Joel Tolvanen

DEVELOPMENT OF TRUSTWORTHY CYBER-PHYSICAL SYSTEMS: ARTIFICIAL INTELLIGENCE'S VIEWPOINT

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

ABSTRACT

Tolvanen, Joel
Development of Trustworthy Cyber-Physical Systems: Artificial Intelligence's Viewpoint
Jyväskylä: University of Jyväskylä, 2020, 91 p.
Information Systems Science, Master's Thesis
Supervisor: Abrahamsson, Pekka

This Master's Thesis assesses how companies developing Artificial Intelligence are pursuing its trustworthiness. Artificial Intelligence is widely used when implementing Cyber-Physical Systems, that are coupling computational capabilities with the ability to control and sense the physical space. By gaining the access of physical environment, unexpected operation of Cyber-Physical Systems with AI capabilities can cause critical damage to environment and even human beings. Due to this, AI systems should be lawful, ethical and reliable. This research was conducted using the Ethics guidelines defined by European Commission, that provided the conceptual framework for assessment of trustworthiness of prevailing practices. Empirical qualitative research was conducted within Finnish companies developing Artificial Intelligence. Findings of the study suggest that trustworthiness is mainly pursued by realizing accountability, transparency and technical robustness of the system, whilst realization of societal and environmental wellbeing and diversity, nondiscrimination and fairness were not brought into attention. For realizing the neglected requirements, managerial advice is provided.

Keywords: Cyber-Physical Systems, Artificial Intelligence, ethics, trustworthy AI

TIIVISTELMÄ

Tolvanen, Joel
Luotettavien kyberfyysisten järjestelmien kehittäminen: Tekoälyn näkökulma
Jyväskylä: Jyväskylän yliopisto, 2020, 91 s.
Tietojärjestelmätiede, pro gradu -tutkielma (esimerkiksi Tietojärjestelmätiede, pro gradu
-tutkielma)
Ohjaaja: Abrahamsson, Pekka

Tämän pro gradu -tutkielman tavoitteena on selvittää millä keinoin yritykset pyrkivät kehittämään luottamusta herättäviä tekoälyjärjestelmiä. Tekoälyä käytetään laajasti toteutettaessa kyberfyysisiä järjestelmiä, jotka yhdistävät fyysisessä maailmassa tapahtuvat prosessit tietokoneiden tarjoamiin kykyihin. Saavuttaessaan kyvyn vaikuttaa fyysiseen ympäristöönsä, tekoälyllä varustetun kyberfyysisen järjestelmän odottamaton toiminta voi aiheuttaa kriittisiä vahinkoja sekä ympäristölleen että ihmisille. Tämän vuoksi tekoälyjärjestelmiä tulisi toimia eettisesti, lainmukaisesti ja vakaasti. Euroopan komission luotettavaa tekoälyä koskevat eettiset ohjeet muodostivat tämän tutkimuksen teoreettisen viitekehyksen, jota käytettiin vallitsevien käytänteiden luotettavuuden arviointiin. Tutkimus suoritettiin empiirisenä laadullisena haastatteluna, jonka osallistuvat koostuivat suomalaisista tekoälyä kehittävien yritysten työntekijöistä. Tutkimuksen tulokset esittävät, että luotettavuutta tavoitellaan ensisijaisesti vastuuvelvollisuuden, avoimuuden ja teknisen vakauden toteuttamisella. Sen sijaan yhteiskunnallisen ja ekologisen hyvinvoinnin sekä monimuotoisuuden, syrjimättömyyden ja oikeudenmukaisuuden tavoittelua ei nostettu esiin. Näiden laiminlyötyjen vaatimusten toteuttamiseksi esitetään liikkeenjohdollisia keinoja.

Asiasanat: kyberfyysiset järjestelmät, tekoäly, etiikka, luotettava AI

FIGURES

TABLES

TABLE OF CONTENTS

1 Introduction

Internet has had a remarkable impact on how people communicate and interact with each other. Further development of technology has made it possible to revolutionize interaction with physical world too. These systems, that communicate and interact with each other and their physical environment, are referred as Cyber-Physical Systems (CPS). (Rajkumar et al., 2010.)

Cyber-Physical System implements computation to physical processes. This is done by measuring real-world actions and surroundings with sensors and influencing them with actuators. CPSs are also capable of networking and communication. This differs them from traditional embedded systems that are so-called black boxes that do not provide outer access to the computational capabilities. (Alur, 2015; Lee, 2008; Rajkumar et al., 2010.)

Cyber-Physical Systems create value by improving operational efficiency, facilitating emerging business models and enabling service-centered manufacturing business. (Herterich et al., 2015.) Smart mobility is one example of a field that can benefit from the development of Cyber-Physical Systems. Mobility is a burning question in highly populated urban areas: it has both economic and environmental impact. It also involves the hard-to-predict human behavior. Smart Mobility is trying to solve these problems with the assistance of information technology. (Benevolo, Dameri, & D'Auria, 2016.)

New business models and services have transformed the development of software products. Costs of product development has declined from millions to thousands. Much of this is thanks to the open source communities and new service ecosystems. The change is also salient in hardware start-ups. There are no more need for building own factories as offshoring has become more accessible. (Blank, 2013.) This has made the development of different Cyber-Physical Systems available to larger audiences than before.

However, combining computation and physical processes is not easy task. For example, concurrent actions happening in physical world are unwieldy implemented in traditional computational abstractions, that have been dealing with user-provided inputs, outputs and only a little to none concurrency. (Lee, 2008.) Complexity of Cyber-Physical Systems has also increased significantly, when comparing them to the first embedded systems. For example, in the weapon systems of United States Airforce the amount of source code has multiplied by hundreds of thousands since 1960's and software has become a vital part of weapon systems with interconnected capabilities. (West & Blackburn, 2017.)

Cyber-Physical Systems are not easy to design. They might operate in physical environment that is inaccessible or they involve physical features that do not yet exist. Engineering has become increasingly computer-aided and different simulations are used in design stage. The integration of physical systems and their cyber counterparts is not easy tasks. Development of cyber counterpart may involve the development of Digital Twin, that is used for modeling various physical attributes forming complex relations. These cyber counterparts can aid the design of physical system. Water pumps are one example of such physical system: their physical implementation has many features that are subordinate to fluid dynamics and other laws of physics and they are very hard to alter after their physical implementation. (Ferguson, Bennett, & Ivashchenko, 2017.)

Cyber-Physical Systems have been up and coming for several years. Gartner Hype Cycle for Emerging Technologies features many technologies that can be considered cyber-physical. These include connected homes, autonomous vehicles, IoT platforms and smart robots. (Gartner, 2017.) The expectation is that soon we will be surrounded with various Cyber-Physical Systems, that are going to have significant impact on our surroundings.

Many of the capabilities of CPS can be achieved with the implementation of Artificial Intelligence (AI). Artificial Intelligence is set of techniques and methods used to implement machines with capabilities that have been previously insisted to require human intelligence. This means that instead of being single technical application, Artificial Intelligence covers various ways of implementing such intelligence. In addition to human-like capabilities, AI can process enormous sets of data, making it also superhuman. (Kaplan, 2016.)

Robotics is one subfield of research for Artificial Intelligence. This refers to development of intelligent agents that operate in physical space. The definition of this field is overlapping with the definition of Cyber-Physical Systems as they are too operating in physical environment with computational abilities. Artificial Intelligences applicability is not limited just to robotics. Other common applications of AI, for example computer vision, can too be implemented as part of a Cyber-Physical System. (Kaplan, 2016.)

Instead of welcoming Artificial Intelligence with open arms, its adoption has raised concerns among the people. People have been feeling anxious about losing their jobs to machines, risking their privacy or being violated by inhuman algorithmic decision. Some might be even afraid of becoming enslaved by evil AI seeking world dominance. (Kaplan & Haenlein, 2020.)

To prevent these fears from becoming reality, development of ethical, fair and cooperative Artificial Intelligence have been discussed by various scholars (Kaplan & Haenlein, 2020; von Krogh, 2018). European Commission (2019) has set guidelines for developing Artificial Intelligence that is lawful, ethical and robust. Following these guidelines is expected to realize AI systems that benefit the mankind and its individuals instead of realizing the dystopic scenarios presented in previous paragraph.

Goal of this study is to find out how trustworthy Cyber-Physical Systems with AI capabilities are being developed. Research question for this study is following:

> How do companies currently pursue trustworthiness when developing Artificial Intelligence?

Literature review was conducted using Google Scholar, because of its relevance and coverage. Literature was search with the keywords "cyber-physical systems" and "artificial intelligence". Other literature was also accepted from publications that were rated at least level 1 by the JUFO Publication Forum of Finnish scientific community. "Ethics guidelines for Trustworthy AI" by European Commission (2019) was chosen as theoretical framework for this study because of its relevance, as European Commission being the executive branch of European Union has considerable authority over companies adopting and developing Artificial Intelligence within the borders of EU. The empirical part of this study was conducted using qualitative methods as structured interview. The data from interviews was analyzed using thematic analysis employing integrated coding (Cruzes & Dybå, 2011). Analysis was performed with ATLAS.ti software.

Following chapters presents the literature review for this study, that discusses the concepts of Cyber-Physical Systems and Artificial Intelligence. After the literature review, the conceptual framework that forms the foundation for analysis is presented. Then, the chosen research method is discussed. This is followed by examination of the empirical findings and discussion, that connects the empirical findings to the theoretical background. In the final chapter, study is concluded with answer to the research question, discussion on the limitations of the study and proposition of future research opportunities.

2 Cyber-Physical Systems

This chapter introduces the concept of Cyber-Physical Systems, discusses the key features it has and examines the architectural solutions for CPS implementation.

2.1 Definition

Cyber-physical systems (CPS) integrate computational capabilities into physical processes. Cyber-physical systems can be monitored and controlled using the computational and communicative capabilities. (Lee, 2008; Rajkumar et al., 2010.) Based on these ground rules, the first functional component of a Cyber-Physical System is computational data management and analysis capabilities that form the cyber domain. To supplement these computational capabilities, the second functional component is connectivity that enables real-time data collection and transfer from physical world to the cyber domain. Together these two components form the core of Cyber-Physical Systems. (Lee, Bagheri, & Kao, 2015.)

Cyber-Physical Systems differ from "traditional" information systems. Information systems process a human provided input and provide an output that is also interpretable by human user. Cyber-Physical Systems interpret sensor data and control actuators independently, thus removing the human user. (Alur, 2015.) This removes the human error, but in other hand makes Cyber-Physical Systems more security critical as they have influence on their physical environment.

For efficient CPSs, the architecture behind systems should reflect their domain (i.e. their physical environment) accurately on low-level and yet be consistent on the meta level of the system (Rajkumar et al., 2010). Rajkumar et al. (2010) call this approach being "globally virtual, locally physical". It has been noticed that domain specific programming languages are efficient when building highly complex systems (Nakajima et al., 2002).

According to Yao et al. (2019) real-time data access, reconfigurable and interoperable capabilities, decentralized decision-making, intelligence and proactivity are characteristic for Cyber-Physical Systems. This differentiates them from traditional embedded systems that have computational capabilities but are lacking the ability to collect and utilize data and act independently and proactively. Koçak (2014) agrees with this by describing the development of embedded systems towards Cyber-Physical Systems by "simple and also single node, task oriented devices mutating into context-aware, multitasking and interactive devices in a network of nodes" (Koçak, 2014).

2.2 Features of Cyber-Physical Systems

Cyber-Physical Systems have key features that are data acquisition, control of physical environment, concurrent operations, time-awareness, security-criticalness and networking capabilities. These features and the corresponding literature that mentions them are depicted in TABLE 1.

TABLE 1 Features of Cyber-Physical Systems and the corresponding literature.

Feature	**Source**
Data acquisition	Alur (2015) Lee (2008) Rajkumar et al. (2010) Xu & Duan (2019) Yao et al. (2019)
Control of physical environment	Alur (2015) Rajkumar et al. (2010) Yao et al. (2019)
Concurrent operations	Alur (2015) Lee (2008) Rajkumar et al. (2010) Rettberg et al. (2017)
Time-awareness	Alur (2015) Ledwaba & Venter (2017) Lee (2008) Rajkumar et al. (2010)
Security-criticalness	Alur (2015) Ledwaba & Venter (2017) Rajkumar et al. (2010)
Networking capabilities	Lee (2008) Rajkumar et al. (2010)

This chapter examines the above-mentioned features in more detail.

2.2.1 Data acquisition

Sensors that measure their environment provide data for the feedback loop of Cyber-Physical Systems. This means that CPSs are able to collect data from their physical environment. (Alur, 2015.) With these sensors deployed in unprecedented locations to allow wide-spanned collection of data (Rajkumar et al., 2010).

Data collected by CPS can be used to control the physical environment. Alur, 2015.) With the insight the data provides, CPS can apply proper response to physical phenomena (Rajkumar et al., 2010). It is also possible to detect patterns on physical phenomena within the data. These patterns can then be utilized in data driven business. (Xu & Duan, 2019.)

Data acquisition capabilities also pose a great challenge. Cyber-Physical Systems are capable of collecting vast amounts of data. Thus, the quality of data should be ensured with reliable capturing, transferring and storing techniques. (Xu & Duan, 2019.)

2.2.2 Control of physical environment

Cyber-Physical Systems combine physical processes and computational capabilities in reactive manner: the computational device interacts with its environment independently without human interference. (Alur, 2015; Lee, 2008.) CPSs are expected to provide significant benefits by transforming the way humans control the physical world (Rajkumar et al., 2010). Yao et al. (2019) refer to the composition of data acquisition and control of physical environment as physical awareness.

Physical control sets many requirements for security, reliability, efficiency and architecture of Cyber-Physical Systems. Acting in physical space inflicts uncertainties on the system's operation. Something unexpected could happen in the physical environment of the CPS, affecting the system's operation from outside. This can also happen the other way around, resulting in unexpected behavior of Cyber-Physical System in its physical environment (e.g. errors in the physical device or inappropriate computational models). (Rajkumar et al., 2010.)

2.2.3 Concurrent operations

Cyber-Physical Systems are complicated and there are lot of automated actions happening concurrently. The high level of automation itself causes complexity in these systems. Furthermore, concurrent actions are contradictory with traditional computational abstractions, and causes confusion and obfuscation as they don't fit the physical world as we observe it. (Lee, 2008; Rettberg et al., 2017.)

Efficient Cyber-Physical Systems presumes that the computational abstractions match the physical properties (Rajkumar et al., 2010). CPSs work in physical environment where multiple processes are working simultaneously and communicating with each other. This means that the same task may be carried out by multiple and separate processors. (Alur, 2015.)

The problem of concurrency has been assessed with multiple models. These models have been both synchronous and asynchronous. In asynchronous models the actions are timed independently, and they are not fixed to a collective time domain. Synchronous models in other hand resemble more the traditional computing models. They feature a fixed logical sequence that is executed in synchronized rounds. (Alur, 2015.)

2.2.4 Time-awareness

In CPSs, real-time data from the physical sensors is processed in the cyber domain. This data is coming concurrently from different sources and actuations are also performed concurrently in reactive manner. (Ledwaba & Venter, 2017.) These concurrent operations and the control of physical environment means that Cyber-Physical Systems should be aware of the time of their physical environment. Control of the composition of cyber and physical resources must happen in real-time. (Rajkumar et al., 2010.)

Achieving reliability of real-time operations of CPSs has been a difficult task since networking technologies and computational abstractions have been more suitable to the cyber-domain and hardly matched the requirements of physical domain. Due to processor architectures, it is very hard to determine how long the execution for individual piece of code takes. The real-time function of a Cyber-Physical System should be predictable, which is challenging to achieve without temporal semantics and concurrent computing. (Lee, 2008.)

2.2.5 Security-criticalness

As a consequence of combining computation with physical processes, physical world is exposed to newborn security threats (Ledwaba & Venter, 2017). Ledwaba and Venter (2017) identified that, in a successful attack against a Cyber-Physical System, the system itself, its environment and human life may become exposed to damage. Risks should be assessed carefully to ensure stable and continuous operation of Cyber-Physical Systems. (Wu, Kang, & Li, 2015.)

Reliability is also a major concern when contemplating the security of Cyber-Physical Systems. The system should operate correctly in all situations. If failing to do so, unacceptable consequences, such as damage to the environment or humans, may emerge. (Alur, 2015)

2.2.6 Networking capabilities

Cyber-physical systems are networked, and their computational capabilities are accessible from outside, which differs them from traditional embedded systems that cannot be accessed from outer network and are programmed using low-level languages. Networking capabilities also pose a challenge when comparing CPS with embedded systems as exposing computational capabilities to the outside world makes them prone to altered behavior. (Lee, 2008.)

Networking is one of the core features of CPSs. One way to achieve networking for Cyber-Physical Systems is using the Internet. This concept of plugging sensors and actuators to internet is called the Internet of Things (IoT), and it can be considered as one instance of Cyber-Physical Systems. In IoT, different sensors and networks they form provide newly discovered information of their environment. With networking-capable sensors and actuators, Internet of Things is providing a bridge over the gap between physical world and information systems. (Haller, Karnouskos, & Schroth, 2009.)

2.3 Architectures for Cyber-Physical Systems

To implement the above-mentioned features, there have been several propositions for the architecture of Cyber-Physical Systems. CPS combines different subsystems into a bigger and far more complex system, arousing the need for architectural patterns.

> It is also expected that CPS is a collection or combination of a secured and efficient systems. Those systems enable individual entities to work together in order to form various complex systems with a new set of applications and capabilities. (Ahmed et al., 2016.)

Complexity of Cyber-Physical Systems makes their architecture especially critical since minor errors in system's behavior may escalate system-wide malfunction (Sha et al., 2008). Khaitan and McCalley (2015) point out that due to complex and multi-layered nature it is necessary to consider the design and architecture of CPSs carefully.

> A CPS is a "system of systems" where complex and heterogeneous systems interact in a continuous manner, and proper regulation of it necessitates careful codesign of the overall architecture of CPSs. (Khaitan & McCalley, 2015.)

Alam and El Saddik (2017) agree with this, by stating that "A CPS is composed of various other independent systems". It can be either composition of few or many subsystems. (Alam & El Saddik, 2017.)

2.3.1 5C architecture

5C architecture is a five-layered architecture for Cyber-Physical Systems introduced by Lee et al. (2015). These five layers are connection, conversion, cyber, cognition and configuration. Overview of the architecture can be seen in FIGURE 1.

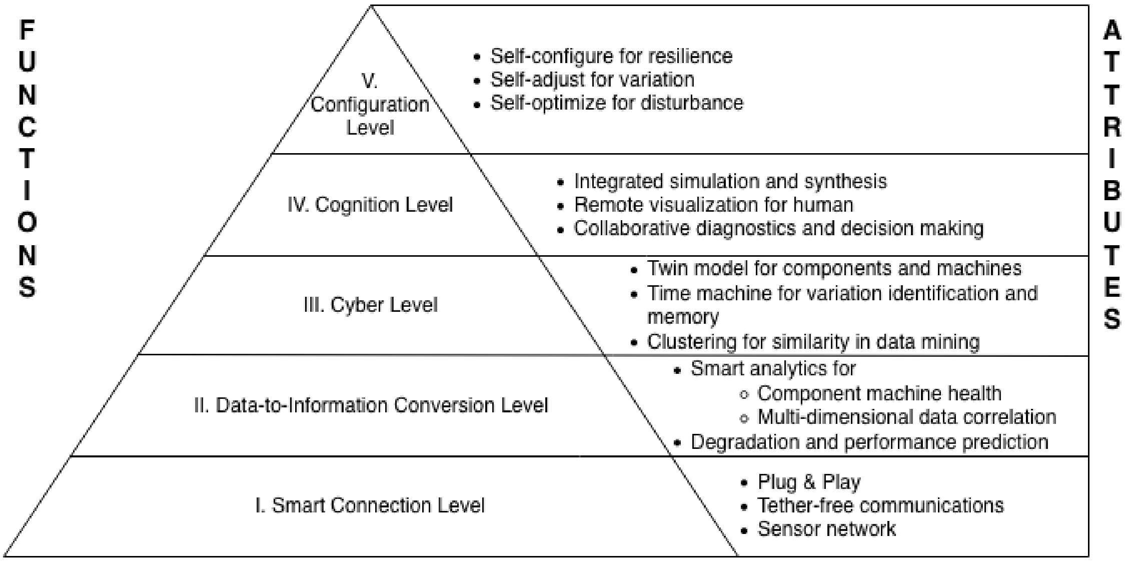

FUNCTIONS
ATTRIBUTES
V. Configuration Level
• Self-configure for resilience
• Self-adjust for variation
• Self-optimize for disturbance
IV. Cognition Level
• Integrated simulation and synthesis
• Remote visualization for human
• Collaborative diagnostics and decision making
III. Cyber Level
• Twin model for components and machines
• Time machine for variation identification and memory
• Clustering for similarity in data mining
II. Data-to-Information Conversion Level
• Smart analytics for
○ Component machine health
○ Multi-dimensional data correlation
• Degradation and performance prediction
I. Smart Connection Level
• Plug & Play
• Tether-free communications
• Sensor network

FIGURE 1 5C architecture for implementation of Cyber-Physical System (Lee et al., 2015).

Connection level handles the reliable and accurate data acquisition from the physical domain to the cyber space. Conversion level transforms the data coming from connection layer into meaningful information, thus making the machines self-aware. Cyber level's responsibilities include data management and analysis. It is the central information hub that combines the information coming from individual machines in the conversion layer. Cognition level generates holistic insight on the monitored system. It supports decision-making by providing both overall information of the Cyber-Physical System and status of individual machines. Configuration level makes the feedback from cyber space to physical space possible and facilitates corrective and preventive actions by providing configuration capabilities for the Cyber-Physical System. (Lee et al., 2015.)

2.3.2 Cloud-Based Cyber-Physical System Architecture

Alam, Sopena and El Saddik (2015) propose a Cloud-Based Cyber-Physical System Architecture. This architecture has its own layers for physical things, cyber things, peer-to-peer relation, intelligent services and system usage and administration. Alam and El Saddik (2017) refer to this architecture as the C2PS architecture. Outline for Cloud-Based Cyber-Physical System Architecture is presented in FIGURE 2.

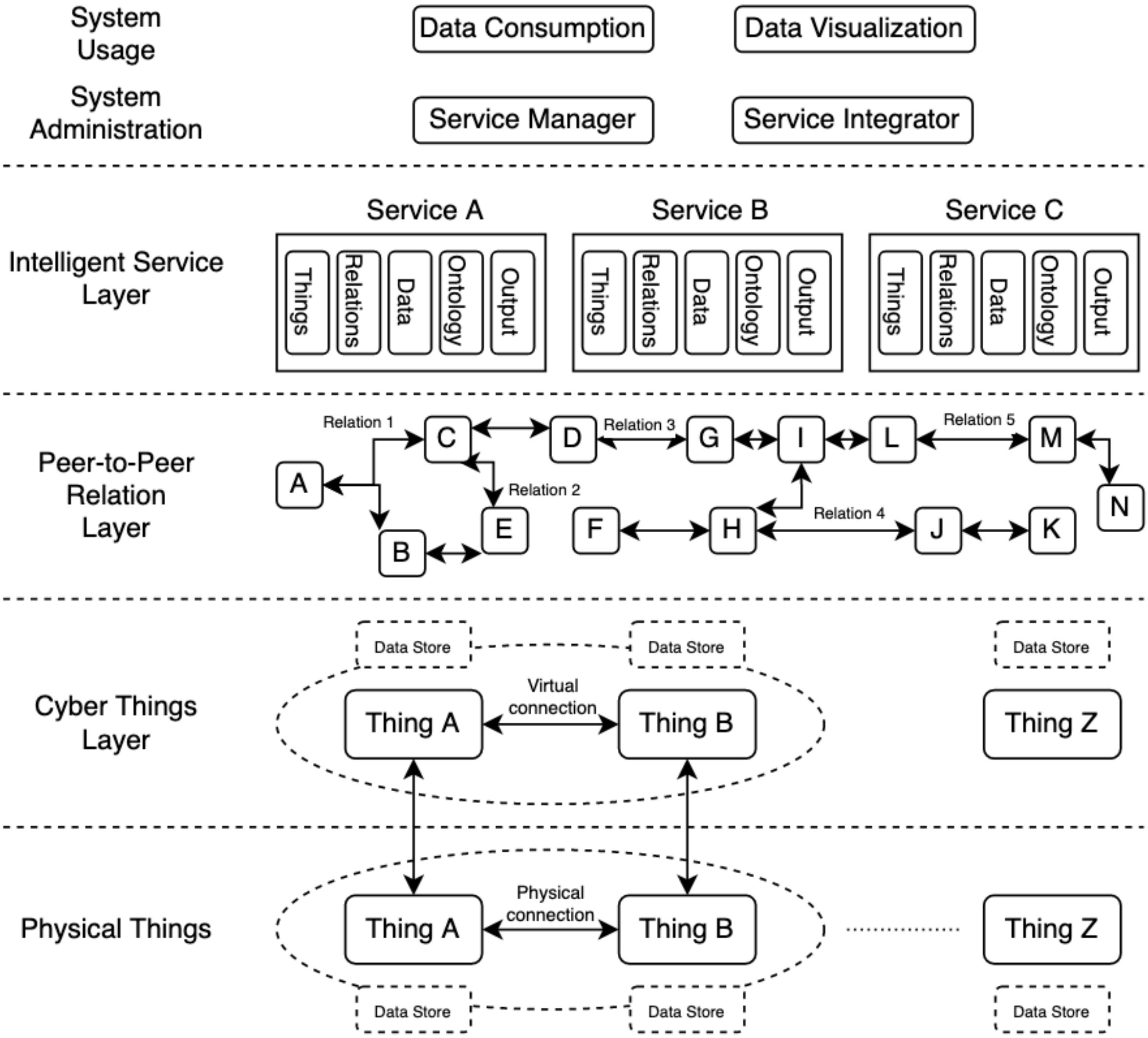


FIGURE 2 Cloud-Based Cyber-Physical System Architecture (Alam et al., 2015).

In C2PS architecture, each physical entity has its own dedicated cloud-based cyber counterpart, thus forming the layers of physical things and cyber things. Entities on both layers are aware about the existence of their counterpart and they both manage their own data store. Each entity that exists on these levels (cyber or physical) has an unique identifier. (Alam et al., 2015.)

Peer-to-Peer Relation Layer handles the relations formed by Cyber-Physical entities. The Cyber-Physical Things form many relations, creating communication groups. These groups have also their own unique identifiers. These communications are created and managed on the Peer-to-peer relation layer. (Alam et al., 2015.)

Intelligent Service Layer forms a middleware that couples all the relations formed on lower level layers with related ontologies. When domain specific ontologies are implemented on the data, intelligent solutions become reality. (Alam et al., 2015.)

In the system usage & administration layer, Service Manager manages the access control and privacy aspects of the Cyber-Physical things. Service

Integrator on the other hand makes it possible to combine multiple services together. (Alam et al., 2015.) Finally, the data is consumed: in the implementation of Alam et al. (2015), the data is consumed as reports and inputs to other systems.

2.3.3 Five-module architecture

Ahmed et al. (2016) outline a CPS architecture that consists of five modules. These five modules are actuators, sensing modules, data management modules, service-aware modules and application modules. The modules and related services of this architecture are depicted in FIGURE 3.

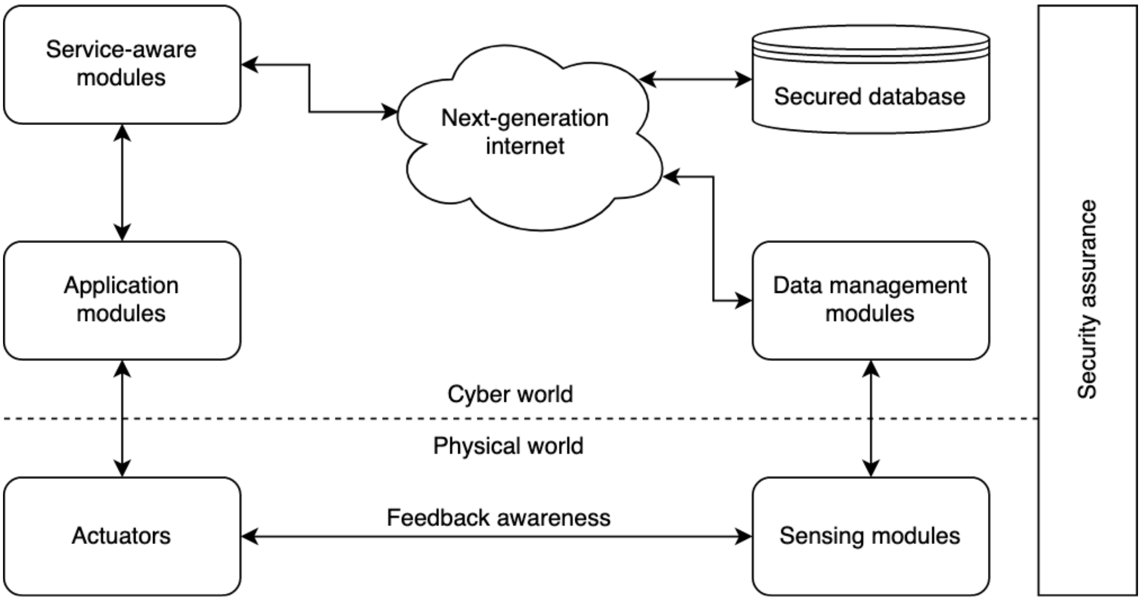


FIGURE 3 Five module CPS-architecture as presented by Ahmed et al. (2016)

To supplement the five modules, the architecture features Next-generation internet with advanced routing capabilities, secured database both locally and in the cloud and security assurance that is considered in implementation of every module in the architecture. (Ahmed et al., 2016.)

Actuators and sensing modules provide a beachhead at the edge of physical world and cyber world. Sensing modules consist of sensors that renders environmental awareness possible. Actuators in the other hand interact with physical environment by executing commands. (Ahmed et al., 2016.)

Environmental data collected by Sensing modules is then fed to the Data management modules. Data management module's duties include data processing (e.g. normalization, noise reduction) and storage. Data processed by Data management modules is then forwarded to Service-aware modules over the Next-generation internet. (Ahmed et al., 2016.)

Service-aware modules offers the overall functionality of the whole Cyber-Physical System. Its capabilities include support for decision making, task analysis and task scheduling. Service-aware modules recognize their incoming data and then forward it to the appropriate services. (Ahmed et al., 2016.)

Application modules deploy various services that interact with Next-generation internet. They also give commands to be executed by Actuators, thus having control over the physical world. (Ahmed et al., 2016.)

3 Artificial Intelligence

This chapter introduces the definition and characteristics of Artificial Intelligence. After these central concepts have been introduced, development of AI systems is discussed.

3.1 Definition

The concept of Artificial Intelligence (AI) is very broad and lacks unanimous definition. Instead of considering AI as a technological application, Von Krogh (2018) describes it as an ubiquitous phenomenon that has significant impact economically and organizationally. Artificial Intelligence encompasses various methods and applications for achieving technology with human-like abilities, the so-called "intelligence" in this context. This definition is still crippled by the difficulty of defining human intelligence in the first place. Another burden for definition of AI comes from the "AI effect": if a machine is able to perform chosen task, the task itself is no longer perceived as something that requires intelligence (Kaplan & Haenlein, 2020.)

In addition to vague overall definition of intelligence, defining Artificial Intelligence through the lens of human abilities does not cover all of its possibilities. For example, AI systems are able to process vast amounts of data and identify patterns in it. With this ability, AI systems can for example recognize cyber-attacks or approaching natural disasters, tasks that require both the computational ability to process data and the humane ability to draw conclusions. In these examples, Artificial Intelligence has capabilities that exceed their human counterparts, suggesting that both human and superhuman capabilities are characteristic for AI. (Kaplan, 2016.)

To overcome the vague relation between the abilities of Artificial Intelligence and human intelligence, it has been proposed that the way AI system approaches the problem is as important as solving the problem when assessing its intelligence. This means that instead of brute forcing all possible alternatives with its computational capabilities in order to find the best solution, AI reaches its outcomes by making context-aware generalizations on the matter. (Kaplan, 2016.) Kaplan (2016) describes this “essence of AI” as “the ability to make appropriate generalizations in a timely fashion based on limited data”. This kind of inductive ability to make generalizations on the grounds of limited observations is shared by both human intelligence and the so-called “superhuman” Artificial Intelligence. Haenlein and Kaplan (2019) reach a similar conclusion with slightly other words, characterizing an AI system with the ability to ”interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation”.

3.2 Adoption

Despite its recent emergence in wide public, AI is an old concept and the roots of the term “Artificial Intelligence” date back to the year 1955 when it was allegedly used for the first time by John McCarthy (Myers, 2011). In the beginning, AI was expected to be a field of study that would grant computers ability to learn, reason, solve problems and make decisions (von Krogh, 2018).

After the concept’s inauguration, the research on AI declined for the following decades due to lack of practical applications and ability to deliver the inflated expectations (von Krogh, 2018). Neural networks, which work in similar way as neurons of the human brain, are one example of early AI’s inability to deliver. Their concept was discussed as early as 1940’s, but research on Neural networks stalled in 1960’s when it was discovered that computational power of the time was not sufficient for their implementation. During the 1960’s, the field of AI witnessed major cuts on funding in United States and United Kingdom. (Haenlein & Kaplan, 2019.) Interest on AI was revived on 1980’s and 1990’s when companies started heavily investing on research and development of expert systems to support decision-making. (Haenlein & Kaplan, 2019; von Krogh, 2018.)

Von Krogh (2018) argues that the remarkably rapid adoption of AI in recent years has been driven by development of underlying technologies, advancement on data collection abilities, increased affordability of computational power and cloud-based services that have made AI available for wide range of operators.

Brock and von Wangenheim (2019) discuss the business impact of AI adoption. Their study implies that, on the contrary to the widespread discussions and so-called “hype”, implementation and adoption of Artificial Intelligence does not drastically differ from other technologies and is similar to other digital transformation projects. They also suggested seven success factors for adoption of AI. Their framework, called DIGITAL, comprises the perspectives of data, intelligence, groundedness, integrality, teaming ability, agility and leadership. This framework is presented in TABLE 2.

TABLE 2 Success factors for adoption of AI proposed by Brock and von Wangenheim (2019).

Perspective	Description
Data	Data in digital format forms fundamental basis for AI systems and its quality should be ensured. Organizations need proper capabilities for acquiring, managing and analyzing relevant data.
Intelligence	Technical and managerial skills and patience constitute the perspective of Intelligence. Strategic, technological and security-related capabilities are needed for achieving proper understanding on how AI can be utilized to benefit the company. Patience is also needed because insightful adoption and implementation of AI is often a non-deterministic process and requires experimentation.
Grounded	Adoption of AI should be grounded to the existing business of the adopting company in order to provide benefit for their core operations. This is achieved by starting AI projects that, instead of pursuing extravagant goals with inflated expectations, are focused on improving existing offerings, reducing costs or improving efficiency of the operations.
Integral	Integral and holistic approach is needed, and this approach should consider strategy, processes, data management, technology alignment, employee engagement and culture of the company. In this perspective, AI should be perceived in its broad definition as a tool for digital transformation that is spanning all units within the company.
Teaming	AI ambitions are not likely to realize if pursued alone. This means that companies should form efficient teams and partnerships in order to build powerful ecosystems. There are no off-the-shelf solutions for AI systems, which requires tight cooperation between companies and their technology partners.
Agile	Companies' ability to adapt to changes was identifying as the most important success factor for AI adoption. Successful adoption of AI systems also provides further improvement on organizations ability to adapt to change. Thus, agility is both prerequisite and outcome of successful Artificial Intelligence adoption.
Leadership	Leaders should support the AI ambitions of the company by active endorsing. Leadership is especially important for facilitating the transformation process and carrying it to the figurative finish line, ensuring the successful adoption.

Data is the foundation for AI systems and its high quality is fundamental for successful implementation of AI, they argue. Intelligence refers to required skills for utilizing AI, as they identified that building successful AI systems requires both technical talent and strategic vision of managers. Companies should also be Groundedness means that companies' AI aspirations should be bounded with their existing operations to avoid falling out-of-scope. Integral approach is also required for successful company-wide AI implementation. With this approach, AI is seen in its broad definition as a paradigm for achieving company-wide transformation efforts. Forming teams and partnerships should produce effective business ecosystems for utilizing AI. Lack of agility was identified as the second most important challenge for AI adoption, and ability to adapt to change was considered as prerequisite for successful exploration of Artificial Intelligence. Leaders should have active role in promotion of AI system and work as change agents to ensure successful adoption. (Brock & von Wangenheim, 2019.)

These findings of Brock and von Wangenheim (2019) suggest that adoption of AI bears great resemblance with other digitalization efforts that companies are doing. AI has partially same drivers as digital transform generally has, such as the improved ability to collect vast amounts of data.

3.3 Application areas

According to Kaplan (2016), robotics, computer vision, speech recognition and natural language processing are currently the most significant research areas for Artificial Intelligence. Each of these areas are characterized with tasks that have previously required human intervention.

Kaplan (2016) defines the field of robotics as development of "machines that are capable of performing physical tasks". These AI-equipped robots should be general purpose machines, that can execute various physical tasks autonomously with AI capabilities. This differs them from traditional mechanic automation, that too can operate in physical space but are built for very specific purposes. (Kaplan, 2016.) This definition of robotics with AI capabilities is considerably equivalent to various definition of Cyber-Physical Systems, which are also expected to combine physical processes with computational capabilities (Lee, 2008; Rajkumar et al., 2010). In other words, AI capabilities can be used to implement intelligent agents with computation and networking capabilities and control of their physical environment, thus forming an instance of Cyber-Physical Systems.

Computer vision is trying to give computers visual ability. In practice this means that computer vision algorithms are able to identify and label objects in images. Such identification tasks can be considered as the first successful frontier of computer vision, as the algorithms can be used for more complex tasks, as constructing three-dimensional models from multiple images. Computer vision can be used to analyze other data and not just visible light, thus differing from the human-centric concept of "vision". This ability can be used to process all kinds of two-dimensional data, meaning that computer vision can be used to analyze infrared and ultrasound, for example. Possible applications include finding underground geological formations, detecting tumors and assessing condition of concrete structures, just to mention some. (Kaplan, 2016.)

Speech recognition is trying to transform the human speech into written text. There are various challenges in this, such as detecting the correct signal from surrounding noise, recognizing correct punctuation and identifying correct written expression for given sounds. From technical perspective, computer vision and speech recognition form a very different problem. Computer vision detects objects using two-dimensional data from given moment of time while speech recognition is trying to label a "single variable (sound waves) that changes over time". (Kaplan, 2016.) Multiple commercial systems with speech recognition capabilities have been developed during 2010's. Due to this, speech recognition is considered as the first triumph of deep learning AI. (Deng, 2018.)

Natural language processing is trying to achieve the ability to automatically process languages used by human beings, which can be used for example to translate, summarize and produce such text. This marks the distinction between natural languages, that humans use when communicating with each other, and programming languages, which are just abstractions that provide a way of giving accurate instructions for machines. (Kaplan, 2016.) According to Deng (2018), machine translation, which refers to the NLP algorithms' capability of translating texts that are written in natural languages, is currently one of the major application areas of AI alongside with computer vision and speech recognition.

3.4 Impact

Adoption of Artificial Intelligence is expected to have impact that is stirring various concerns. Systems that act autonomously in reactive and non-deterministic manner could cause unprecedented questions. Responsibility and accountability considering AI system's actions still remain debatable. Adverse outcomes caused by intelligent and autonomous agents could result in situations where liability is disputed, thus resulting in legal vacuum. (Kaplan, 2016.) AI systems are expected to provide outputs that are easily perceived. However, how the chosen output was produced often remains concealed from the end user. To understand the underlying mechanisms, one should be familiar with the implemented algorithms. (von Krogh, 2018.)

Kaplan and Haenlein (2020) conducted an analysis on the ethical concerns. Using the PESTEL framework, which is a framework for strategic analysis of business environment, they assessed adoption of AI from the perspectives of politics, economics, society, technology, environment and law.

Political impact of AI goes beyond traditional warfare as it can be used for disturbing political processes. AI can be utilized for military purposes with use cases spanning beyond advanced and autonomous military robots and drones. These possibilities include utilization for hybrid warfare that brings cyber and political dimensions beside the traditional means of warfare. For example, algorithms are already selecting the content that is shown to users of social media. This has made these algorithms unexpectedly influential on political matters. Affecting the decisions made by these algorithms can make it possible to obstruct democratic processes. Still, AI can as well be utilized for better informed decision-making also in politics and for example help citizens find the political party that is closest to their values. (Kaplan & Haenlein, 2020.)

Discussions of AI's economic impact have been revolving around AI systems replacing human employees and increasing unemployment. It has been argued that AI can perform tasks more efficiently and cheaper than human employees. Still, AI systems are very specialized on the tasks that they can perform, meaning that even though being adaptable and having the ability to create generalizations, they are still able to overcome tasks that are tightly scoped. AI systems require significant investments before complete replacement of human labor is realistic. These remarks suggest that replacing human employees with Artificial Intelligence still lies far in the future and rather it is more likely that AI will improve humans' work quality via automating meticulous routine tasks. (Kaplan & Haenlein, 2020.) Von Krogh (2018) agrees with this notion by calling Artificial Intelligence a technology that in the near future AI applications are expectedly going to augment human labor instead of substituting it.

Societal impact of AI depends on its accessibility. Inequality and loneliness are considered to be the most common social problems among modern day societies. Capital required for AI adoption makes the benefits of AI inaccessible for people that are already in weaker position within the society, thus promoting inequality. However, AI can also reduce disparity by making services (e.g. healthcare) accessible for broader range of people. AI systems also have the ability to both promote and reduce loneliness within societies. If human employees are replaced with AI systems, people dependent on them could end up in isolation. On the other hand, AI systems that have been designed to act in human-like manner already exist and they can, for example, recognize human emotions making it possible to react to them. This could relieve people with limited social contact. (Kaplan & Haenlein, 2020.)

Technological point of view concerns the relation between technology and its creators. World dominance of computers is one of the most discussed dystopic scenarios relating to adoption of Artificial Intelligence. Retaining human agency when increasing number of tasks are put under control of AI is one of these concerns. There are several issues obstructing stability of human control over AI: human interpretation, biased training data and complexity of the system. AI systems might react unexpectedly to unclear demands and instructions that human users may present them, as they falsely assume the task to be self-evidently clear. However, AI might not act accordingly to these assumptions thus operating against human intention. AI systems could also be biased from their outset. Training them with historical data might result in AI systems that amplify historical biases and latently discriminate people. Rapid development of AI systems has also increased their complexity. This makes it harder for humans to understand their underlying mechanisms and explain how the system produced the outcome in question. Some of these systems are perceived to be black boxes making human authority over them questionable. (Kaplan & Haenlein, 2020.)

Environmentally the adoption of AI can also happen both for bad and good. Each technological advancement during human history has put cumulative burden on the environment. Adoption of Artificial Intelligence is no exception of this trend. Increased need for computational power increases energy consumption and manufacturing semi-conductor components used in computers requires various natural resources. This all is added on top of the existing strain on the environment. However, utilizing AI could also have a positive impact on environment. Adopting AI solutions is expected to lesser the environmental burden caused by human economy via improving energy efficiency and facilitating resource-wise operations. (Kaplan & Haenlein, 2020.)

Law and regulation are another major concern on AI. Adoption of AI involves various legal challenges. AI systems process vast amounts of data, and significant portion of this data is generated on individual human beings. This makes privacy one of the major concerns from legal point of view: respect for privacy is considered as a civil right but too tight regulation could lead to loss of investments to places with lesser regulation. Other legal concern is the liability of Artificial Intelligence's operations. AI system's supply chain involves various stakeholders starting from people who developed the underlying algorithms to its end users. This makes it especially important to define liability for AI system's operations if it causes harm to its environment (both cyber or physical) or human beings. (Kaplan & Haenlein, 2020.)

As regulations have a significant say in how AI is going to shape the society and environment, Haenlein and Kaplan (2019) examined the need for new regulation on three levels. First, they assess the need of regulation from the micro-perspective that assesses the AI systems' impact on individuals, then they move on to meso-perspective that discusses the impact on societal level. Finally, they discuss the macro-perspective of regulation that considers the impact of AI globally.

From the point of view of individual human (micro-perspective), AI will increasingly impact individuals as AI systems are being employed for decision-making. This could lead into situations where existing historical biases are being reinforced, thus leading to increased discrimination. Haenlein and Kaplan's (2019) suggestion for avoiding these micro-level adversaries is the regulation of development processes and organizations instead of regulating the AI itself. This calls for clear requirements for training and testing AI system's underlying algorithms and rules for ensuring the accountability of the companies that are developing such algorithms. On societal level, Haenlein and Kaplan (2019) expect that AI systems will replace human employees in expert work at least to some extent, thus resulting in significant transformation of the labor market. They suggest that new regulations should be adapted for mitigating the by-products of the shift in labor market. These regulations could include entailing companies to re-educate their employees with the money saved by AI systems and imposing limits on the use of automation. On macro level, Haenlein and Kaplan (2019) argue that regulation of AI systems should ensure the respect to democracy and peace. Several legislative bodies have taken various paths on the regulation of AI from this point of view, as the government of China has utilized AI systems for surveillance and, on the other end of this spectrum, the European Union has issued legislation that restricts the use of personal data, thus resulting in significant restrictions on possible use cases of Artificial Intelligence. These opposite legislative approaches could result in regional disparity on the use of AI, making international coordination of regulations necessary in the future.

3.5 Approaches for implementation

This chapter presents two common approaches for implementing Artificial Intelligence: machine learning and neural networks.

3.5.1 Machine learning

Machine learning (ML) is a set of methods for detecting patterns, developing predictions and making decisions based on probabilistic models. All of this is done under uncertainty, meaning that there are no explicitly correct answers to the problems that the algorithm is solving. Machine learning can be divided further into Supervised learning, Unsupervised learning and Reinforcement learning. (Murphy, 2012.)

Supervised learning (sometimes referred as predictive learning) is approach where AI is trained with labeled input-output pairs. Input can be a set of numerical data that represent one subject's features and attributes. Output, also known as the response variable, is often a categorical value but it can also be numeric. Algorithms with categorical and nominal output variable are used to solve classification problems as they are used for classifying the input data. If the output variable is numeric, the algorithm is solving a regression problem. The label pairs with input-output form the training set for the ML algorithm, that can then be used to classify other input data. (Murphy, 2012.)

Unsupervised learning discovers patterns from the data on its own. Instead of being trained with a training set consisting of labeled input-output pairs, Unsupervised learning algorithms try to discover conspicuous patters from the given input data. This is done in purely data-oriented manner without any *a priori* conception on which patterns the data might contain. Not being a well confined problem, assessing the success of unsupervised learning algorithm is difficult as given input has no explicit output pair that could be observed by other means. Thus, the comparison of unsupervised learning algorithms outcome and expected output is not possible. (Murphy, 2012.)

Reinforcement learning uses reward and punishment signals for achieving desired behavior of AI (Murphy, 2012). Practically this means that the outcome produced by Machine Learning algorithm is validated, thus giving a reward or punishment signal. One way to implement validation is to have human operators determining whether the outcome was correct (i.e. teaching the system), but other ways for validation are also possible. This way AI system utilizing a reinforcement learning algorithm learns by trial and error when it is acting in changing environment. Reinforced learning models should inherently explore their environment in order to learn and find their optimal way to solving the problem. The trade-off between exploring the environment (i.e. trying new actions) and exploiting (i.e. repeating an action) the environment for most optimal is one of the biggest concerns of reinforced learning. (Kaelbling, Littman, & Moore, 1996.) Existing reinforcement learning algorithms can for example learn to play computer games on their own, with input that only consists of the pixels of the game screen and their current score. These scenarios do not require human interference for training, as the algorithm gets explicit reward signal from the game as its score increases. (Mnih et al., 2015.)

3.5.2 Neural networks

Neural networks are trying to implement Artificial Intelligence by imitating the way nervous systems operate. They combine multiple artificial neurons, simple computational nodes resembling the neurons of animals' nervous system, into networks that can perform complex tasks and have the ability to learn. (Nielsen, 2015; Schmidhuber, 2015.)

History of neural networks goes back to 1950's, when a simple artificial neuron called perceptron was developed. A perceptron receives multiple binary values as input factors, and it outputs one binary value. Each of the input factors are assigned a weight for determining their importance on perceptron's output. When these weights are defined, perceptron's inner logic for producing the output is simple: it is determined by calculating the weighted sum of the input factors. If the weighted sum is greater than chosen threshold, perceptron outputs 1 and otherwise it outputs 0. Perceptron's output can be used as input for another perceptron to form network of perceptron. This forms the basic structure of Neural network, as depicted in FIGURE 4. (Nielsen, 2015.)

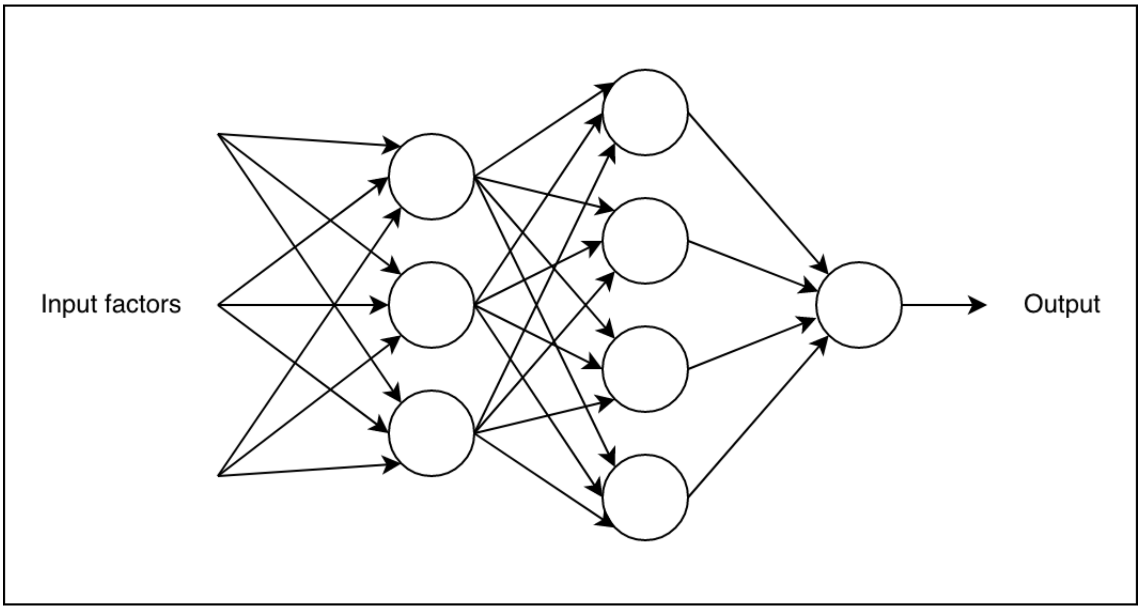


FIGURE 4 Architecture for neural network combining multiple perceptrons (Nielsen, 2015).

Weakness of perceptrons is that small adjustments in one perceptron causes significant change on the output on the rest of the network. This makes networks of perceptrons very hard to adjust, thus obstructing the learning possibility of the network. The concept of perceptrons have been developed further to form more advanced neural networks. Fundamental building block for these networks is a Sigmoid neuron that can also handle non-binary values, as their input and output can be anything between 0 and 1.

Neural networks combine multiple layers of neurons. There are three kinds of neuron layers: input layer, hidden layer and output layer. Input layer of neural network consists of input neurons and output layer has output neurons. Hidden layers lie between the input and output layers and they are called hidden because they are not directly exposed to the outside world of the network. Neural network can have multiple hidden layers. This architecture is depicted in FIGURE 5.

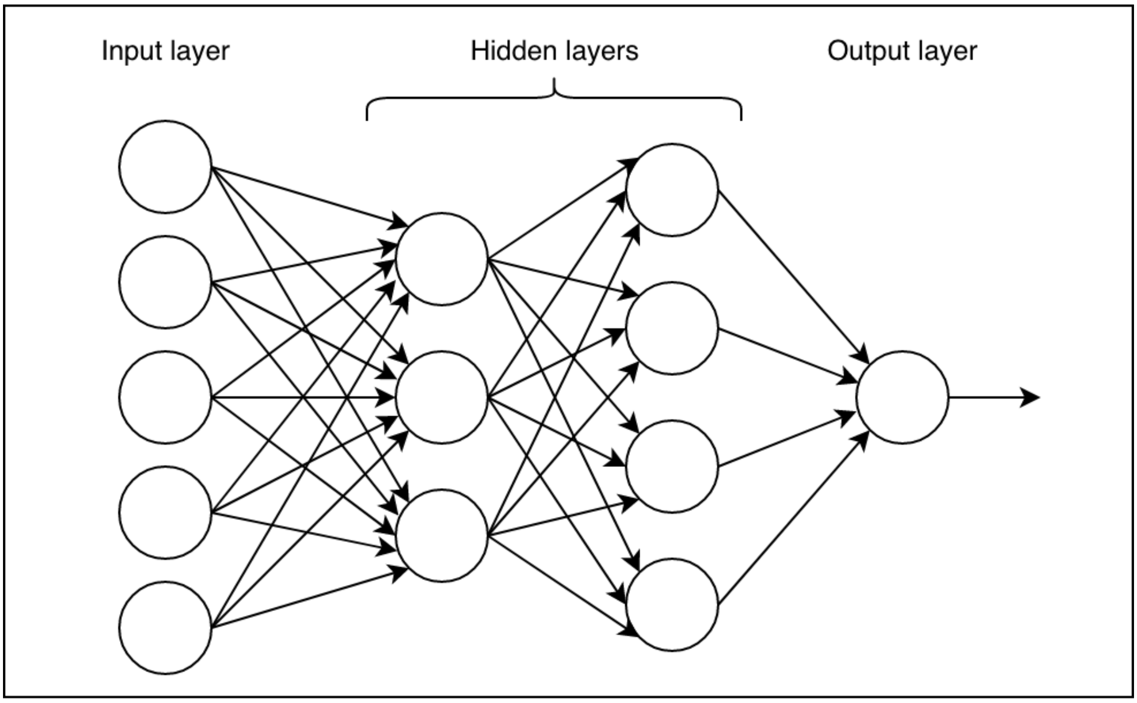


FIGURE 5 The layers of Neural network (Nielsen, 2015)

According to Nielsen (2015), the design of input and output layers is generally straightforward. On the other hand, design of the hidden layers is tightly bound to the context and various heuristics for their implementation have been developed. Yet, the conceived nature of hidden layers can make behavior of the neural network hard to predict and trace.

Network in figure above is a Feedforward neural network. In these networks, outputs of given layer are used as input for the next layer, and no circular references nor feedback loops are employed meaning that information from neurons is always passed forward to the neurons on next layer. (Nielsen, 2015.) However, this is not the only type of neural networks. Recurrent neural networks are utilizing feedback loops and via this they can "learn programs that mix sequential and parallel information processing in a natural and efficient way" (Schmidhuber, 2015).

3.5.3 Explainable AI

There are two reasons why AI solutions can be seemingly black box to their users. First is that the algorithm is too complicated and complex for human to understand. Second reason is that the algorithm or underlying mechanisms is kept secret because the company that developed it considers it a trade secret. To tackle the uncertainty, unpredictability and consequential loss of trust on AI systems, approaches for developing explainable AI have been discussed. (Rudin, 2019.)

Using black box AI raises various concerns especially when they are employed for decisions that have significant impact on individual human beings, but without any clear conception on what the system is actually doing for reaching the outcome. User who is making an AI supported decision may be unaware of which factors the AI system has already taken into consideration when using the AI system in support for decision-making. Such unawareness can result in poor decisions that are founded on unbalanced grounds. This is the result of user's poor understanding on how the system is operating as they could have a misconception on what kind of problems the black box model can actually be applied and what to expect from them. This ambiguity could interfere with decision-making process in situations where the user has to consider exceptional circumstances in addition to the AI system's suggested outcome when making decisions. (Rudin, 2019.)

Using another model to explain a black box model has been suggested to overcome the issues raising from black box. Inaccurate explanation models could actually lower users' trust on the system, acting against their initial purpose. Explanation models for such systems lack details and provide only shallow, or even none, understanding on what the black box model is actually doing. This could unnecessarily complicate decision-making and elevate the risk of human error. Such consequences could again erase the expected benefits of the AI system. (Rudin, 2019.)

Rudin (2019) argues that this can be overcome by developing AI systems with models that are inherently interpretable. Various ways for interpretable model implementation have been examined. Implementation of prototype layer could provide interpretability for AI systems. During its training phase, the AI system picks up prototypical features for each class. These prototypes are used to explain the patterns identified by the AI system, providing explanations that are directly related to the system's underlying mechanism instead of being so-called "educated guesses" on how the system might have produced its outcome.

4 Guidelines for Trustworthy AI

To overcome the ethical challenges presented in previous chapters, European Commission's High-Level Expert Group on Artificial Intelligence (2019) defined principles for trustworthy Artificial Intelligence development. The basic principles that comprise trustworthy systems are lawfulness, ethicalness and robustness. Sometimes these principles can be in contradiction with each other. For example, legal regulations don't match the ethical expectations from time to time. From these principles, the Expert Group has formulated seven key requirements, that should be fulfilled in order for the system to be considered trustworthy. These requirements are linked to the Charter of Fundamental Rights of the European Union. The Expert Group also proposed practices for fulfilling these requirements.

4.1 Key requirements for trustworthy AI

These seven key requirements introduced by European Commission (2019) are used as theoretical framework for this thesis. The requirements and their interrelated nature are depicted in FIGURE 6.

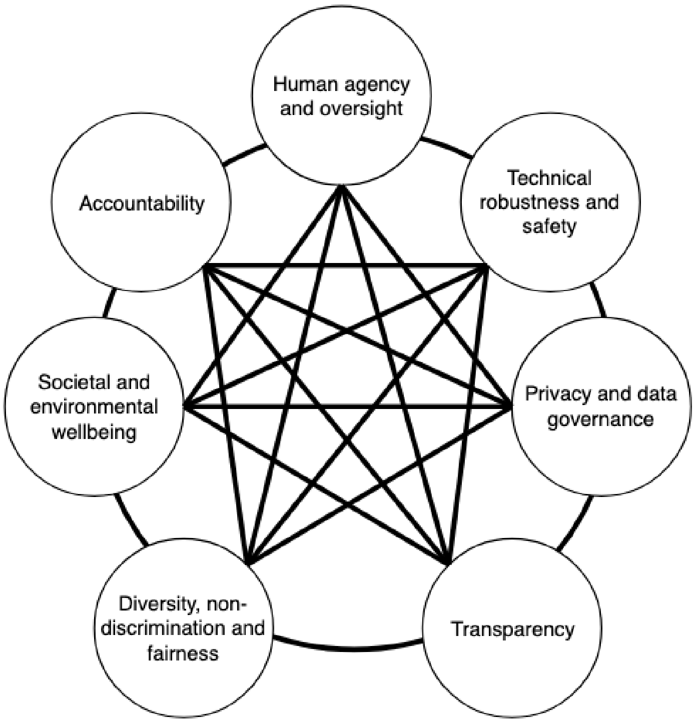

Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability

FIGURE 6 The key requirements for trustworthy AI and how they relate to each other (European Commission, 2019).

In this chapter, these requirements introduced in European Commission's (2019) guidelines are discussed in their original domain Artificial Intelligence (AI) systems.

4.1.1 Human agency and oversight

AI systems should not compromise human autonomy. Thus, AI systems should be developed so that they take the fundamental rights into account. Avoiding infringement of other people's rights and freedom is important. Risks for such infringements should be assessed during every phase of the system's lifecycle. After assessment it should be evaluated whether those risks could be either reduced or justified as necessary. Assessment of the impact on fundamental rights can be reinforced with external feedback mechanism. (European Commission, 2019.)

AI systems should help human users to achieve their individual goals. They do this by helping the users to make better informed decisions. Users should be able to self-assess and challenge the AI system when necessary. When making decision that have significant impact on individuals, every one of them is entitled to not being a subject of solely automatic processing. This means that human user should be the agent when AI systems are used, and user autonomy should be maintained in every situation. (European Commission, 2019.)

To ensure the fundamental rights and human agency, the AI system should include human oversight in its' processes. There are various governance mechanisms that can be used for achieving oversight. The mechanisms mentioned in Commission's report include human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC) approach. In HITL, human user has capability to review every decision the AI system makes. In HOTL, human user can monitor AI system's operation and intervene with the system during its' design cycle. HIC approach means that user has the ability decide when and how to use the system and whether to use it at all. (European Commission, 2019.)

4.1.2 Technical robustness and safety

Ensuring technical robustness and safety is the second mentioned key requirement in Commission's report. This means that the chances of an AI system operating unintentionally and unexpectedly should be minimized and any unacceptable harm that it could cause is prevented. (European Commission, 2019.)

AI systems are liable to vulnerabilities like other software systems. Common attacks performed on AI systems include data poisoning, model leakage and attacks against the software or hardware infrastructure. Resilience to attacks is crucial for technical robustness and safety as such attacks may cause unexpected behavior or even shut down the AI system completely. (European Commission, 2019.)

However, if there occurs some problem within the system, a fallback procedure should be in place. This procedure could require a human involvement before the system continues operation. Fallback procedure should be in place and general safety considered so that the AI system won't cause any physical harm to people or environment. (European Commission, 2019.)

The decisions made by AI systems should be accurate, reliable and reproducible. Accuracy refers to the system's ability to make proper predictions and classifications. Reliable AI systems work properly with various inputs and situations. Reproducibility in the other hand means that AI system should produce the same result when exposed to similar conditions. (European Commission, 2019.)

4.1.3 Privacy and data governance

All personal data and information that user hands over to the system and is collected on them should be protected. Maintaining trust to the data collection process is crucial. It should be clear that any data collected by the system is used only in lawful and non-discriminative way. Access to data should also be limited only to the people who are qualified and circumstances that are necessary. (European Commission, 2019.)

AI systems are often fueled with data. This means that data guides their behavior. Due to this, the quality and integrity of data is very important for proper operation of the system. Perpetrators may try to alter AI system's behavior by tampering the data (data poisoning). Integrity of the data should be ensured by testing and documenting used data sets. (European Commission, 2019.)

4.1.4 Transparency

Transparency and open courses of action are considered as fundamental part of good governance. Prerequisite for transparency is that the user is aware of the fact that they are interacting with AI system at all times. System's capabilities and limitations should also be clearly communicated to the user. (European Commission, 2019.)

Processes, data sets and algorithms that produce a decision made by AI system should be documented comprehensively. This should make it possible to trace why and how AI system produced such outcome. In case of erroneous AI-supported decision, traceable AI systems provide possibility to identify factors that constitute to the occurred mistake. (European Commission, 2019.)

The decisions and technical processes behind them should be explainable. In technical point of view, the decisions and the process that lead to them should be understandable for human users. Explainability is a major concern especially when the decisions made by AI system affect living beings. (European Commission, 2019.)

4.1.5 Diversity, nondiscrimination and fairness

AI systems should enable inclusion and diversity. There is possibility for unfair biases in historic data. Using such data in AI systems would reinforce these old biases and promote discrimination against certain people. Underlying biases may affect AI system also in its' development phases. This means that not only data but used algorithms too should be assessed to prevent discriminatory biases. (European Commission, 2019.)

AI systems should be accessible to people regardless of age, gender, abilities or characteristics. European Commission considers that accessibility for people with disabilities is paramount and principles of Universal Design should be used to serve extensive range of people. (European Commission, 2019.)

Stakeholders that are affected by the AI system during its' lifecycle should be involved in its' development. There should also be a feedback mechanism that can be used for soliciting user feedback when the system is in use. Thus, different stakeholders should be included in AI system development at all times. (European Commission, 2019.)

4.1.6 Societal and environmental wellbeing

Above-mentioned key requirements considered mainly human beings as stakeholders for AI systems. However, developing trustworthy AI also requires assessment from the point of view of society, nature and environment. When employed appropriately, benefits of sustainable AI systems will last for future generations. (European Commission, 2019.)

Environmental sustainability should be ensured during the system's lifecycle. Impact on environment should be assessed in every stage of AI system's supply chain. (European Commission, 2019.)

AI systems can also have social impact. They have been perceived useful in enhancing social interaction (e.g. interacting with autistic children). However, it should be noted that they can impair social skills as well, if the social aspect is not taken into account in development of social AI systems. (European Commission, 2019.)

Using AI has also impact on our society as a whole. If used in democratic processes or decision making in political context, users should be particularly cautious when implementing AI systems in societal context. AI system's impact on democracy and its' institutions should be assessed carefully. (European Commission, 2019.)

4.1.7 Accountability

Finally, AI systems should be held accountable for their actions in order to be considered trustworthy. There should be mechanisms that facilitate auditability, reporting and redressing. Accountability should be maintained in all phases. (European Commission, 2019.)

Algorithms, data and design processes of AI systems should be available for auditing. This evaluation can be done by both internal and external auditors. Open audit reports increase trustworthiness further. Independent auditing is especially crucial for safety-critical systems that have impact on fundamental rights mentioned above. (European Commission, 2019.)

Negative impacts of AI system should be minimized and reported. Possible negative outcomes that may come from using the AI system should be comprehensively documented and assessed. Various impact assessment methods can be used proportionally to the posed risk. (European Commission, 2019.)

Fulfilling all the requirements mentioned above can lead to situations where decision-makers are obliged to make trade-offs. These trade-offs should be explicitly brought forward and solved in rational and methodological way. Implemented trade-offs should be continuously reviewed during the lifecycle of the AI system. (European Commission, 2019.)

Despite all the effort mentioned above, usage of AI system might result in harmful outcomes. In these situations, adequate redress mechanisms should be ensured. The possibility to redress is vital for securing the trust on AI systems. (European Commission, 2019.)

4.2 Practices for implementation of Trustworthy AI

To fulfill the requirements mentioned above, European Commission (2019) proposes a set of technical and non-technical practices. These practices should be used in every phase of AI system's lifespan.

4.2.1 Technical practices

Technical practices implied by European Commission (2019) are trustworthy architectures, ethics and lawfulness by design, explanation methods, testing and validation and service quality indicators. These practices are depicted below in TABLE 3.

TABLE 3 Technical practices for fulfilling the key requirements of Trustworthy AI (European Commission, 2019).

Method	**Description**
Trustworthy architectures	Requirements should be translated into procedures and constraints that are implemented directly into the AI system's architecture. These procedures and constraints should be adapted in every step of the sense-plan-act cycle of non-deterministic AI systems.
Ethics and lawfulness by design	Ethical norms and legislation should be implemented into the design of AI systems. Fail-safe shutdown and continuation mechanisms should also be designed.
Explanation methods	Underlying mechanisms of AI systems and their outputs should be explainable. This is still an open challenge especially for neural networks.
Testing and validation	Novel testing methods are required for AI systems because of their non-deterministic nature. Unexpected and undesired behavior of AI system may occur only when used with realistic data (i.e. in the real-life environment). Due to this, the data and model encompassed in the system should be validated as early as possible in addition to traditional software testing.
Service quality indicators	Quality of Service should be appropriately measured. This includes traditional software metrics such as functionality, performance, usability, reliability, security and maintainability. Novel metrics can also be used to ensure that AI system is

developed in safe and secure manner.

The maturity of these practices varies. Due to this, some of them are straightforward to apply into daily processes of software development while the others (e.g. explanation methods) still remain an open problem. (European Commission, 2019.)

4.2.2 Non-technical practices

European Commission (2019) also imply non-technical practices that support trustworthy development of AI. These are regulation, codes of conduct, standardization, certification, accountability via governance systems, education and awareness of ethical mindset, stakeholder participation and social dialogue and diverse and inclusive design teams. Non-technical practices are shown in TABLE 4.

TABLE 4 Non-technical methods for fulfilling the key requirements of Trustworthy AI (European Commission, 2019).

Method	Description
Regulation	New regulations should be adopted, and prevailing regulations should be assessed and updated to match the requirements of trustworthy AI systems. This means that regulation should both enable and restrict the use of AI systems.
Codes of conduct	Organizations should develop clear guidelines to be followed. These guidelines should consider corporate responsibility, display organizational intentions and emphasize the desirable values when using AI systems.
Standardization	Standardization of e.g. design patterns, manufacturing processes and business practices can be used to ensure and reinforce ethical use of AI systems.
Certification	Certification by independent organizations can provide people with shallow understanding of AI systems' functions and impact a decent conception of its trustworthiness.
Accountability via governance systems	Accountability from the ethical point of view should be ensured with governance frameworks that assess the decisions made in each stage of AI systems' lifespan. Governance should encourage open discussion about possible dilemmas, emerging issues and ethical concerns.
Education and awareness of ethical mindset	The potential impact of AI systems should be communicated, educated and trained to every stakeholder, so they know that they have chance to shape societal development when utilizing AI.
Stakeholder participation and social dialogue	Open discussion with stakeholders, societal institutions and the general public should be encouraged. Active participation of stakeholders and dialogue on AI

	systems' impact contribute to validating the implemented approaches within the AI system.
Diverse and inclusive design teams	Design teams for AI systems should reflect the diversity of their users and the society where the system operates. This is important for developing AI systems that take all the necessary perspectives into account.

In order to develop Trustworthy AI, these non-technical practices should be used and assessed in every phase of AI system's lifecycle (European Commission, 2019).

4.3 Conceptual framework

European Commission's (2019) guidelines for trustworthy AI, that were presented in this chapter, form the theoretical framework for this study. The guidelines include seven key requirements, that should be realized for implementation of trustworthy AI systems, and list of thirteen suggested practices that should contribute towards realization of the requirements.

Human agency and oversight combine the aspects of "fundamental rights, human agency and human oversight". These aspects mainly consider the rights of human beings and human autonomy over the AI system. (European Commission, 2019.) Practices that are expected to contribute to their realization are presented in TABLE 5.

TABLE 5 Practices with expected contribution to Human agency and oversight

Practice	**Relation**
Trustworthy architectures	Architectures ensuring human oversight
Ethics and lawfulness by design	Inclusion of fundamental rights to the design
Explanation methods	Facilitating oversight with explanation methods
Regulation	Regulation for ensuring fundamental rights
Standardization	Standards for human oversight
Accountability via governance systems	Governance systems for achieving oversight and agency
Education and awareness of ethical mindset	Educate and create awareness on AI's impact on fundamental rights and human agency

Technical robustness and safety are realized along the aspects of "resilience to attack and safety, fallback plan and general safety, accuracy and reliability and reproducibility". These aspects are concerned on the predictable, reliable, accurate and safe technical operation of the system. (European Commission, 2019.) Practices that are expected to contribute to their realization are presented in TABLE 6.

TABLE 6 Practices with expected contribution to Technical robustness and safety

Practice	Relation
Trustworthy architectures	Architectures increasing general safety by mitigating complexity
Ethics and lawfulness by design	General safety and prevention of harm with designed fail-safe mechanisms
Explanation methods	Assessing the accuracy of the system
Testing and validation	Ensuring aspects of general safety and accuracy
Service quality indicators	Indicators to ensure the aspects of accuracy, safety and security
Standardization	Standardization of systems design to ensure overall robustness
Certification	Independent certifying to ensure overall robustness and safety

Privacy and data governance have the aspects of "respect for privacy and data protection, quality and integrity of data and access to data". These aspects are especially concerned on proper use of personal data and ensuring that data is handled appropriately. (European Commission, 2019.) Practices that are expected to contribute to their realization are presented in TABLE 7.

TABLE 7 Practices with expected contribution to Privacy and data governance.

Practice	Relation
Trustworthy architectures	Architectures that provide mechanisms for privacy
Ethics and lawfulness by design	Ensuring respect for privacy by design
Testing and validation	Validation for quality of data
Regulation	Regulating the use of data
Codes of conduct	Codes of conduct for handling personal data
Accountability via governance systems	Governance frameworks for data governance
Diverse and inclusive design teams	Diverse teams to ensure inclusive data governance

Transparency includes the aspects of "traceability, explainability and communication". These require that companies are transparent both on their own intentions and their AI systems' inner mechanisms. (European Commission, 2019.) Practices that are expected to contribute to their realization are presented in TABLE 8.

TABLE 8 Practices with expected contribution to Transparency.

Practice	Relation
Trustworthy architectures	Architectures that ensure traceability and explainability
Ethics and lawfulness by design	Designing systems to be traceable
Explanation methods	Ensuring the aspect of explainability and traceability

Testing and validation	Ensuring explainability
Regulation	Enforcing communication and traceability via regulation
Codes of conduct	Documentation of organizational intentions to establish transparent operations
Standardization	Standards to improve explainability and communication
Certification	Independent certifying to ensure traceability and explainability
Accountability via governance systems	Governance frameworks that demand traceability
Education and awareness of ethical mindset	Ethical mindset for encouraging transparent communication
Stakeholder participation and social dialogue	Realizing the aspects of explainability and communication with open discussions

Diversity, nondiscrimination and fairness considers aspects of “unfair bias avoidance, accessibility and universal design and stakeholder participation”. The major concern on these aspects is that diversity is taken into account on decisions related to the AI system. (European Commission, 2019.) Practices that are expected to contribute to their realization are presented in TABLE 9

TABLE 9 Practices with expected contribution to Diversity, nondiscrimination and fairness.

Practice	Relation
Ethics and lawfulness by design	Avoiding unfair biases via design
Testing and validation	Validating system’s operation to avoid biases
Codes of conduct	Considering responsibility from the point of view of avoiding biases and discrimination
Stakeholder participation and social dialogue	Establishing inclusiveness by open discussion with stakeholders
Diverse and inclusive design teams	Diverse design teams to ensure realization of diversity

Societal and environmental wellbeing combines the aspects of “sustainable and environmentally friendly AI, social impact and society and democracy”. These aspects state that the widespread impact of AI systems on both environment and society should be discussed. (European Commission, 2019.) Practices that are expected to contribute to their realization are presented in TABLE 10.

TABLE 10 Practices with expected contribution to Societal and environmental wellbeing.

Practice	Relation
Trustworthy architectures	Developing architectures that consider sustainability
Ethics and lawfulness by design	Including aspects of sustainability, social impact and democracy into design
Regulation	Regulation for ensuring sustainable AI

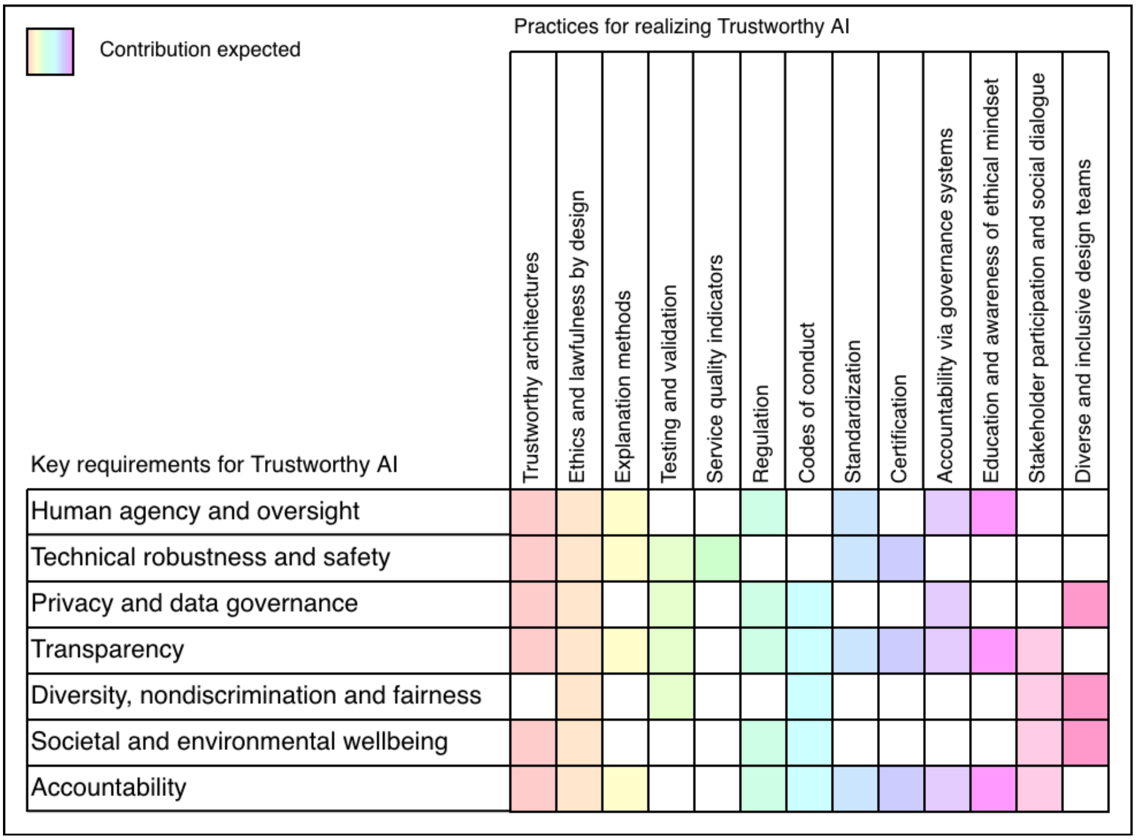
Codes of conduct	Documenting intentions from the point of view of society
Stakeholder participation and social dialogue	Social dialogue and stakeholder participation on AI impacting society
Diverse and inclusive design teams	Diverse design teams to include various viewpoints on social impact

Accountability has the aspects of "auditability, minimizing and reporting negative impact, documenting trade-offs and ability to redress". They consider the accountability of AI systems operations. (European Commission, 2019.) Practices that are expected to contribute to their realization are presented in TABLE 11.

TABLE 11 Practices with expected contribution to Accountability.

Practice	Relation
Trustworthy architectures	Architectural solutions for ensuring auditability
Ethics and lawfulness by design	Including aspects of auditability, reporting and ability to redress into design
Explanation methods	Explanation methods provide traceability, which is making auditability possible
Regulation	Regulation for auditability, redress and trade-offs
Codes of conduct	Documenting trade-offs and negative impact
Standardization	Standards to improve auditability
Certification	Certification for auditability
Accountability via governance systems	Governance frameworks that enforce auditability, reporting, documentation and redress
Education and awareness of ethical mindset	Ethical mindset for reporting of negative impact, trade-off documentation and redress
Stakeholder participation and social dialogue	Improving auditability with involving stakeholders

These remarks are formed into Primary conceptual conclusion, which is depicted in FIGURE 7, where colored square indicates that practice is expected to contribute towards realization of each key requirement. As seen in the figure, each practice could contribute towards realization of multiple requirements. This forms the theoretical framework for this study.



■ Contribution expected

Key requirements for Trustworthy AI	Practices for realizing Trustworthy AI												
	Trustworthy architectures	Ethics and lawfulness by design	Explanation methods	Testing and validation	Service quality indicators	Regulation	Codes of conduct	Standardization	Certification	Accountability via governance systems	Education and awareness of ethical mindset	Stakeholder participation and social dialogue	Diverse and inclusive design teams
Human agency and oversight	■	■	■			■		■		■	■		
Technical robustness and safety	■	■	■	■	■			■	■				
Privacy and data governance	■	■		■		■	■			■			■
Transparency	■	■	■	■		■	■	■	■	■	■	■	
Diversity, nondiscrimination and fairness		■		■			■					■	■
Societal and environmental wellbeing	■	■				■	■					■	■
Accountability	■	■	■			■	■	■	■	■	■	■	

FIGURE 7 Primary conceptual conclusion for this study that shows expected practices for realizing key requirements for Trustworthy AI.

5 Research method

Objective of the empirical part of this study was to identify which practices for development of trustworthy AI are employed in the industry and assess the consideration of requirements for trustworthy AI.

5.1 Data collection

Data for the study was extracted from a larger set of data considering Artificial Intelligence ethics. Data collection for set was conducted as a structured interview. This was done as a survey.

In the original survey had demographic questions, quantitatively measured Likert-scale questions and open-ended questions. This study focused on the open-ended questions of the study. In the survey outline, Artificial Intelligence ethics was defined via four principles. In alphabetical order, these principles were Accountability, Predictability, Responsibility and Transparency. In the questions, the respondents were asked how their organizational policies and practices take these principles into account when developing AI systems.

From the total 249 responses of the survey, 39 companies were chosen for the study as they had answered the open-ended questions. As the data consisted of responses to open-ended questions, research was conducted as qualitative analysis.

5.2 Description of the data

Sample consisted of 39 companies, that were developing AI systems. These companies were selected to the sample as they had answered the open-ended questions in more than three words. Companies' sizes ranged from small (1-9 employees) to large (over 500 employees). Companies and the roles of respondents are depicted in TABLE 12.

TABLE 12 Companies and respondents for the study.

ID	Company size	Role of respondent(s)
C1	10-49	Sales Director
C2	10-49	Technical service responsible
C3	10-49	Software engineering, product development
C4	10-49	Front end developer
C5	10-49	CEO and product owner/designer
C6	10-49	CEO
C7	10-49	Product owner
C8	1-9	CTO
C9	1-9	Supervisor
C10	1-9	Specialist (creating business development related content/requirements)
C11	1-9	Consultant
C12	1-9	Developer
C13	1-9	Creative director
C14	1-9	Developer, architect
C15	250-499	Full Stack Developer
C16	250-499	Product manager
C17	250-499	Business line director
C18	500+	N/A
C19	500+	Lead consultant/Architect
C20	500+	Administrator and development specialist
C21	500+	Image preprocessing, training data set validation dataset
C22	500+	Team lead
C23	500+	Integration specialist
C24	500+	Requirements engineer/consultant
C25	500+	Product manager/owner
C26	500+	N/A
C27	500+	Senior testing specialist
C28	500+	Business analyst.
C29	500+	N/A
C30	500+	BI/AI services Area Manager, Project Manager
C31	500+	Project Manager
C32	500+	Senior systems architect
C33	50-249	Software designer
C34	50-249	Project manager in delivery projects
C35	50-249	CTO (leading whole RNG incl. SW development)
C36	50-249	AI specialist
C37	50-249	Head of a competence organization
C38	50-249	Software Developer
C39	50-249	Project manager

5.3 Data analysis

Data was analyzed with qualitative thematic analysis. Theming was used in order to identify central concepts from the data. Aim of the thematic analysis was to find out which practices companies within the sample use when pursuing trustworthiness. From the identified concepts, a list of practices was formed. Coding approach for identifying the concepts is presented later in this chapter. After the list of practices was formed, they were mapped to the key requirements in order to identify which practices contribute to fulfilling each requirement.

According to Cruzes and Dybå (2011), there are three approaches to coding the data: deductive, inductive and integrated. Deductive coding uses list of codes that is defined *a priori* to the actual coding of the data. Inductive coding lets the data speak of itself, and the codes are assigned and organized as the reading of data goes on and perceivable concepts emerge from the data. Integrated approach uses both of these, as codes can be created on basis of both emerging concepts from the data (inductive approach) and pre-defined structure of codes (deductive approach). (Cruzes & Dybå, 2011.)

Cruzes and Dybå (2011) identified three common problems for coding of data: coding is too general, identifying desired concepts instead of seeing what the data is saying and out of context coding. Codes should also have limited scope and clear definition. They should have unique semantics, meaning that the codes are not interchangeable. (Cruzes & Dybå, 2011.)

Data for this study was coded using integrated approach. List of practices by European Commission (2019) was used as start list for coding, but novel codes were also formed when clear concepts were identified during reviews of data. Deductive coding of integrated approach ensures that coding stays in the context and inductive coding inhibits too general codes and leaves room for concepts that emerge from the data.

6 Empirical findings

This chapter depicts the empirical findings of this study. First, an overview of the data and its coding is presented. Then realization of each requirement from European Commission's (2019) report is assessed.

6.1 Overview

Goal of the analysis phase was to identify practices within the data and assess their contribution towards realization of trustworthy AI. Each piece of data was assigned both an inductive and deductive code during the analysis. First, each quotation was assigned inductive code. This was done when perceivable concepts emerged from the data as the analysis moved on. These inductive codes were then grouped under deductive codes, which were formed from the practices mentioned in guidelines of European Commission (2019). Pieces of data that did not indicate any practice were assigned the code "Observation". Assigned codes and their occurrences are presented TABLE 13.

TABLE 13 Assigned codes and their occurrences within the data

Deductive code	Inductive code	Occurrences
Accountability via governance	Accountability guidelines	2
Accountability via governance	Defined responsibilities	7
Certification	Audits	1
Certification	Certification organizations	1
Codes of conduct	Company policies	4
Codes of conduct	Contract	5
Codes of conduct	Decision-making practices	2
Codes of conduct	Documentation	7
Codes of conduct	Operational guidelines	5
Codes of conduct	Preparation	1
Education and awareness of ethical mindset	Change agency	2

Education and awareness of ethical mindset	Trainings	1
Explanation methods	Reviewable models	1
Explanation methods	Software audit trail	2
Observation		8
Regulation	Following regulations	3
Service quality indicators	Cost tracking	1
Service quality indicators	Time tracking	2
Stakeholder participation	Customer involvement	4
Stakeholder participation	Informed customer	3
Standardization	Agile methods	3
Standardization	Unified processes	6
Testing and validation	Code reviews	3
Testing and validation	Testing	5
Trustworthy architectures	Proper architectural solutions	2

Based on thematic analysis, codes of conduct were the practice that emerged in most of the responses. Some of the practices contributed to realization of multiple requirements. There were also few practices that were not identified within the data: Ethics and lawfulness by design and Diverse and inclusive design teams. This finding forms the first empirical conclusion.

> EC1: Practices of Ethics and lawfulness by design and Diverse and inclusive design teams were not present within the data.

It was also noted that requirements of Societal and environmental wellbeing and Diversity, nondiscrimination and fairness were not operationalized in the survey outline, forming the first primary empirical conclusion PEC1.

> PEC1: Requirements of Societal and environmental wellbeing and Diversity, nondiscrimination and fairness were not discussed within the data.

Realization of the rest of the requirements from the theoretical framework are being assessed in later chapters.

6.2 Human agency and oversight

Requirement of Human agency and oversight is supposed to ensure the respect for human autonomy. AI system should support human decisions and human oversight of system's operation should be enabled. (European Commission, 2019.)

Regulation is one of the contributing practices for Human agency and oversight. In the data there was no practices that related to regulation itself.

This probably is due to the fact, that regulation is done by governments and not by software companies that formed the sample for this study. Nevertheless, present practice of following regulations is consequence of regulatory power executed by legislative bodies. Some companies mentioned that their main practice for achieving trustworthiness was following regulations. One respondent considered their system's criticality so minor, that to achieve trustworthiness they relied only on following regulations. This could imply, that following regulations is the least that companies should do in order to achieve trustworthiness.

> "We follow regulations, but since our software is not a very risky one, we haven't taken much caution." - Respondent C36

Another response supports the interpretation, that following regulations is the minimal effort for trustworthiness. When asked if their organization's policies consider trustworthiness, they responded:

> "I think so, following regulations at least" - Respondent C15

Using the expression "at least" could suggest that following regulations is the least one should do. Thus, it can be seen as something that companies must do in order to achieve trustworthiness. Despite being the least effort for achieving trustworthiness, prevailing regulations do not rule in Human agency and oversight. This finding forms the following empirical conclusion:

> EC2: Following regulations is considered to be the least that companies must do for achieving trustworthiness, but they do not take Human agency and oversight into account.

Explanation methods is considered to contribute towards Human agency and oversight. This practice did emerge only in one response. Reason behind this could be the overall immaturity of explanation methods. Only one respondent discussed the reviewability of their AI models, highlighting that they have the possibility to inspect the underlying mechanisms, thus providing possibility of human oversight.

> "Machine learning models can be inspected afterwards by feeding in different data and analyzing the output." - Respondent C14

Same respondent also threw into relief another practice linked to explanation methods: Audit trail of software's operation. However, they pointed out that this doesn't solve the problem of explainability entirely.

> "Software is designed to maintain an audit trail log of all actions made by or with it, analyzable afterwards in some detail but of course not everything." - Respondent C14

With only one respondent discussing explainability methods, it can be concluded that explainability is not a common practice for achieving trustworthiness. This forms the empirical conclusion EC3.

> EC3: Explainability of AI system's underlying mechanisms is not highlighted when discussing trustworthiness and current explanation methods do not consider Human agency and oversight.

Accountability via governance systems is expected to provide overall oversight on AI system's operation and the company behind its implementation. The overall oversight can also be expected to cover Human oversight on AI systems' operation. Accountability guidelines make up one governance system that was expected to contribute towards trustworthiness.

> "[COMPANY NAME] management foundation sets rules and guidelines for everything we do so that we are trustworthy partner" - Respondent C19

> "The goal is that we would have clear decision making practices, documentation practices and well defined organisation and roles. There is still some work on this but direction is clear." - Respondent C35

Clearly defined responsibilities formed another data emergent practice that was grouped under Accountability via governance systems. With this practice, responsibilities of each party are explicitly defined, thus making it easier to assess who is accountable on system's operation.

> "The responsible and accountable parties for each development objects are defined are defined in the beginning of the project and also before signing the contract of the development work to be done." - Respondent C18

> "The written contracts with the end customers directly specifies their rights and our responsibilities." - Respondent C9

It is still somewhat vague whether these practices of Accountability via governance systems are taking Human agency and oversight into account or not.

> "We are responsible for every possible mistakes made during development process" - Respondent C4

> "Usually mentioned in contracts with customers. Each developer is responsible for the delivered feature. The features can easily be tracked to the individuals involved in the implementation. So yes, accountability and responsibility are considered in the software development." - Respondent C27

Quotation above suggests that responsibility is mainly perceived as software provider's responsibilities towards the paying customer, thus leaving e.g. end users out of its scope. Responsibilities considering the impact on human beings caused by AI system is not given that much attention. This forms the empirical conclusion EC4.

> EC4: Governance frameworks are mainly used to ensure responsibilities and oversight between software providers and their customers and not for ensuring Human agency or oversight over the system.

Standardization can also be used to achieve Human agency and oversight. There emerged two practices related to Standardization in the data. From these, unified processes can contribute to Human agency and oversight. Below are some examples of unified processes that respondents expected to contribute towards trustworthiness.

> “It's the reason the processes exist and are document and trained.” - Respondent C37

> “Processes are unified and responsibilities are defined.” - Respondent C30

> “We use process and tools and those are transparent to all in company.” - Respondent C30

However, emerged Standardization practices within the data were depicted in such general level that it is hard to assess whether they take Human agency into account or not, thus forming the conclusion EC5.

> EC5: Standardization practices don’t take Human agency and oversight into consideration.

Education and awareness of ethical mindset was expected to contribute towards Human agency and oversight in the conceptual framework. There were few occurrences of such companies, which embraced strong change-oriented core values. Their contribution towards realization of Human agency and oversight was however left unclear.

> EC6: Education and awareness of ethical mindset is practiced in only few companies, where they are based on strong change-oriented core values, with no explicit contribution to Human agency and oversight.

Trustworthy architectures and Ethics and lawfulness were considered practices that contribute to realization of Human agency and oversight. As stated in empirical conclusion EC1, there were no quotations coded as Ethics and lawfulness by design. Though there were practices coded as Trustworthy architectures within the data, no architectural solutions for mechanisms to ensure human oversight was recognizable within the responses.

> EC7: Architectural solutions for trustworthiness exist but they are not used for realizing Human agency and oversight.

In total, there were five empirical conclusions for Human agency and oversight. Based on these conclusions, primary empirical conclusion PEC2 was formed.

PEC2: Companies employ practices that could contribute towards the requirement of Human agency and oversight, but they do not consider the requirement particularly, leaving their contribution towards Human agency and oversight vague.

Empirical conclusions and the primary empirical conclusion for Human agency and oversight are presented in TABLE 14.

TABLE 14 Empirical conclusions for Human agency and oversight

Identifier	**Empirical conclusion**
EC2	Following regulations is considered to be the least that companies must do for achieving trustworthiness, but they do not take Human agency and oversight into account.
EC3	Explainability of AI system's underlying mechanisms is not highlighted when discussing trustworthiness and current explanation methods do not consider Human agency and oversight.
EC4	Governance frameworks are mainly used to ensure responsibilities and oversight between software providers and their customers and not for ensuring Human agency or oversight over the system.
EC5	Standardization practices don't take Human agency and oversight into consideration.
EC6	Education and awareness of ethical mindset is practiced in only few companies, where they are based on strong change-oriented core values, with no explicit contribution to Human agency and oversight.
EC7	Architectural solutions for trustworthiness exist but they are not used for realizing Human agency and oversight.
PEC2	Companies employ practices that could contribute towards the requirement of Human agency and oversight, but they do not consider the requirement particularly, leaving their contribution towards Human agency and oversight vague.

6.3 Technical robustness and safety

AI systems should be robust and safe from the technical point of view. This means that any harm that could possibly result from the use of AI system should be prevented. AI systems should be resilient to adversary actions by external agents, have procedures that ensure fail-safe operation in unexpected situations and operate in predictable and explainable manner. (European Commission, 2019.)

Software architecture is fundamental for building robust software. European Commission (2019) proposes that companies should take trustworthiness of AI into account when designing the system and embed its requirements in its architecture. Two respondents within the data stated that they rely on proper architectural solutions to achieve trustworthiness. Robustness was pursued by seeking architectural solutions that can stand the test of time instead of providing short term returns.

> “We always try to select long term architectural solutions” - Respondent C39

Another respondent did put emphasis on the same subject by stating that their goals is to find appropriate architectural solutions, even if its implementation required additional effort.

> “We focus on finding the right solution, not the easiest.” - Respondent C17

As discussed in previous chapter, it is unclear whether architectures take the ethical dimension of AI into account or not. When considering Technical robustness and safety, it can be said that architecture is used to ensure development of technically robust and safe systems as the companies are seeking for long-term solutions for their systems. This lays the groundwork for conclusion EC8.

> EC8: Architectures are used in some extent for realization of Technical robustness and safety.

Service quality indicators provide different technical metrics, such as functionality, performance and reliability. These metrics can be used when assessing technical robustness and safety. Beyond these traditional software metrics, new indicators to assess AI systems technical Quality of Service should be adopted. (European Commission, 2019.) Survey data featured three practices that involved explicit metrics. However, the metrics present in the data were more concerned about measuring software development from the viewpoint of project management. This was done by tracking time and costs spent on developing software.

> “Cost are reported to owner and follower” - Respondent C30

> “Tracking all development tasks and time used for them with various tools” - Respondent C14

> “People responsible of each project deliverable will deliver those in time” - Respondent C18

This could suggest that companies consider QoS indicators as a tool for project management rather than ensuring trustworthiness, forming the empirical finding EC9.

> EC9: Quality of Service is primarily measured from the perspective of project management.

Practices of Testing and validation were highlighted in nine cases within the data. Code reviews among developers was the most common Testing and validation practice within examined companies.

> “Heavy testing and continuous code reviewing.” - Respondent C5

> “Full documentation and tools where we can address who has made what, code reviews between developers.” - Respondent C34

> “No code gets to the production if no other developer has given it a review and approved it first” - Respondent C27

Underlying assumption supporting code reviews could be the perception that two pairs of eyes are better than one. When the same piece of code is run through by several developers, potential bugs are more likely to be identified. This way code reviews and, software testing in general, are expected to provide herd immunity for the code and ensure robustness and reliability.

> “All the operations are tested in test environments by several coders before publishing.” - Respondent C7

> “People accountable will review the solutions to documents and provide an auditable track record.” - Respondent C18

Software testing in general was identified as major contributor to trustworthiness. Testing was described as a critical part of development process, and it was expected to be executed actively with clockwork precision.

> “We have strong "process" culture where are described responsibilities etc. Rigorous and thorough testing process.” - Respondent C23

> “Active testing during the process and sprints” - Respondent C4

Testing and validation are also executed by testing the system using real-life scenarios. This should provide validation for the system and give insight on how it is going to act in given situations.

> “Development work is always done by testing practical examples.” - Respondent C20

Insight given by validation could prove crucial especially for non-deterministic systems that AI systems often are. This kind of validation could ensure system’s safe and predictable operation, benefiting technical robustness and safety. Testing and validation practices are concluded with conclusion EC10.

> EC10: Traditional software testing methods and validation by code reviews are common practices used for ensuring Technical robustness and safety.

Explanation methods are also expected to provide reliability and reproducibility, that are requisites for building safe and robust AI systems. As stated in previous chapter, one respondent presented, that they have ability to validate system's behavior using different sets of data. This makes it possible to analyze the underlying mechanisms of their software.

> "Software is designed to maintain an audit trail log of all actions made by or with it, analyzable afterwards in some detail but of course not everything. Machine learning models can be inspected afterwards by feeding in different data and analyzing the output" - Respondent C14

In addition to Reviewable models, the quotation above also includes another practice coded as Software audit trail. This means that their software maintains an audit log of its operation. Audit log could contribute towards Technical robustness and safety, as it can be used to track down how the system operated in the conditions in question. This comprises the conclusion EC11.

> EC11: Explanation methods can be used to validate Technical robustness and safety, but they are not common practice.

Certification by independent organizations can provide people who are not familiar with AI systems a way to assess system's trustworthiness. In the data, there was no certification practices that assessed the operation of AI system itself. However, there was one instance of certification practice that was related to software development. This respondent told that they used certified software development method called Scaled Agile Framework. This method is expected to maintain the quality of the software when developing increasingly complex systems (Leffingwell, 2018).

> "SAFe model offers policies for things that mentioned above." - Respondent C28

Using certified agile methods is expected to have impact on quality of the software. Thus, employing certified methods should contribute to Technical robustness and safety of AI systems.

> EC12: Certified software development methods are used for realization of Technical robustness and safety.

In addition to using certified software development frameworks, commitment to agile principles was identified as a contributor to trustworthiness by various companies within the sample. These codes were organized under practice of Standardization as they provide coherent practices for software development.

> "We aim to use best practices of the agile SW development and increasingly also include documentation as part of the SW process." - Respondent C35

Respondents suggest that use of agile methods contributes to trustworthiness from various perspectives. From the perspective of technical robustness, agile methods and the subsequent software quality can be perceived as a contributing factor towards robust and reliable systems.

> “Our project aim to follow agile methods and as a part of them, the processes like planning sprints and reviewing the developed solutions are very transparent to the customer. Also predictability is high because of these operations.” - Respondent C26

The quotation above suggests that agile methodologies contribute to predictability. Yet, it’s open for interpretation whether predictability is covering AI system’s operation, the software development process itself or both. When putting agile methods aside and considering software development in more general level, more Standardization practices on software development processes were identified.

> “Development process model includes partially these.” - Respondent C39

> “Processes are unified and responsibilities are defined.” - Respondent C30

This suggests that software development is done with practices that are standardized on company-level. These unified development processes could benefit especially the aspects of reliability and reproducibility. Based on these remarks on Standardization, empirical conclusion EC13 is formed.

> EC13: Software development practices (including agile methods) are used for realizing Technical robustness and safety.

As already stated in EC1, practice of Ethics and lawfulness by design did not emerge from the sample when coding the data.

Based on empirical conclusions on Technical robustness and safety, it can be said that Technical robustness and safety are mainly being realized with practices of Standardization and Testing and validation. Based on this, primary empirical conclusion PEC3 for Technical robustness and safety is proposed.

> PEC3: Technical robustness and safety are currently realized with standard software development processes (incl. agile methods) and testing and validation practices (e.g. code reviews).

As conclusion, six new empirical conclusions and one primary empirical conclusion considering Technical robustness and safety were proposed. These are presented in TABLE 15.

TABLE 15 Empirical conclusions for Technical robustness and safety

Identifier	**Empirical conclusion**

EC8	Architectures are used in some extent for realization of Technical robustness and safety.
EC9	Quality of Service is primarily measured from the perspective of project management.
EC10	Traditional software testing methods and validation by code reviews are common practices used for ensuring Technical robustness and safety.
EC11	Explanation methods can be used to validate Technical robustness and safety, but they are not common practice.
EC12	Certified software development methods are used for realization of Technical robustness and safety.
EC13	Software development practices (including agile methods) are used for realizing Technical robustness and safety.
PEC3	Technical robustness and safety are currently realized with standard software development processes (incl. agile methods) and testing and validation practices (e.g. code reviews).

6.4 Privacy and data governance

According to European Commission (2019), organizations that employ AI systems should ensure privacy, protect system's data from tampering, assure its quality and confine access only to appropriate persons and circumstances.

Regulation is expected to provide Privacy and data governance. In recent years, several regulations for data governance have been issued. General Data Protection Regulation (GDPR) issued by European Union and California Consumer Privacy Act (CCPA) issued by California State Legislature are examples of such regulation. Within the survey, one company mentioned following GDPR as one of their main practices for achieving trustworthiness.

> "All employees are trained to follow organizational policies. Company also follows GDPR strictly." - Respondent C31

In general, following regulations for achieving trustworthiness was mentioned two more times. Also, as proposed in EC2, following regulations is something that must be done to achieve trustworthiness, thus suggesting that the existing regulation is acknowledged within the industry. This establishes empirical conclusion EC14.

> EC14: Regulation for realizing Privacy and data governance already exist and they are acknowledged within the industry.

Trustworthy architectures can be employed to ensure privacy and provide capabilities for good data governance. Architectural solutions should be adapted to the required constraints of the AI system's target environment in order to ensure appropriate use of data. Respondents that discussed architectures did highlight their effort for choosing proper architectural solutions.

> “We always try to select long term architectural solutions” - Respondent C39

> “We focus on finding the right solution, not the easiest.” - Respondent C17

Still, these statements cover architectural solutions on general level, making it hard to draw detailed conclusions on architectures for realizing Privacy and data governance. This suggests that architectures exist for taking realization of Privacy and data governance further, but the means for doing so are not highlighted. These findings are incorporated in empirical conclusion EC15.

> EC15: Trustworthy architectures for realizing Privacy and data governance exist but the means for their realization are not highlighted.

By utilizing Codes of conduct, companies are expected to provide clear outlook on their intentions and values, maintain guidelines to ensure corporate responsibility and develop policies striving towards trustworthy AI (European Commission, 2019). Six different practices of Codes of conduct were identified inductively within the data: company policies, contracts, decision-making practices, documentation, operational guidelines and preparation.

Documentation was the most highlighted practice within the data with eight occurrences. However, documentation practices were mainly related to documentation of software development and documenting corporate intentions was not mentioned in the responses.

Contracts were also a common Codes of conduct practice to occur within the data. They were mainly used as a practice for establishing responsibilities with different parties involved in the development of the system.

> “Responsibilities are covered by contract.” - Respondent C1

Contracts did not discuss data governance in particular and they were mainly used to ensure responsibilities between software vendor and customer. Still, there were some responses suggesting that contracts can be used for ensuring privacy and proper data governance.

> “The written contracts with the end customers directly specifies their rights and our responsibilities.” - Respondent C9

> “The responsible and accountable parties for each development objects are defined are defined in the beginning of the project and also before signing the contract of the development work to be done” - Respondent C18

These excerpts suggest that contracts can be used to set ground rules for governance of data. As the regulation for Privacy and data governance already exists, contracts may be useful practice for defining the responsibilities required by the law.

Operational guidelines that provide clear course of action within the company formed another Codes of conduct related practice that was identified within the data.

> “Operational guidelines and rules that are monitored by audits.” - Respondent C19

> “We use process and tools and those are transparent to all in company. Cost are reported to owner and follower” -Respondent C30

These kinds of guidelines that set boundaries to day-to-day operations within the companies could facilitate different mechanisms for oversight, data management and privacy. One respondent especially mentioned that the guidelines are expected to ensure that customers are appreciated.

> “The instructions state that customers should be appreciated, and transparency is highlighted.” - Respondent C33

The concept of appreciation can be expected to cover the respect for customer privacy and appropriate handling of personal data. Another respondent also agreed that the guidelines exist to ensure trustworthiness of the company.

> “Of course. As a large company, we must keep everything above in order. [COMPANY NAME] management foundation sets rules and guidelines for everything we do so that we are trustworthy partner” - Respondent C19

This suggests that operational guidelines exist, and they are used for realization of some aspects of Privacy and data governance.

Closely linked to operational guidelines, Company policies were also suggested to have impact on trustworthiness. In these cases, the policies were company-wide and determined on the upper level.

> “All employees are trained to follow organizational policies.” - Respondent C31

With the principles provided by such policies, companies expected to improve their relationship with various stakeholders. This could also have impact on Data governance and privacy.

> “We are a premium product provider, which puts quality on high priority. Large part of our policies, processes and guidance is there for ensuring our customers and other stakeholders can put their trust on us.” - Respondent C37

Finally, one company especially mentioned that their Company policies take the requirements of handling personal data into account. This suggest that outspoken efforts towards realization of Privacy and data governance exist in the industry.

> “We handle personal information and demand high information security” - Respondent C7

Preparation for future was the least emergent code grouped under Codes of conduct. Trustworthiness was considered as something that companies should prepare for, meaning that aspects of trustworthiness should be taken into account in advance. This suggests that companies should be prepared as new problems and questions on trustworthiness will emerge when developing AI systems.

> “These factors have now been added eg to our developmental process, because it has been seen that it is an advantage to increase our customers and our own knowledge about these, especially when facing new challenges with the near future AI technology. It is good to prepare in advance, not after something has happened.” - Respondent C29

Realization of Data privacy and governance requires various mechanisms to ensure trustworthiness in different stages of AI system’s lifecycle. Implementation of these mechanisms requires preparation for incidents occurring in any phase of this lifecycle.

In conclusion for Codes of conduct, the data suggests that various Codes of conducts are widely employed within the industry. From these practices, especially contracts, operational guidelines, company policies and preparation are contributing towards realization of European Commission’s definition of Privacy and data governance. This forms the conclusion EC16.

> EC16: Codes of conduct (e.g. contracts, operational guidelines, company policies and preparation) are widely contributing towards realization of Privacy and data governance.

Testing and validation are also expected to ensure Privacy and data governance, for example by validating the quality and integrity of system’s data. As previously stated, testing and validation practices do appear in the data.

> “Development work is always done by testing practical examples.” - Respondent C20

One respondent tells that practical examples are used in development phase to ensure trustworthy operation. This could contribute to realization of Privacy and data governance by assuring that data used in development phase is correspondent with the real-life use case of the system. Still, there was no explicit reference of testing practices directly related Privacy and data governance, thus empirical conclusion EC17 is formed.

> EC17: Prevailing Testing and validation practices do not explicitly consider the realization of Privacy and data governance.

Accountability vie governance systems can contribute to Privacy and data governance. Appointing Data Privacy Officer is one example of such practice that is directly related to Privacy and data governance. In the data, there were two inductive codes appointed under Accountability via governance systems: accountability guidelines and defined responsibilities. Already stated in conclusion EC4, these systems mainly consider responsibilities between software vendor and their client.

> “I will take e.g. responsibility of the code I write and what it does.” - Respondent C27

> “We are responsible for every possible mistakes made during development process” - Respondent C4

> “The responsible and accountable parties for each development objects are defined are defined in the beginning of the project and also before signing the contract of the development work to be done.” - Respondent C18

These responses suggest that responsibility is appointed to different parties during the software development. The responsibilities in question should also include aspects of Privacy and data governance, as they are being reinforced by regulation mentioned in EC14. Yet, no explicit frameworks considering e.g. data privacy or accountability on data governance was not brought into attention. From these vague premises, emerging suggestion is that aspects of data privacy can be included in accountability via governance systems. This leads to drawing empirical conclusion EC18.

> EC18: Governance systems for accountability are used for defining responsibilities in software development and data privacy should be included in these.

Diverse and inclusive design teams were also considered to contribute to realization of Privacy and data governance, for example by preventing use of biased data. However, as stated in EC1, there was no practices identified as Diverse and inclusive design teams within the data. The same conclusion applies to the expected contributions of Ethics and lawfulness by design, which did not occur within the data.

In total, there were five empirical conclusions for Privacy and data governance. Regulations was explicitly considering the requirement of Privacy and data governance and other practices were reinforcing its impact. Based on these remarks, primary empirical conclusion PEC4 was formed.

> PEC4: Privacy and data governance are explicitly being realized by Regulations while other practices that could contribute to data privacy exist but do not explicitly consider its aspects.

Primary empirical conclusion along with empirical conclusions for Privacy and data governance are presented in TABLE 16.

TABLE 16 Empirical conclusions for Privacy and data governance

Identifier	Empirical conclusion
EC14	Regulation for realizing Privacy and data governance already exist and they are acknowledged within the industry.
EC15	Trustworthy architectures for realizing Privacy and data governance exist but the means for their realization are not highlighted.
EC16	Codes of conduct (e.g. contracts, operational guidelines, company policies and preparation) are widely contributing towards realization of Privacy and data governance.
EC17	Prevailing Testing and validation practices do not explicitly consider the realization of Privacy and data governance.
EC18	Governance systems for accountability are used for defining responsibilities in software development and data privacy should be included in these.
PEC4	Privacy and data governance are explicitly being realized by Regulations while other practices that could contribute to data privacy exist but do not explicitly consider its aspects.

6.5 Transparency

AI system's transparency should be considered from different points of view. These include the data used within the systems, system's behavior and the business models employed when using it. (European Commission, 2019.)

Explanation methods should provide transparency on underlying mechanisms of the AI system. As stated earlier, there were two identified practices belonging to Explanation methods: Reviewable models and Software audit trail. Both of these practices were identified in only one company.

> "Software is designed to maintain an audit trail log of all actions made by or with it, analyzable afterwards in some detail but of course not everything. Machine learning models can be inspected afterwards by feeding in different data and analyzing the output." -Respondent C14

Both of these practices are contributing towards making the AI system's outcomes more understandable. Audit trail logging helps to assess how the system operated in given conditions and reviewable models facilitates better oversight on underlying mechanisms of the AI system. From this premise, it can be said that Explanation methods contribute directly to realization of Transparency, but they are not in wide use yet. This is noted in empirical conclusion EC19.

> EC19: Explanation methods contribute directly to the realization of Transparency, but they are not in wide use.

Trustworthy architectures were also expected to contribute towards Transparency. Architectural solutions and patterns can be used for enforcing traceability and explainability of the systems operation. There were two companies that mentioned architectures as part of their pursues towards realization of trustworthiness. However, there were no clear depiction of chosen architectures, thus leaving their contribution towards realization of Transparency vague. Based on these remarks, the empirical conclusion EC20 is drawn.

> EC20: Trustworthy architectures are employed but their impact to Transparency remains vague.

Accountability via governance systems are contributing towards trustworthiness by defining responsibilities. One respondent told that developers take responsibility for the parts that they have programmed.

> "Each developer is responsible for the delivered feature. The features can easily be tracked to the individuals involved in the implementation." - Respondent C27

> "I will take e.g. responsibility of the code I write and what it does." - Respondent C27

With defined responsibilities, it should possible to trace how the chosen features and models were implemented within the AI system. Such traceability was brought into attention by two more respondents.

> "The responsible and accountable parties for each development objects are defined are defined in the beginning of the project and also before signing the contract of the development work to be done. Yes, changes are tracked back to person." - Respondent C18

> "The goal is that we would have clear decision making practices, documentation practices and well defined organisation and roles." - Respondent C35

This suggests that accountability frameworks are contributing towards Transparency by improving traceability of the system's operation. This forms the basis for the conclusion EC21.

> EC21: Accountability via governance systems improve traceability of system's operation, thus contributing towards realization of Transparency.

Stakeholder participation and social dialogue were highlighted by seven companies within the data. There were two inductive codes grouped under Stakeholder participation and social dialogue: customer involvement and informed customer. From these practices, customer involvement means that customers are taking part in development process, while informed customer refers to lesser participation of customers. Excerpts for customer involvement can be seen below.

> “Changes are accepted by customers.” - Respondent C18

> “Mutual understanding and verbal contract” - Respondent C12

> “As I am in close cooperation with the customer, it is my duty and responsibility to take care that the factions state in [COMPANY NAME] processes are reviewed and discussed with the customer.” - Respondent C29

These cases show that the customers were expected to provide their input for the development process by recurring two-way communication, suggesting that customers are actively taking part in software development. Customers participation was also considered important among software vendors.

> “We are developing software together with our customers and their feedback is very important when it comes to software development decisions.” - Respondent C1

There were also cases, where the activity of customer was not emphasized to same extent: Some respondents told about practices that required less involvement from their customers. This was coded as informed customer.

> “For customers it shows when we have meetings where we discuss about the project.” - Respondent C6

> “People want, that the customers know about the progress, and what is actually done.” - Respondent C33

> “Our aim is to be very transparent for the customer. Our project aim to follow agile methods and as a part of them, the processes like planning sprints and reviewing the developed solutions are very transparent to the customer.” - Respondent C26

In these cases, the relationship with customers was mostly taken care with one-way communication. Responses suggest that in these cases customers were only aware of software development without their own contribution to development process. While the participation of customer was not as extensive as in customer participation, customer was still involved in the development process. Both of these practices, customer participation and informed customer, enhanced the communication between software developers and their customers and provided mechanisms for informing the user. However, wider social dialogue with different stakeholders was not identified to data. This suggests that practices of Stakeholder participation and social dialogue exist and already contribute to realization of Transparency towards customers, but social dialogue for achieving wider transparency is not highlighted. This forms the conclusion EC22.

> EC22: Transparency is already being realized by practices of Stakeholder participation and social dialogue from the customer’s point of view, but social dialogue for achieving wider transparency is not highlighted.

Certification by independent organizations can contribute to Transparency. There were two different Certification practices identified within the data: audits and certification organizations. Independent audits may help people with no expertise on AI systems to assess trustworthiness.

> “Operational guidelines and rules that are monitored by audits.” - Respondent C19

In this case, audits were used to ensure that operational guidelines were followed. This should contribute towards realization of Transparency by raising traceability of company’s operations. Certification organizations were also used for achieving trustworthiness, by employing certified agile software development methods. They are contributing towards realization of Transparency by providing mechanisms for communication.

> “SAFe model offers policies for things that mentioned above.” - Respondent C28

Judging by its occurrences, Certification is not common practice for achieving trustworthiness. While the widespread recognition is still missing, they still have mentionable impact on realization of Transparency in the companies that mentioned them. This is forming the empirical conclusion EC23.

> EC23: Existing Certification practices are contributing towards realization of Transparency, but they are not widely considered.

In addition to certified development methods, Standardization of software development methods could be contributing towards realization of Transparency. Standard software development (including agile methods) were mentioned in multiple cases within the study. Especially principles of agile software development can be seen as contributing practice for realization of Transparency, certified or not. In total, there were three occurrences of agile software development methods within the data.

> “Our aim is to be very transparent for the customer. Our project aim to follow agile methods and as a part of them, the processes like planning sprints and reviewing the developed solutions are very transparent to the customer. Also predictability is high because of these operations” - Respondent C26

This respondent explicitly says that transparency is being pursued by agile methods, emphasizing that it makes the development process transparent for their customers. Besides agile methods, standardized software development methods were also being employed as the other Standardization practice identified. These practices were depicted on such general level, that it is hard to assess their impact on Transparency. This comprises the empirical conclusion EC24.

> EC24: Standardized agile development methods contribute towards realization of Transparency from the customers point of view.

Codes of conduct are widely used within the industry, as stated already in empirical conclusion EC16. The most common Codes of conduct practice was Documentation, that could contribute towards Transparency significantly.

> “Transparecy is taken into account internally through documentation but is not really considered from the other view points.” - Respondent C6

> “In my decision-making it is very important to be transparent, why certain content of function promotes our cause. I am expected to report it very precisely - it is the organization policy. The definition of predictability - if I understood correctly, doesnt apply so much. We create new and innovative solutions but of course certain factors are universal.” -Respondent C25

By documenting software development, companies could improve all three aspects of transparency: traceability, explainability and communication.

> EC25: Codes of conduct are widely contributing towards realization of Transparency.

Education and awareness of ethical mindset was identified in three cases within the data. Promoting company’s role as a change agent was one of these practices. These cases suggested that companies were on a journey of building a better future with trying new solutions.

> “One of our core values is integrity and the mission is "observation for a better world". These reflect to training an general mentality of the entire company and corporate responsibility - not only software development.” - Respondent C37

> “In [COMPANY NAME] everyone is encouraged to try new things and when doing so even failures are option. We may always learn from failures and make things and processes better, which supports predictability, accountability and responsibility nicely.” - Respondent C2

Operations of these companies were founded on core values that were considered crucial for trustworthiness. They seem to be promoting a change in the world. Whether Transparency itself is considered to be included in these values, is still left open for discussion. However, ethical questions are considered in these companies. In addition to promotion of change agency, the practice Education and training were also present in one company.

> “[COMPANY NAME] offers basic knowledge on these factors as mandatory trainings for all, and if one likes, it is possible to add your knowledge in other trainings.” -Respondent C29

In this case, the company in question was providing some training on trustworthiness for its employees. These trainings were mandatory, which suggests that ethical questions are considered important. Despite the lack of wide use for ethical awareness and mindset, they are already in use in few companies. In these companies, the ethical mindset is laid on foundation of strong core values that strive for change. This forms empirical conclusion EC26.

> EC26: Education and awareness of ethical mindset is practiced in only few companies, where they are based on strong change-oriented core values.

Testing and validation were also expected to contribute transparency to AI system's operation. One respondent suggested that transparency is achieved by peer surveillance. This means that code is reviewed by other developer and nothing is executed without up-front inspection of another employee.

> "Transparency is achieved through code reviews, so no code gets to the production if no other developer has given it a review and approved it first." – Respondent C27

Another respondent supported this suggestion, elaborating that above mentioned practice of documentation also enhances the transparency achieved from code reviews. In this case, the person who had implemented features was supposed to be declared within the documentation. This is supposedly providing transparency on the development phase of AI system.

> "Full documentation and tools where we can address who has made what, code reviews between developers." – Respondent C34

Based on this, it can be said that Testing and validation benefit the realization of Transparency via code reviews that improve Transparency on how the system was built. This is the basis for empirical conclusion EC27.

> EC27: Testing and validation contribute to Transparency by increasing the visibility of how the system was built via code reviews.

Regulation practices should also contribute towards realization of Transparency. There are existing regulations that call for companies to declare their intentions on the personal data they are collecting. General Data Protection Regulation (GDPR) was one example of such and it laid the foundation for empirical conclusion EC14.

> "Company also follows GDPR strictly." - Respondent C31

This suggests that Regulations that require companies to inform the end users about how their data is being used and stored. This can be seen as direct contribution towards realization of Transparency. Thus, empirical conclusion EC28 is formed.

> EC28: Regulations contribute directly towards realization of data-related aspects of Transparency.

Ethics and lawfulness by design was also expected to contribute towards realization of Transparency, but as already stated in EC1 there was no such practices identified within the data. Based on the nine empirical conclusions for transparency, a primary empirical conclusion was formed.

PEC5: Transparency is primarily perceived as a matter between software provider and their customer on the development process and it is already realized with various practices, with 10 practices out of 11 having at least a partial contribution to its realization.

Empirical conclusions for Transparency are depicted in TABLE 17 along with the primary empirical conclusion considering realization of Transparency.

TABLE 17 Empirical conclusions for Transparency

Identifier	Empirical conclusion
EC19	Explanation methods contribute directly to the realization of Transparency, but they are not in wide use.
EC20	Trustworthy architectures are employed but their impact to Transparency remains vague.
EC21	Accountability via governance systems improve traceability of system's operation, thus contributing towards realization of Transparency.
EC22	Transparency is already being realized by practices of Stakeholder participation and social dialogue from the customer's point of view, but social dialogue for achieving wider transparency is not highlighted.
EC23	Existing Certification practices are contributing towards realization of Transparency, but they are not widely considered.
EC24	Standardized agile development methods contribute towards realization of Transparency from the customers point of view.
EC25	Codes of conduct are widely contributing towards realization of Transparency.
EC26	Education and awareness of ethical mindset is practiced in only few companies, where they are based on strong change-oriented core values.
EC27	Testing and validation contribute to Transparency by increasing the visibility of how the system was built via code reviews.
EC28	Regulations contribute directly towards realization of data-related aspects of Transparency.
PEC5	Transparency is primarily perceived as a matter between software provider and their customer on the development process and it is already realized with various practices, with 10 practices out of 11 having at least a partial contribution to its realization.

6.6 Accountability

Accountability and responsibility should be carefully considered when using and developing AI systems. These aspects include auditability, minimizing and reporting negative impact, documenting trade-offs and ability to redress. (European Commission, 2019.)

Accountability via governance systems is manifested within the industry by defining responsibilities and accountability guidelines. Both of these practices are expected to contribute towards Accountability of the system. Responsible parties were defined in multiple companies.

> "The written contracts with the end customers directly specifies their rights and our responsibilities." - Respondent C9

> "The responsible and accountable parties for each development objects are defined are defined in the beginning of the project and also before signing the contract of the development work to be done." - Respondent C18

These excerpts suggest that distribution of responsibilities and accountability are explicitly defined beforehand. This explicit responsibility could be beneficial in situations where system's operation leads to adversarial results. One respondent told that they have the responsibility for every mistake that can be traced back to the development phase, thus making them accountable on the system's operation.

> "We are responsible for every possible mistakes made during development process" - Respondent C4

These practices suggest that governance systems are used in the development phase for determining the accountable parties. The realization of Accountability in later phases after development however is left unclear.

> EC29: Accountability is being realized with governance systems in the development phase of AI systems.

Trustworthy architectures can be used to implement mechanisms that for example provide accountability in trade-off situations. As stated in previous chapters, companies that discussed architectures mentioned that they are looking for appropriate architectural solutions for their systems. These architectures should include mechanisms for realization of Accountability. However, as stated in conclusions EC7, EC8, EC15 and EC20, detailed features of these architectures remain unknown, making it hard to determine whether architectures take Accountability into consideration. This forms empirical conclusion EC30.

> EC30: Companies seek for Trustworthy architectures but their current impact on realization of Accountability remains vague.

Explanation methods provide way for ensuring traceability of AI system's operation as is required for realization of Accountability. There was one company within the study practicing Explanation methods. These methods were coded as Software audit trail and Reviewable models.

> "Software is designed to maintain an audit trail log of all actions made by or with it, analyzable afterwards in some detail but of course not everything. Machine learning models can be inspected afterwards by feeding in different data and analyzing the output." - Respondent C14

As seen in this quotation, Explanation methods are providing auditable log of actions and opportunity to analyze the possible outcomes of AI system. This proposes that existing explanation methods contribute to realization of Accountability. As already stated in conclusions EC11 and EC19, Explanation methods are not in wide use. Empirical conclusion EC31 is drawn from these remarks.

> EC31: Explanation methods are contributing towards realization of Accountability by their practitioners, but they are not in wide use.

Stakeholder participation and social dialogue can provide external oversight on the AI system and thus contribute towards realization of Accountability. Seven companies within the data told they practice Stakeholder participation. Four cases included intense involvement of customer while three other companies just kept their customer informed of the development process.

> "We are developing software together with our customers and their feedback is very important when it comes to software development decisions." - Respondent C1

> "People want, that the customers know about the progress, and what is actually done." - Respondent C33

Intense involvement of customer directly contributes to realization of Accountability with external guidance and oversight. Informing the customer can also work as a mechanism for external oversight. Thus, empirical conclusion EC32 is formed.

> EC32: Stakeholder participation and social dialogue with customers already provide mechanisms for realization of Accountability.

Education and awareness of ethical mindset was practiced by three respondents. As stated in EC26, the ethical mindset of these companies was founded on change-oriented core values in few companies. However, no values that directly contribute to Accountability was present.

> “One of our core values is integrity and the mission is "observation for a better world". These reflect to training an general mentality of the entire company and corporate responsibility - not only software development.” - Respondent C37

> “In [COMPANY NAME] everyone is encouraged to try new things and when doing so even failures are option. We may always learn from failures and make things and processes better, which supports predictability, accountability and responsibility nicely.” - Respondent C2

It is still unclear whether these companies consider change as something that can be achieved by any means. Realization of Accountability requires minimization of negative impact, which could be contradictory with strong advocacy for change. Yet, this does not mean that minimizing negative impact is ignored. These remarks form the empirical conclusion EC33.

> EC33: Education and awareness of ethical mindset is practiced in only few companies, where they are based on strong change-oriented core values, that could be contradictory with mechanisms required for realization of Accountability.

Codes of conduct included various common practices within examined companies. Company policies, contracts, decision-making practices, documentation, operational guidelines and preparation were all inductively identified within the data as Codes of conduct that could contribute to realization of Accountability.

> ”We are a premium product provider, which puts quality on high priority. Large part of our policies, processes and guidance is there for ensuring our customers and other stakeholders can put their trust on us.” - Respondent C37

In example above, company policies are used for ensuring trustworthiness from the point of view of customers and other stakeholders. Contracts were also used for maintaining accountability and responsibility between software vendors and their customers.

> “Responsibilities are covered by contract.” - Respondent C1

> “Mutual understanding and verbal contract” - Respondent C12

In one case, contracts were used especially for enforcing governance frameworks’ influence on Accountability.

> “The written contracts with the end customers directly specifies their rights and our responsibilities.” - Respondent C9

Decision-making practices in conjunction with Documentation were also used for maintaining Accountability. Clear mechanisms for decision-making provide auditability to companies’ operations by improving traceability of decisions.

> “In my decision-making it is very important to be transparent, why certain content of function promotes our cause. I am expected to report it very precisely - it is the organization policy. The definition of predictability - if I understood correctly, doesnt apply so much. We create new and innovative solutions but of course certain factors are universal.” - Respondent C25

> ”The goal is that we would have clear decision making practices, documentation practices and well defined organisation and roles. There is still some work on this but direction is clear.” - Respondent C35

In overall, documentation practices were widely used in the examined companies. These companies kept record on the development process, including what had been done and who carried out the implementation.

> ”Transparecy is taken into account internally through documentation but is not really considered from the other view points.” - Respondent C6

> ”Full documentation and tools where we can address who has made what” - Respondent C34

> “By using and documenting all SW work to different databases and softwares.” - Respondent C17

These were also supplemented by operational guidelines that provide clear instructions on how to conduct operations within the company.

> “Mostly by using good methods in workplanning, recoursing and reporting in SW development.” - Respondent C17

> “The instructions state that customers should be appreciated, and transparency is highlighted.” - Respondent C33

One company noted that they should continuously prepare for the unforeseen questions and challenges that could occur when developing AI systems. This code of conduct sets trustworthiness subordinate to continuous assessment.

> “These factors have now been added eg to our developmental process, because it has been seen that it is an advantage to increase our customers and our own knowledge about these, especially when facing new challenges with the near future AI technology. It is good to prepare in advance, not after something has happened.” - Respondent C29

These quotations suggest that Codes of conduct are extensively used for realizing Accountability of AI systems. Especially the aspect of auditability is taken into consideration when practicing Codes of conduct. These remarks form the empirical conclusion EC34.

> EC34: Codes of conduct are contributing towards realization of Accountability by providing mechanisms for auditability, traceability and documentation.

Standardization was practiced by various companies and most of these practices were unified software development practices that were standardized on the company level.

> “Processes are unified and responsibilities are defined.” - Respondent C30

> “Development process model includes partially these.” - Respondent C39

> “It's the reason the processes exist and are document and trained.” - Respondent C37

Some companies described their standardized processes in more detail. These respondents told that they commit to agile software development for achieving trustworthiness.

> ”Our project aim to follow agile methods and as a part of them, the processes like planning sprints and reviewing the developed solutions are very transparent to the customer.” - Respondent C26

> “Our aim is to be very transparent for the customer. Our project aim to follow agile methods and as a part of them, the processes like planning sprints and reviewing the developed solutions are very transparent to the customer.” - Respondent C26

Standardized processes are extensively used within industry. These processes are expected to provide various mechanisms that could also contribute to realization of Accountability.

> EC35: Standardization practices are extensively used but their contribution towards realization of Accountability remains vague.

Certification was used for reinforcing standardized agile methods. As stated in previous chapters (e.g. conclusion EC12), one company expressed that they use certified agile method in order to achieve trustworthiness. This suggests that Certification provides further auditability and oversight on the existing standardization practices. The other certification practice within the data, audits, was also mentioned by one company.

> “Operational guidelines and rules that are monitored by audits.” - Respondent C19

Auditability being one of the aspects of Accountability, this certification practice should benefit realization of the requirement directly. This suggest that Certification contributes to Accountability in the companies that practice it.

> EC36: Certification practices contribute directly to realization of Accountability, but they are not in wide use.

Regulation was expected to contribute towards realization of Accountability. As previously noted in conclusions EC14 and EC28, current regulations consider mainly the use of data. New legislation that considers new accountability and liability issues caused by Artificial Intelligence was not found. Conclusion EC29 stated that accountability is achieved with contracts. These contracts expectedly get their power from legislation, which makes them binding. Nevertheless, regulation for emerging Accountability issues was not found, leaving the contribution partial. This makes the foundation for empirical conclusion EC37.

> EC37: Regulation for emerging AI related issues for Accountability do not exist, but existing Regulations have partial contribution towards its realization.

Ethics and lawfulness by design was also expected to contribute towards realization of Accountability but, as previously stated in EC1, this practice was not identified in the data.

There were eight empirical conclusions considering Accountability. These conclusions suggested that accountability is perceived to be a matter between the software provider of the AI system and their customer. Based on these conclusions, primary empirical conclusion PEC6 was formed.

> PEC6: Accountability is considered to be a matter between software provider and their customer, and it is already realized with major contributions by Codes of conduct, Governance systems and Stakeholder participation and this is supported with partial contributions by other practices.

Empirical conclusions and primary empirical conclusion for Accountability are depicted in TABLE 18.

TABLE 18 Empirical conclusions for Accountability

Identifier	Empirical conclusion
EC29	Accountability is being realized with governance systems in the development phase of AI systems.
EC30	Companies seek for Trustworthy architectures but their current impact on realization of Accountability remains vague.
EC31	Explanation methods are contributing towards realization of Accountability by their practitioners, but they are not in wide use.
EC32	Stakeholder participation and social dialogue with customers already provide mechanisms for realization of Accountability.
EC33	Education and awareness of ethical mindset is practiced in only few companies, where they are based on strong change-oriented core values, that could be contradictory with mechanisms required for realization of Accountability.
EC34	Codes of conduct are contributing towards realization of Accountability by providing mechanisms for auditability, traceability and documentation.
EC35	Standardization practices are extensively used but their contribution towards realization of Accountability remains vague.
EC36	Certification practices contribute directly to realization of

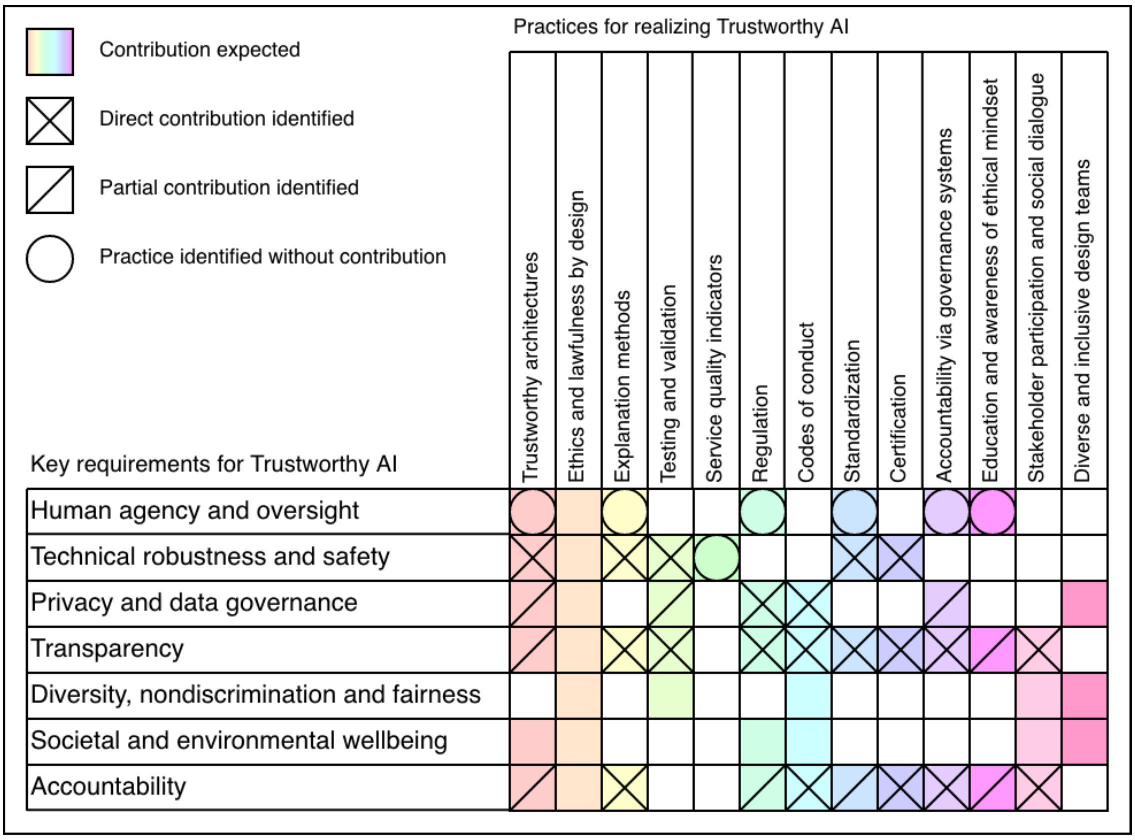
Accountability, but they are not in wide use.

EC37 Regulation for emerging AI related issues for Accountability do not exist, but existing Regulations have partial contribution towards its realization.

PEC6 Accountability is considered to be a matter between software provider and their customer, and it is already realized with major contributions by Codes of conduct, Governance systems and Stakeholder participation and this is supported with partial contributions by other practices.

6.7 Summary

This chapter included the analysis of empirical data and its outcome: the six primary empirical conclusions. Contribution towards realization of each key requirement was assessed for every practice identified within the data. These 34 empirical conclusions are the remarks that form the primary empirical conclusions. The contributions by each practice are depicted in FIGURE 8, where Checked square means direct contribution by the practice, square with one line refers to vague or minor contribution and square with circle means that practice was present within the data but no contribution towards realization of the practice was identified. Colored square with no insignia means that practice was expected to contribute towards the requirement, but it was not identified within the data.



▨ Contribution expected

☒ Direct contribution identified

⧄ Partial contribution identified

○ Practice identified without contribution

Practices for realizing Trustworthy AI

Key requirements for Trustworthy AI	Trustworthy architectures	Ethics and lawfulness by design	Explanation methods	Testing and validation	Service quality indicators	Regulation	Codes of conduct	Standardization	Certification	Accountability via governance systems	Education and awareness of ethical mindset	Stakeholder participation and social dialogue	Diverse and inclusive design teams
Human agency and oversight	○	▨	○			○		○		○	○		
Technical robustness and safety	☒	▨	☒	☒	○			☒	☒				
Privacy and data governance	⧄	▨		⧄		☒	☒			⧄			▨
Transparency	⧄	▨	☒	☒		☒	☒	☒	☒	☒	⧄	☒	
Diversity, nondiscrimination and fairness		▨		▨			▨					▨	▨
Societal and environmental wellbeing	▨	▨				▨	▨					▨	▨
Accountability	⧄	▨	☒			⧄	☒	⧄	☒	☒	⧄	☒	

FIGURE 8 Contributions towards realization of trustworthy AI's requirements (comparison of primary conceptual conclusion and primary empirical conclusions).

The remarks presented on the figure above are formed into the last primary empirical conclusion of this study. This is presented as the conclusion PEC7.

PEC7: Two practices proposed by European Commission, Ethics and lawfulness by design and Diverse and inclusive design teams, were not brought into attention by companies developing AI. Other practices proposed by European Commission are already employed for achieving trustworthiness with contributions of various extent.

Outcome of this chapter are the seven primary empirical conclusions that are based on empirical evidence presented in this chapter. The conclusions are presented in TABLE 19.

TABLE 19 Primary empirical conclusions formed from the data.

Identifier	Empirical conclusion
PEC1	Requirements of Societal and environmental wellbeing and Diversity, nondiscrimination and fairness were not discussed within the data.
PEC2	Companies employ practices that could contribute towards the requirement of Human agency and oversight, but they do not

	consider the requirement particularly, leaving their contribution towards Human agency and oversight vague.
PEC3	Technical robustness and safety are currently realized with standard software development processes (incl. agile methods) and testing and validation practices (e.g. code reviews).
PEC4	Privacy and data governance are explicitly being realized by Regulations while other practices that could contribute to data privacy exist but do not explicitly consider its aspects.
PEC5	Transparency is primarily perceived as a matter between software provider and their customer on the development process and it is already realized with various practices, with 10 practices out of 11 having at least a partial contribution to its realization.
PEC6	Accountability is considered to be a matter between software provider and their customer, and it is already realized with major contributions by Codes of conduct, Governance systems and Stakeholder participation and this is supported with partial contributions by other practices.
PEC7	Two practices proposed by European Commission, Ethics and lawfulness by design and Diverse and inclusive design teams, were not brought into attention by companies developing AI. Other practices proposed by European Commission are already employed for achieving trustworthiness with contributions of various extent.

These primary empirical conclusions lay the foundation for discussion in the following chapter. To clarify these conclusions, context enriched PECs are presented in TABLE 20.

TABLE 20 Context-enriched primary empirical conclusions

Identifier	Context-enriched conclusion
PEC1	Societal and environmental wellbeing and Diversity, nondiscrimination and fairness were not brought up by companies developing Artificial Intelligence.
PEC2	Human agency and oversight over Artificial Intelligence are not widely considered by software providers of AI systems, but practices that could contribute to it are already employed.
PEC3	Current practices for development of technically robust and safe Artificial Intelligence follow the common methods for software development like code reviews and agile methods to mention some.
PEC4	Development of trustworthy Artificial Intelligence from the perspective of data privacy and governance is mainly realized by legislation regulating the use of data.
PEC5	Transparency between software providers of Artificial Intelligence and their customers is widely considered during the development process and multiple practices are used for maintaining traceability, communication and explainability, that form its aspects. Transparency towards wider public is not as much considered.
PEC6	Accountability concerns of Artificial Intelligence are currently mitigated by various codes of conduct, using governance frameworks

to define accountable parties and involving customers in the development process. Accountability of AI systems is still seen as matter between the software provider and their customers.

PEC7 Companies developing AI systems employ various practices for achieving their trustworthiness, but they are not putting emphasis on design methods that include ethics in the core processes and diverse design teams.

The context-enriched PECs extend the original primary empirical conclusions with information that make them understandable in their very self. They can be used when communicating the empirical findings in mediums where the context of the study is not available.

7 Discussion

This chapter connects the seven primary empirical conclusions that formed the outcome of the previous chapter to the theoretical foundation of this study.

7.1 Practical implications

European Commission (2019) states, that "AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom."

As stated in PEC1, requirements of Diversity, nondiscrimination and fairness and Societal and environmental wellbeing were not identified in the data. This suggests that industry is lagging behind in these aspects, as they were not even brought into attention. If Diversity, nondiscrimination and fairness and Societal and environmental were under company-wide consideration, expectedly it should reflect to employees' perception of trustworthiness. This implies that companies should put additional emphasis on ensuring that AI systems treat different stakeholders equally and act in manner that is sustainable in both societal and environmental point of view.

The requirement of Human agency and oversight is another projection of the above statement. Empirical conclusion PEC2 suggests that despite its importance, Human agency and oversight are not seen as fundamental aspects when striving for trustworthiness. This implies that companies should include aspects of Human agency and oversight into company-wide discussion when considering trustworthiness.

Regulations were perceived as major contributor for realization of Privacy and data governance. However, this suggest that currently they mainly cover data related issues such as respect for privacy, traceability of data and controlling access to personal information.

PEC7 suggested that practices of Ethics and lawfulness by design and Diverse design teams were not brought up by any company within the study.

This implies that companies should adopt such practices and make them more salient within their operations. Especially the inclusion of ethical norms into the design of AI systems should be promoted, as such practices were expected to contribute towards realization of all the requirements of trustworthy AI in the conceptual framework.

TABLE 21 Practical implications of the study

Empirical conclusion	Implication for practice
PEC1	Company-wide discussions should put additional emphasis on Diversity, nondiscrimination and fairness and Societal and environmental wellbeing.
PEC2	Aspects of Human agency and oversight should be included in company-wide discussion when considering trustworthiness.
PEC4, PEC5	Regulations should be imposed to realizing other aspects of trustworthy AI besides of merely data related issues. Global pursuits for regulation are needed.
PEC7	Ethics and lawfulness should be explicitly embedded in the design of AI systems by companies developing them.
PEC7	Companies should keep diversity in mind when establishing design teams to make sure that the wide spectrum of people affected by AI systems is represented among the people developing them.

7.2 Theoretical contributions

Goal of this study was to assess realization of the key requirements for trustworthy AI within software industry. Empirical evidence suggests that maturity for realization of the seven key requirements vary. Thus, the realization is in more advanced phase for some of them than the others.

TABLE 22 Theoretical contributions of the study

Identifier	Empirical conclusion	Relation to existing research
PEC1	Requirements of Societal and environmental wellbeing and Diversity, nondiscrimination and fairness were not discussed within the data.	Contradicting, environmental and societal issues were considered as major concerns of AI (Kaplan & Haenlein, 2020).
PEC2	Companies employ practices that could contribute towards the requirement of Human agency and oversight, but they do not consider	Contradicting, ensuring human authority was discussed in multiple studies (Kaplan & Haenlein, 2020; von Krogh, 2018).

	the requirement particularly, leaving their contribution towards Human agency and oversight vague.	
PEC3	Technical robustness and safety are currently realized with standard software development processes (incl. agile methods) and testing and validation practices (e.g. code reviews).	Corresponding with previous research (Brock & von Wangenheim, 2019).
PEC4	Privacy and data governance are explicitly being realized by Regulations while other practices that could contribute to data privacy exist but do not explicitly consider its aspects.	Corresponding with previous studies (Haenlein & Kaplan, 2019).
PEC5	Transparency is primarily perceived as a matter between software provider and their customer on the development process and it is already realized with various practices, with 10 practices out of 11 having at least a partial contribution to its realization.	Contradicting, Transparency should encompass other stakeholders (e.g. people affected by AI supported decisions) than only the paying customer. (Holzinger et al., 2018; Kaplan & Haenlein, 2020; Rudin, 2019).
PEC6	Accountability is considered to be a matter between software provider and their customer, and it is already realized with major contributions by Codes of conduct, Governance systems and Stakeholder participation and this is supported with partial contributions by other practices.	Contradicting, Accountability of AI system's operation should cover other parties in addition to the vendor and customer (Kaplan, 2016; Kaplan & Haenlein, 2020; von Krogh, 2018).
PEC7	Two practices proposed by European Commission, Ethics and lawfulness by design and Diverse and inclusive design teams, were not brought into attention by companies developing AI. Other practices proposed by European Commission are already employed for achieving trustworthiness with contributions of various extent.	Novel, previous research on how trustworthiness is pursued by companies developing AI systems was not identified.

Societal and environmental wellbeing and Diversity, nondiscrimination and fairness were not discussed at all within the survey data. (PEC1) This contradicts with other research, as societal and environmental issues have been identified as major concerns for AI adoption. Kaplan and Haenlein (2020) suggested that AI systems can either reinforce or mitigate the effects of societal problems such as loneliness and inequality. They also proposed the same effect on AI system's impact on environmental burden of human economy. Against this background, it can be stated that both Societal and environmental wellbeing and Diversity, nondiscrimination and fairness are major concerns that should not be forgot.

Human agency and oversight were not taken into account in practices for achieving trustworthiness, as was suggested in PEC2. This contradicts with theoretical background, as there were multiple studies stating that ensuring human authority and autonomy over AI systems was major concern for adoption of AI. Delegation of tasks that previously required human reasoning to Artificial Intelligence results in reduction of human authority. This shift of authority from humans to AI could have unexpected consequences. (von Krogh, 2018.) Proper education could be useful in mitigating the risks for such shift on authority. Learning programming could help people understand how Artificial Intelligence works, thus forming better premises for their oversight In the end, Artificial Intelligence should be cooperating hand in hand with human beings. (Kaplan & Haenlein, 2020.) This suggests that aspects of Human agency and oversight should be carefully considered during development of AI systems.

From the point of view of Technical robustness and safety, agile software development methods are the major contributor towards trustworthiness (PEC3). Organizational agility and ability to adapt to change was also identified as a success factor for AI adoption by Brock and von Wangenheim (2019). Comparing the empirical conclusion and findings of Brock and von Wangenheim (2019) suggest that agility could contribute to other aspects of trustworthy AI instead of just contributing towards realization of Technical robustness and safety. Brock and von Wangenheim (2019) also propose that technological alignment is required to ensure integrity while the tasks and scope for AI adoption are getting more complex and wider. This kind of alignment was not identified as contributor towards trustworthiness within the empirical findings. Realizing Technical robustness in the sense of reliability and reproducibility could be further increased with creation of AI systems with interpretable and transparent underlying models (Rudin, 2019).

Regulation was identified as major contributor towards realization of Privacy and data governance (PEC4). This is partially corresponding with the remarks of Haenlein and Kaplan (2019) who noted that governments have chosen opposite paths on regulation of AI's operational environment, mentioning China and European Union as examples from different ends of this path. As the data consisted of companies affected by European Union's legislation, the findings of this study are bound to the context of EU. In addition to privacy, Kaplan and Haenlein (2020) also mention other fields that require new regulation from the perspective of AI. It could be demanded that AI systems (e.g. autonomous vehicles) collect and store data of their operations to ensure availability of objective information in case of adverse incidents. Such regulation already exists for aviation, as airplanes are required to record the operations that happen in the cockpit. Kaplan and Haenlein (2020) also suggest that regulation could be used to prevent consolidation of the AI industry. Nevertheless, there is still plenty of space to be regulated from Artificial Intelligence's perspective and Privacy and data governance can be considered as the first milestone of emerging regulations.

It was also found out that Transparency is perceived as a matter between the provider developing the AI system and their customer (PEC5). Customers were often involved with the development process and it was considered important. However, the finding is contradictory with previous literature, as transparency of AI systems should encompass other stakeholders too. Black box AI systems lower people's trust on them overall as their decisions may not be understandable (Holzinger et al., 2018). Open dialogue and transparent leadership are needed for mitigating change resistance. True or not, human employees might be afraid of becoming substituted with adoption of Artificial Intelligence. Raising awareness about AI's impact should help people to adjust to the resulting change. (Kaplan & Haenlein, 2020.) Rudin (2019) states that Transparency towards the end users and people who are affected by AI system's decisions is important when using AI for support in decision making. Kaplan and Haenlein (2020) also suggest that adoption of AI could increase transparency of the inner mechanisms and structures of companies by making their underlying biases more obvious. Even though these findings can be considered as contradictory from the perspective of wider public and even end users, Transparency is still widely considered within the industry when considering the practices that are contributing towards it.

Accountability and liability of AI systems actions was one of the disputable ethical concerns, that culminated with the question of who is being held responsible for actions of AI systems operating autonomously (Kaplan, 2016). AI systems may also engage in trade-off situations requiring ethical decisions, which could be even impossible for individual human beings. When compared to human physicians, Machine Leaning algorithms have already proved to be better in various risk assessment tasks, such as recognizing potential strokes in patients and determining their risk of renewal. But, AI systems' ability to make final decisions between variety of possibilities is still very limited and the ability to interpret and justify the given conditions still remains too "humane" for AI. This calls for clear definition of accountability and mechanisms for ensuring it from multiple perspectives, as the outcomes of AI systems eventually track back to the people developing, training and employing them. (von Krogh, 2018.) Kaplan and Haenlein (2020) state that AI systems are just following orders very precisely. Combination of such literal precision and human instructions, which can be vague from AI's perspective, makes Accountability one of the central issues of trustworthy AI. The empirical conclusion PEC6 suggests that these liability issues were examined only from the perspective of providers, who develop the software, and their paying customers. This remark is contradicting with the previous literature, which considers Accountability as a much extensive issue.

PEC7 provided an overview on how trustworthy AI development is currently being realized among the industry. This can be compared with the primary conceptual conclusion that was presented in FIGURE 7. There were two practices, Ethics and lawfulness by design and diverse and inclusive design teams, that were expected to contribute towards trustworthy AI but were not identified within the empirical data. Ethics and lawfulness by design was expected to contribute to realization of all of the seven requirements, pointing out a remarkable gap between the conceptual framework and empirical results. Lack of Diverse and inclusive design teams cause a similar gap for three requirements, which may not be as drastic but is still remarkable for the concerned requirements. This finding can be considered as novel one, since there was not any previous research on how trustworthiness is being pursued by companies.

This study used the guidelines for trustworthy AI defined by European Commission's High-Level Expert Group on Artificial Intelligence (2019) as theoretical framework. These guidelines provided holistic and comprehensive implications for ensuring trustworthiness of Artificial Intelligence from the perspectives of lawfulness, ethics and robustness. The guidelines were in accordance with the research, as they assessed the various concerns presented by Kaplan and Haenlein (2020), von Krogh (2018), Rudin (2019) and Brock and von Wangenheim (2019) to mention some.

8 Moving on

This chapter presents the final conclusions for the study. These conclusions include the answer to research questions, limitations of the study and future research opportunities.

8.1 Answer to research question

Aim of this study was to identify how companies developing AI are currently pursuing trustworthiness in their operations. To establish a clear conception of this, this study was trying to answer the following research question:

> How do companies currently pursue trustworthiness when developing Artificial Intelligence?

The question was answered by assessing the fulfillment of seven key requirements for trustworthy AI, which were defined by European Commission (2019) in their report titled "Ethics Guidelines for Trustworthy AI". The realization of each requirement was assessed by mapping the practices, which were expected to contribute towards the realization of trustworthiness by companies employing them.

Empirical findings of the study suggest that trustworthiness is mainly being pursued with contributions towards realization of Transparency, Accountability and Technical robustness and safety, as these requirements had multiple practices contributing towards their realization. Trustworthiness was also pursued by contributions towards Privacy and data governance, which was mainly realized by regulations of European Union within the sample. On the other hand, trustworthiness was not pursued with contributions towards realization of Societal and environmental wellbeing and Diversity, nondiscrimination and fairness.

8.2 Limitations of study

Survey was also conducted before the guidelines for trustworthy AI by European Commission (2019) were published. This means that operationalization of trustworthiness was not performed on the basis of theoretical framework. To inhibit limitations caused from this, integrated coding approach was used (Cruzes & Dybå, 2011). On the other hand, this mitigated the guidelines from constructing its own reality as it prevented the survey data from reflecting the guidelines itself, which is a common pitfall for qualitative research (Myers & Newman, 2007).

The employed structured interview did not leave any room for improvisation on the interviews. Due to this, there was no possibility to present elaborating questions for the respondents in order to attain an exhaustive description of the employed practices. This leads to the remark that the findings of this study do not present exhaustive description of all the practices employed for development of trustworthy AI. Rather, the findings provide insight on what practices have the foremost contribution towards realization of trustworthy AI and which are the requirements that benefit the most from these practices.

Sample consisting of respondents primarily from companies operating in Finland limits the generalizability of the conclusions. For example, all of the companies within the study were operating under the same regulations, limiting the ability to draw conclusions on companies that operate outside of European Union.

8.3 Future research opportunities

Three future research opportunities were identified. These are research on inclusiveness, nondiscrimination and fairness, developing predictable and understandable AI and AI literacy.

When discussing trustworthiness, inclusiveness, nondiscrimination and diversity were not thrown into relief by companies developing Artificial Intelligence. This calls for further research on the current state of these aspects within the industry. Both descriptive research for achieving better view on the current state of inclusiveness, nondiscrimination and diversity and research explaining why such aspects are being ignored are required.

Research to ensure predictable and understandable operations of AI are also required. Aspect of predictability was important for technical robustness and transparency of AI systems. Yet, developing black box systems without clear conception on the underlying model is considered somewhat acceptable when developing Artificial Intelligence for commercial purposes. For improving the realization of transparency and technical robustness and safety of AI systems, further research focusing on the development of AI that operates in comprehensible manner is required. After comprehensible and predictable AI models mature, their impact on trustworthiness could be studied too.

Adoption of AI could democratize knowledge by providing people capabilities that previously required vast amounts of capital. On the other hand, it could also widen the gap between the people having the capabilities to domesticate AI and people that do not have access to its benefits. This calls for research on so-called "AI literacy", that would assess both people's environmental factors that affect the use of AI (e.g. socio-economic status) and individual's own ability to successfully employ it (e.g. ability to use technology). The concept of AI literacy could be studied even further to assess its impact on perceived trustworthiness of Artificial Intelligence.

REFERENCES

Ahmed, S. H., Bouk, S. H., Kim, D., & Sarkar, M. (2016). Cyber-physical systems: Basics and fundamentals. *Cyber-Physical System Design with Sensor Networking Technologies*, 21–46.

Alam, K. M., & El Saddik, A. (2017). C2PS: A Digital Twin Architecture Reference Model for the Cloud-Based Cyber-Physical Systems. *IEEE Access, 5*, 2050–2062. https://doi.org/10.1109/ACCESS.2017.2657006

Alam, K. M., Sopena, A., & El Saddik, A. (2015). Design and Development of a Cloud Based Cyber-Physical Architecture for the Internet-of-Things. In *2015 IEEE International Symposium on Multimedia (ISM)* (pp. 459–464). https://doi.org/10.1109/ISM.2015.96

Alur, R. (2015). *Principles of cyber-physical systems*. Cambridge, Massachusetts: The MIT Press. Retrieved from http://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=3339972

Benevolo, C., Dameri, R. P., & D'Auria, B. (2016). Smart Mobility in Smart City. In T. Torre, A. M. Braccini, & R. Spinelli (Eds.), *Empowering Organizations* (pp. 13–28). Cham: Springer International Publishing.

Blank, S. (2013). Why the lean start-up changes everything. *Harvard Business Review*, *91*(5), 63–72.

Brock, J. K.-U., & von Wangenheim, F. (2019). Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence. *California Management Review*, *61*(4), 110–134. https://doi.org/10.1177/1536504219865226

Cruzes, D. S., & Dybå, T. (2011). Recommended Steps for Thematic Synthesis in Software Engineering. In *2011 International Symposium on Empirical Software Engineering and Measurement* (pp. 275–284). https://doi.org/10.1109/ESEM.2011.36

Deng, L. (2018). Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives]. *IEEE Signal Processing Magazine, 35*(1), 177–180.

European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. Brussels. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

Ferguson, S., Bennett, E., & Ivashchenko, A. (2017). Digital twin tackles design challenges. *World Pumps*, *2017*(4), 26–28. https://doi.org/https://doi.org/10.1016/S0262-1762(17)30139-6

Gartner. (2017). Gartner Identifies Three Megatrends That Will Drive Digital Business Into the Next Decade. Retrieved January 18, 2018, from https://www.gartner.com/newsroom/id/3784363

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California*

Management Review, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Haller, S., Karnouskos, S., & Schroth, C. (2009). The Internet of things in an enterprise context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5468, pp. 14–28). https://doi.org/10.1007/978-3-642-00985-3_2

Herterich, M. M., Holler, M., Uebernickel, F., & Brenner, W. (2015). Understanding the Business Value : Towards a Taxonomy of Industrial Use Scenarios enabled by Cyber-Physical Systems in the Equipment Manufacturing Industry. In *CONF-IRM 2015 Proceedings*. Retrieved from https://aisel.aisnet.org/confirm2015/31

Holzinger, A., Kieseberg, P., Weippl, E., & Tjoa, A. M. (2018). Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 1–8).

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, *63*(1), 37–50.

Kaplan, J. (2016). *Artificial Intelligence*. Oxford, England ; New York, New York: Oxford University Press. Retrieved from https://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=4705973

Khaitan, S. K., & McCalley, J. D. (2015). Design Techniques and Applications of Cyberphysical Systems: A Survey. *IEEE Systems Journal*, *9*(2), 350–365. https://doi.org/10.1109/JSYST.2014.2322503

Koçak, D. (2014). Thinking embedded, designing cyber-physical: Is it possible? In *Applied Cyber-Physical Systems* (pp. 241–253). Springer.

Ledwaba, L., & Venter, H. S. (2017). A Threat-Vulnerability Based Risk Analysis Model for Cyber Physical System Security. In *Proceedings of the 50th Hawaii International Conference on System Sciences* (pp. 6021–6030). https://doi.org/10.24251/HICSS.2017.720

Lee, E. A. (2008). Cyber Physical Systems: Design Challenges. In *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)* (pp. 363–369). https://doi.org/10.1109/ISORC.2008.25

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*, 18–23. https://doi.org/https://doi.org/10.1016/j.mfglet.2014.12.001

Leffingwell, D. (2018). *SAFe 4.5 Reference Guide: Scaled Agile Framework for Lean Enterprises*. Addison-Wesley Professional.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Murphy, K. P. (2012). *Machine learning : a probabilistic perspective*. (1970- Murphy

Kevin P., Ed.). Cambridge, MA: MIT Press. Retrieved from https://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=3339490

Myers, A. (2011). Stanford's John McCarthy, seminal figure of artificial intelligence, dies at 84. Retrieved from https://news.stanford.edu/news/2011/october/john-mccarthy-obit-102511.html

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, *17*(1), 2–26. https://doi.org/10.1016/j.infoandorg.2006.11.001

Nakajima, T., Ishikawa, H., Tokunaga, E., & Stajano, F. (2002). Technology challenges for building Internet-scale ubiquitous computing. In *Proceedings of the Seventh IEEE International Workshop on Object-Oriented Real-Time Dependable Systems. (WORDS 2002)* (pp. 171–179). https://doi.org/10.1109/WORDS.2002.1000050

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 2018). Determination press San Francisco, CA, USA:

Rajkumar, R., Lee, I., Sha, L., & Stankovic, J. (2010). Cyber-physical Systems: The Next Computing Revolution. In *Proceedings of the 47th Design Automation Conference* (pp. 731–736). New York, NY, USA: ACM. https://doi.org/10.1145/1837274.1837461

Rettberg, A., Pereira, C. E., & Soares, M. S. (2017). A Model-Based Engineering Methodology for Requirements and Formal Design of Embedded and Real-Time Systems. In *Proceedings of the 50th Hawaii International Conference on System Sciences* (pp. 6131–6140). Retrieved from https://aisel.aisnet.org/hicss-50/st/metrics_and_models_for_systems/2

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003

Sha, L., Gopalakrishnan, S., Liu, X., & Wang, Q. (2008). Cyber-Physical Systems: A New Frontier. *2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (Sutc 2008)*, 1–9. https://doi.org/10.1109/SUTC.2008.85

von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*.

West, T. D., & Blackburn, M. (2017). Is Digital Thread/Digital Twin Affordable? A Systemic Assessment of the Cost of DoD's Latest Manhattan Project. *Procedia Computer Science*, *114*, 47–56. https://doi.org/https://doi.org/10.1016/j.procs.2017.09.003

Wu, W., Kang, R., & Li, Z. (2015). Risk assessment method for cybersecurity of cyber-physical systems based on inter-dependency of vulnerabilities. In *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 1618–1622).

https://doi.org/10.1109/IEEM.2015.7385921

Xu, L. Da, & Duan, L. (2019). Big data for cyber physical systems in industry 4.0: a survey. *Enterprise Information Systems*, *13*(2), 148–169. https://doi.org/10.1080/17517575.2018.1442934

Yao, X., Zhou, J., Lin, Y., Li, Y., Yu, H., & Liu, Y. (2019). Smart manufacturing based on cyber-physical systems and beyond. *Journal of Intelligent Manufacturing*, *30*(8), 2805–2817. https://doi.org/10.1007/s10845-017-1384-5