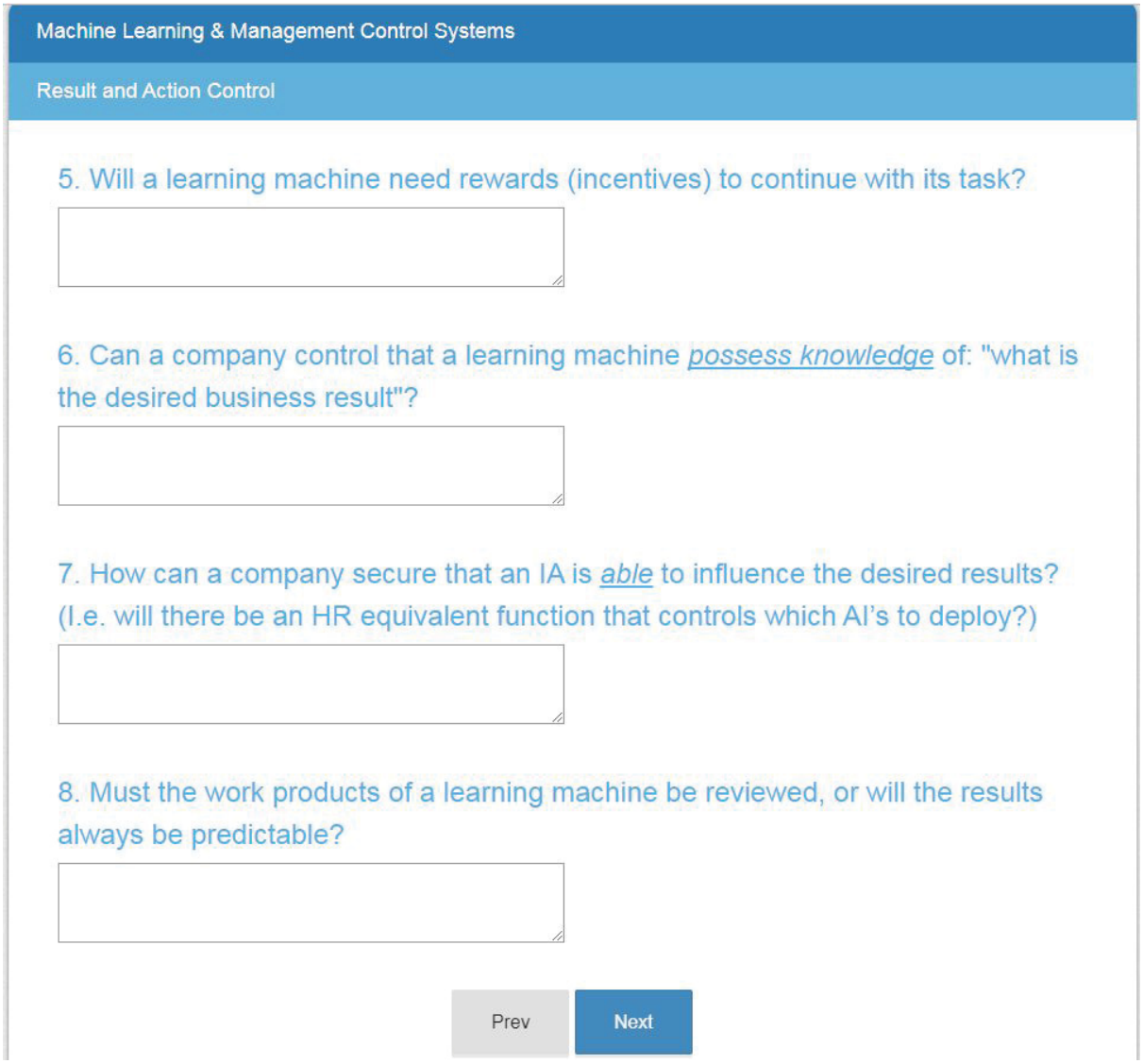
Ability and capability of AI can be evaluated through compositional analysis of every underlying algorithm. Are the F1 scores reasonable? are the p-values reasonable? is the training set representative of the real world? what contextual elements in the training set have an outsized influence on the predictive quality of the algorithm? At Legal Robot, questions like these are exactly what we ask our secondary review algorithms to process before a human makes a judgment call on the fitness of the algorithm.

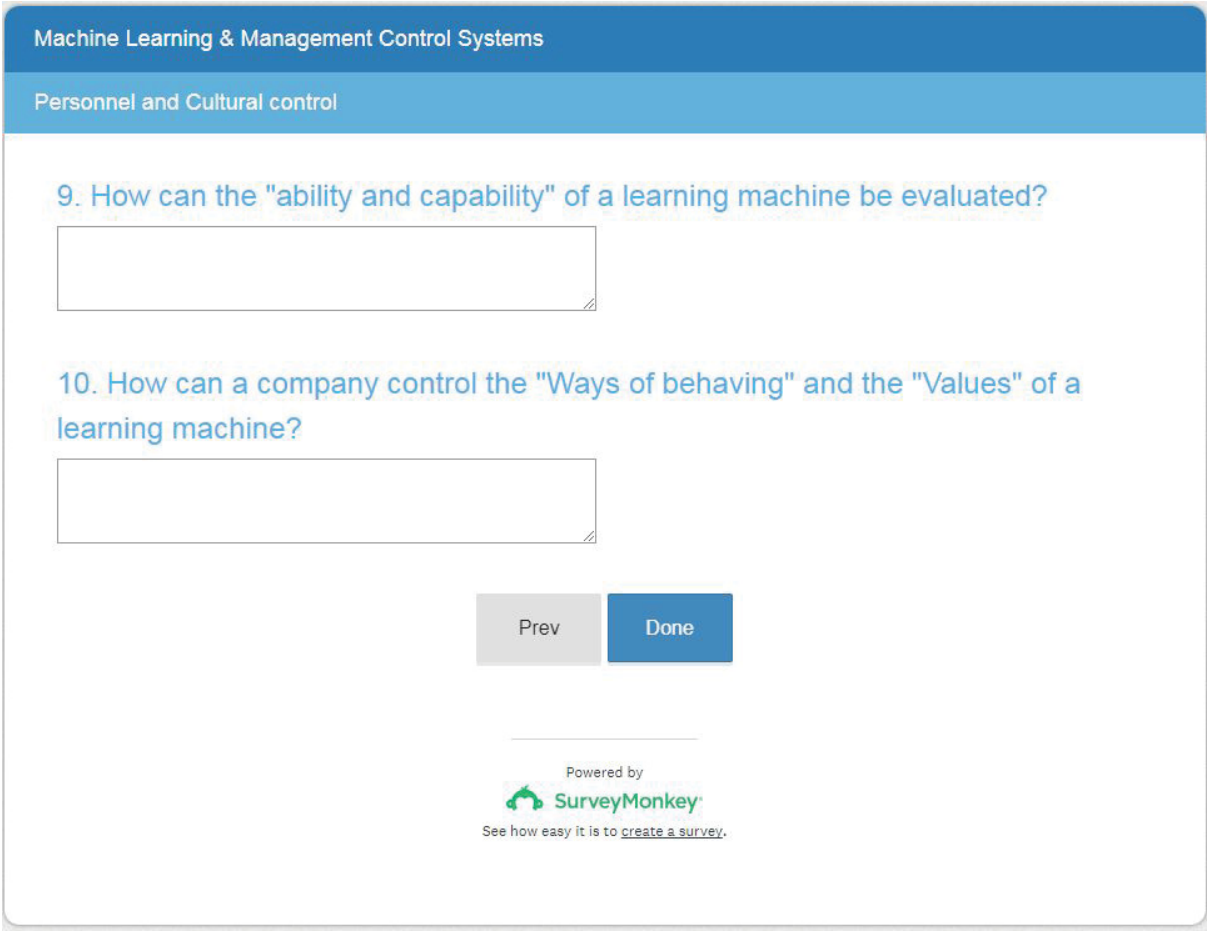
F10: How can a company control the "Ways of behaving" and the "Values" of a learning machine?

First, I would ask if a company should be controlling this, or if the responsibility should fall to independent experts that function outside of the corporate hierarchy. Similarly, I would also strongly argue against government-administered control since there are obvious political issues that could poison independence. Assuming there are independent human experts with sufficient power to administer control, mechanisms like independent code review, (or more conceptual reviews with non-technical folks) can help put humans in the loop before deployment. These are not enough, access and redress mechanisms need to be put in place by judicial systems. Companies need to start disclosing how they source their data, how they chose their algorithms, and actively search for adverse effects. For example, at Legal Robot, we made a public commitment to Algorithmic Transparency and started publishing a quarterly Transparency Report with details about these issues.

8.4.4 Used Monkey survey form




Machine Learning & Management Control Systems
Result and Action Control
5. Will a learning machine need rewards (incentives) to continue with its task?
6. Can a company control that a learning machine possess knowledge of: "what is the desired business result"?
7. How can a company secure that an IA is able to influence the desired results? (I.e. will there be an HR equivalent function that controls which AI's to deploy?)
8. Must the work products of a learning machine be reviewed, or will the results always be predictable?
Prev
Next



https://sv.surveymonkey.com/r/7BKPK3T