Trust facilitating mechanisms in metahuman systems

A Case Study

Stian Grimsrud

Thesis submitted for the degree of Masters in Informatics:
Design, use and interaction
60 credits

Department of Informatics
Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO
May / 2022

Trust facilitating mechanism in metahuman systems

Stian Grimsrud

2022

2022

Trust facilitating mechanisms in metahuman systems

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

With new configurations of machine learning systems developed, the machine learning systems are getting responsibilities that humans usually had. This shift of responsibilities has highlighted the importance of trust in the dynamic for the human to the machine learning. However, there is little research done specifically on how to facilitate trust in metahuman can be done successfully.

This thesis extends existing literature concerning how to facilitate trust in metahuman systems by addressing the research questions: *Which mechanisms facilitate trust in metahuman systems?* Through a one and a half year engaged research project, I have collaborated with DNV that have developed a metahuman system. Metahuman systems refers to a hybrid system of humans and machines that learn and potentially amplify capabilities (Lyytinen, Nickerson, & King, 2021). The focus has been on examining DNVs metahuman system and their struggles on facilitating trust. Based on analysing the empirical findings, and discussion with related literature, I identify five trust facilitating mechanisms: (1) Constantly providing feedback to the system, (2) User getting a greater understanding of the system by using the system, (3) Involving users, (4) Accurate predictions of historically answered cases and (5) Producing service documents. I argue that the five trust facilitating mechanisms contribute to the existing literature.

Keywords: *metahuman system, mechanisms, agent, machine learning agent, human agent, agentic*

Acknowledgement

First, I would like to thank my supervisor, Alexander Kempton, for guiding me through the process of writing this thesis. Our discussions have been invaluable for the result of the thesis.

I would also like to thank the participants in my research project who took time out of their schedule to talk to me about their work. I would like to thank the participants from DNV, your input throughout the project was valuable.

Finally, I would like to thank my family and friends for all the love and support.

Stian Grimsrud

University of Oslo

May 2022

Contents

List of Tables

List of Figures

List of Pictures

Abbreviations

AI - Artificial Intelligence

ML - Machine learning

DNV - Det Norske Veritas

FAQ - Frequently asked questions

UCD - User centred design

DATE - Direct Access to Technical Experts

UX - User experience

HLEG-AI - European Commission High-Level Expert Group of AI

PD - Participatory Design

IS – Information systems

1. Introduction

Recent advances in technology have led to the rise of machine learning technology (ML), and these are further implemented or adopted by organisations. These intelligent systems solve complex problems commonly associated with human activities, including decision making, learning, and pattern recognition. Where humans usually would analyse large datasets, ML are now advanced enough to examine these vast datasets and even automating complex tasks earlier performed by humans. Some examples of this include autonomous vehicles, chatbots, filtering of social media content, intelligent wearables, etc.

Nevertheless, the rise of ML presents a challenge to trust research in Information Systems (IS) research, which earlier has been dominated by the conceptual and empirical assessment of interpersonal and organisational trust in non-intelligent technologies (Glikson & Wolley, 2020, Nissen & Jahn, 2021, Lockey, Gillespie, Holm, & Someh, 2021). There are some immediate concerns regarding this subject. How do new configurations of machine learning change the way we trust and trust-related processes in business-to-business connections? What forms will trust take in intelligent organisational systems? How do the decision-makers interpret and build trust in the application of ML? When and how do trust in ML, trust in organisations, trust in humans, and trust in technology augment and substitute each other? What shapes trust in ML and its consequences, furthermore the costs of the organisations?

Existing literature has discussed a generational shift in machine learning technology, regarding struggles in the new generation. This shift has introduced more responsibilities assigned to the machine learning technology, for example contributing to critical decision making and assisting virtually (Baird & Maruping, 2021, Glikson & Wolley, 2020). While these new configurations of machine learning technology have gotten new responsibilities and improved capabilities, humans still need to interact with them in a new way. Historically humans have controlled the machine learning system and ensured it did what it was assigned. As of now, humans need to cooperate with the machine learning agent in task solving and

can be perceived more as a teammate than a tool. These new configurations of machine learning systems and perception of the machine learning system, creates new challenges regarding facilitation of trust towards the new machine learning technology.

1.1 Research question

I look at the configuration between humans and machine learning systems as metahuman systems. As Lyytinen, Nickerson, and King (2021) define metahuman systems, they consist of humans and machines that learn, and amplify the capabilities and make the systems better at learning that humans or machines provide separately.

I extend the literature in trust facilitating mechanisms by addressing the following research question:

"Which mechanisms facilitate trust in metahuman systems?"

Trust facilitating mechanisms is seen as a process of action performed in organisations, to help the end users gain knowledge about the system, how to use the system, and be involved in the system. Secondly, there are processes of producing service documents, so the end users have the possibility to understand the system, without the system experts being present. Lastly, the system developers want the end users to be part of the knowledge sharing between the machine learning agent and the human agent, by constantly providing feedback.

In this thesis, I will examine the research question by first developing an understanding of the concept of trust in machine learning technology. Secondly, with this basis of understanding, I will analyse the empirical data from a case study, where I have conducted interviews with employees who are developing, maintaining, and using the machine learning configuration named Direct Access to Technical Experts (DATE). The focus is which mechanisms they have used to facilitate trust for the human agent to the machine learning configuration. In recent years DNV has developed an assistant, called DATE, for their case

handlers. The DATE-assistant helps the customers by taking their unsolved cases and gives it a category, for then to send the unsolved case to a case handler. Additionally, the DATE-assistant supports the case handlers with information related to the case they are working on. The DATE-assistant's responsibilities are to remove some of the manual work done earlier, such as categorising the case, finding related cases, information about the customer, which ship it is, the age of the ship, have there been problems with the ship in the past, and so on. Founded in 1864, DNV has a long history of helping customers in the shipping industry. They have over hundred thousand customers in over hundred different countries where the customer group involves a span of different disciplines: maritime, power and renewables, oil and gas, healthcare, etc. It is within this wide range of disciplines DNV operate and navigate; optimisation is therefore in their interest. All the shipowners with class agreement are offered access to the DATE-assistant if they desire, which makes the DATE-assistant widely used and consist of vast amount of information. When categorising cases to the experts, the DATE-assistant has seven hundred different categories to choose from.

1.2 Thesis structure

My thesis is further structured as followed:

Chapter 2: Theoretical Background

This chapter describes literature related to trust facilitating mechanisms in metahuman systems.

Chapter 3: Research approach and method

This chapter describes the background of my case, focusing on describing DNVs machine learning system DATE (Direct Access to Technical Experts). It is also elaborating on my focus during the study and chosen methodology - interpretive case study. Further I describe the methods used for data collection and the process of my analysis to highlight how I came to my contribution to the current literature.

Chapter 4: Findings

This chapter describes my empirical findings. I present which trust facilitating mechanisms DNV has used in their metahuman system.

Chapter 5: Discussion

This chapter starts by defining and describing the five trust facilitating mechanisms in DNVs metahuman system, and how these are related to findings from existing literature. I further discuss how the mechanisms serve different roles in facilitating trust in metahuman systems. Finally, I present how the five mechanisms contribute to both literature and practice in facilitating trust in metahuman systems.

Chapter 6: Conclusion

This final chapter summarises my thesis with concluding remarks.

2. Theoretical Background

The main goal of this thesis is to explore which mechanisms are used in facilitating trust in new configurations of machine learning technology. The research builds on, and aims to extend, the stream of literature concerning trust in systems where humans and machine learning technology collaborate. First, the aspect of trust will be explored, and how trust is valued in the metahuman system. The second section concerns metahuman systems, human-in-the-loop systems etc., and aims to create an understanding of these systems. Concludingly, I will emphasise why trust in new configurations of machine learning technology is highly needed.

2.1 What is trust?

Essential for this thesis, a definition of trust is needed. In this case, trust for the human user to the ML-system implemented by DNV are examined to understand how the human

user trust the answers given and how a trustworthy relationship for the human user to the machine learning technology is built.

First, we must define trust. One of the most cited definitions of trust is from Mayer et al. (1995). The authors argued that trust is "*the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*" (Mayer, Davis & Schoorman, 1995). This definition emphasises a willingness to be vulnerable and the importance of the actions at stake. Lockey et al. (2021) also conceptualise trust in this manner, where the defining components of trust are the intention to accept vulnerability based on positive expectations. This is not limit human-human interactions, which allows us to consider trust regarding technology; here, AI. Trust is relevant in human-AI interactions because of the risk embedded in AI relations due to the complexity and limited information about the outcome of the AI algorithm.

Because trust is a big part of AI, the European Commission High-Level Expert Group of AI (HLEG-AI) have defined some guidelines concerning trustworthy AI. More specifically, they define trustworthy AI as lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values), and robust (both from a technical perspective while also considering the social environment) (European Commission, 2021). Thus, what is the difference between trustworthy AI and responsible AI? By looking at Virgina Dignums' definition of responsible AI, "*Responsible Artificial Intelligence is then an approach that aims to consider the ethical, moral, legal, cultural, and socio-economic consequences during the development and deployment of AI systems*." Since the literature use the terms of responsible AI and trustworthy AI about each other (Arietta, et al., 2020), it is hard to separate them.

Microsoft has also been forward leaning and developed a set of guidelines for responsible and trustworthy AI. These six guidelines include accountability, ethics, inclusiveness, reliability, and safety, explainability, fairness, and transparency. However, what do these guidelines include? The ethical perspective mentioned by Microsoft regards fair and inclusive AI which should not discriminate against, or hinder, different races, disabilities, or people with diverse backgrounds. The AI should also be accountable for its decisions. This is

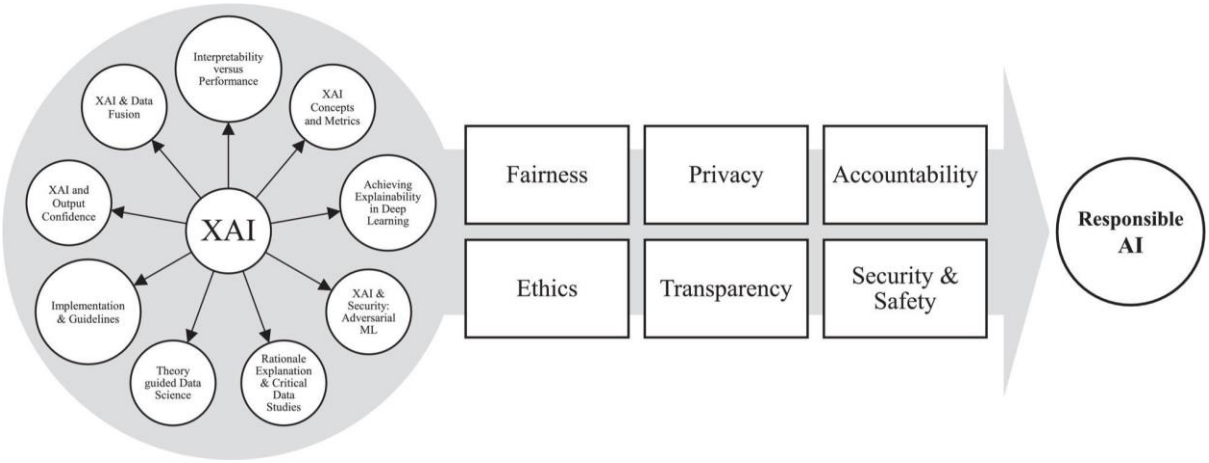
an important aspect that can lead to both moral and juridical difficulties. When Microsoft mentions accountability as one of their responsible and trustful AI guidelines, they mean that the people who design and implement the AI system need to be accountable for its decisions and actions. This could be problematic regarding who is responsible for the outcome; the programmer, the company delivering the product, or the consumer? Kirsten Martin (Martin, 2019) notes that an organisation must take responsibility for the ethical implications algorithm they have created. She argues that an organisation, by creating an algorithm that works in a particular manner, willingly becomes a party to the decision-making and therefore should be accountable for the decisions and outcomes. However, one can argue that although the organisation develops the algorithm, they are not part of the decision-making.

As for inclusiveness, AI should consider all human races and experiences. With the help of this inclusive design, the developers can acknowledge and address potential barriers, for instance, excluding specific populations. There are some examples of this type of exclusion. For example, the algorithm Amazon used during their recruitment process excluded all women, and consequently, they got rejected based on sex rather than knowledge and skills.

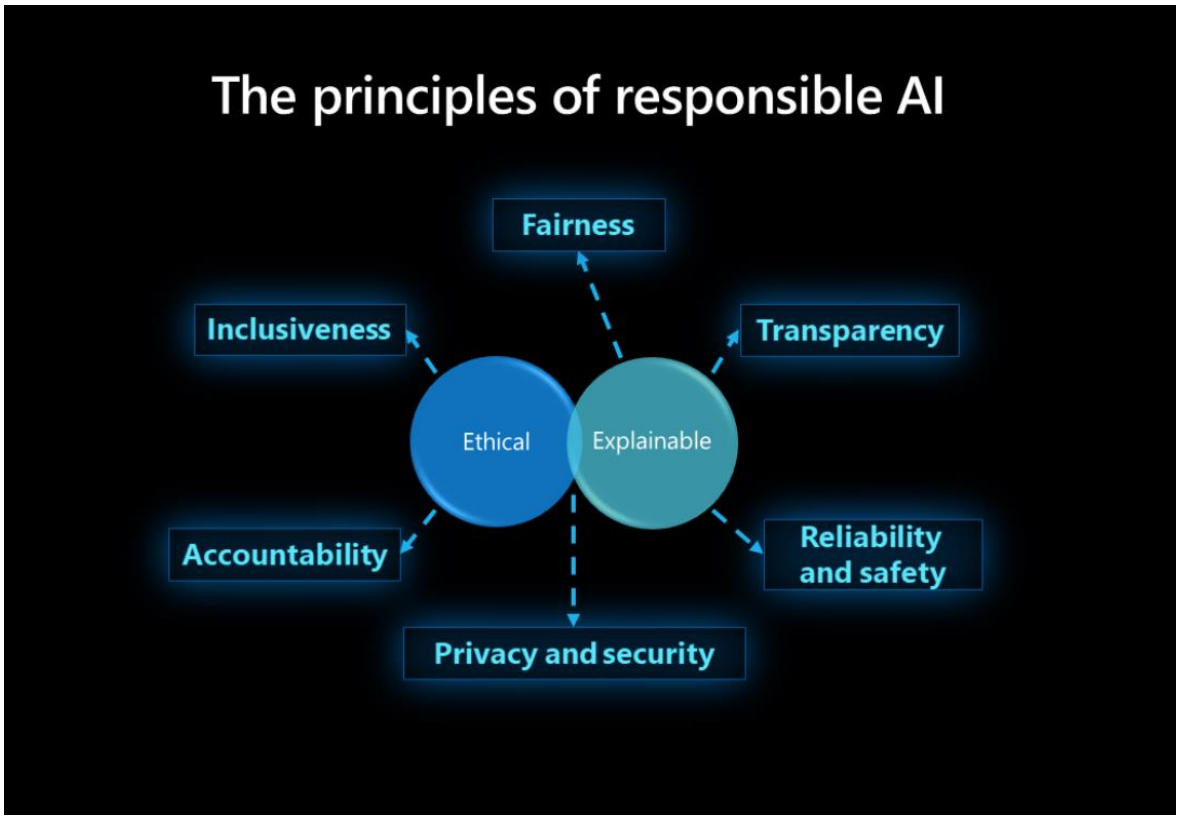
For the users of an AI-system, it is essential that the system is reliable and safe to use. That means that the system acts in the way it was supposed to and responds safely to new situations. The system should be tested rigorously and validated to ensure that the system responds in each manner. But the performance of an AI-system can degrade over time, which means that the organisations need to monitor the performance and then, if necessary, modernise the system.

The explainability guideline is there to help the data scientist, auditors, and decision-makers to ensure that the system can reasonably make the decisions made and how the system has reached a conclusion. Microsoft has even made some tools for making the AI-system explainable. These tools consist of glass-box models, black-box models, and one of their tools that indicates what influences the model; Fairline. The authors of *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI (Arietta, et al, 2020)* also argue that explainability is necessary for AI. The authors say that people use different terms within AI to describe the same concept, which is

problematic. Consequently, it is a need for a consistent use of terms in guidelines and frameworks. For example, the authors of the article have made a figure quite like Microsoft's, where Microsoft is referencing the AI being responsible and the authors to the AI being explainable. See the figures below, Figure 1 is the authors, and Figure 2 is Microsoft.



Picture 1 - Figure made by Arietta et al. about explainable AI



Picture 2 - Microsoft's principles of responsible AI

The last guideline regards transparency in AI. This guideline helps the team to understand the data and algorithm to train the model. Including what transformation logic is applied to the data, the final model, and the associated assets. This type of information offers insight into how the model was created and can recreate the model transparently. Nevertheless, as Glikson, E. & Woolley, A. (2020) argue, transparency is also an essential characteristic of interacting with the model to gain cognitive trust. Where they base the term *cognitive trust* on previous research on trust in technology, which states that *"when researchers examine cognitive trust in AI, they measure it as a function of whether user are willing to take factual information or advice and act on it, as well as whether they see the technology as helpful, competent, or useful"* (Glikson & Wolley, 2020).

2.2 What are machine learning systems?

During my academic studies, I have been introduced to artificial intelligence several times, though with different definitions and meanings according to the person introducing it. Therefore, this chapter aims to create a precise definition of the term "artificial intelligence", and what the terms refer to in this thesis.

2.2.1 Definition

Alzubi, Nayyar, and Kumar suggest that machine learning models have come to stay since the world has generated a huge amount of data. When using machine learning on huge amount of data, computers can imitate human-like behaviours, such as learning from experience. When using machine learning, each interaction with the machine learning system and each action performed becomes something the system can learn from. *"This virtual world has generated vast amount of data which is accelerating the adoption of machine learning solutions & practices. Machine Learning enables computers to imitate and adapt human-like behaviour. Using machine learning, each interaction, each action performed, becomes something the system can learn and use as experience for the next time. This work is an overview of this data analytics method which enables computers to learn and do what comes naturally to humans, i.e., learn from experience" (Alzubi, Nayyar, & Kumar, 2018).*

Seen in the Table 1 made by Benbya, Pachidi and Jarvenpaa (Benbya, Pachidi, & Jarvenpaa, 2021) there are a lot of different AI technologies and the different AI technologies are used in different domains. In my study, the focus is on machine learning systems. We can see from Table 1, that machine learning technology consists of reinforcement learning, supervised learning, and unsupervised learning. To describe the technology used in machine learning systems, the authors use bullet points seen in Table 1. The bullet points show that machine learning systems "*learns from experience", "learns from a set of training data", and "detects patterns in data that are not labeled and for which the results are not known".* All these are characteristics that can be seen in the DATE-assistant. Since the DATE-assistant is dependent on a set of training data to learn the different categories. Then the DATE-assistant can make categorisations based on the experience gained from the set of training data.

Table 1. AI Technologies and Domains of Application

Technology	Brief description	Example application
Machine learning Reinforcement learning Supervised learning Unsupervised learning	Learns from experience Learns from a set of training data Detects patterns in data that are not labeled and for which the result is not known	Highly granular marketing analyses on big data
Deep learning	A class of machine learning that learns without human supervision, drawing from data that is both unstructured and unlabeled.	Image and voice recognition, self-driving cars
Neural networks	Algorithms that endeavor to recognize the underlying relationships in a set of data through a process that mimics the way the human brain operates.	credit and loan application evaluation, weather prediction
Natural language processing	The ability of a computer program to understand human language as it is written or spoken	speech recognition, text analysis, translation, generation
Rule-based expert systems	A set of logical rules derives from human experts	Insurance underwriting, credit approval
Robotic process automation	Automates structured digital tasks and interfaces with systems	Credit card replacement, validating online credentials
Robots	Automates a physical activity, manipulates and picks up objects	Factory and warehouse tasks

Table 1: Benbya, Pachidi and Jarvenpaas Overview over AI Technologies and Domain of Application

2.3 Metahuman systems

In this subchapter I will explain what metahuman systems are, as well as how and which additional capabilities metahuman systems have, and how metahuman systems differentiate from the regular machine learning systems.
Before I discuss why trust is important in metahuman systems, I will explain what metahuman systems are.

2.3.1 Definition

Referring to Lyytinen, Nickerson and Kings (Lyytinen, Nickerson, & King, 2021) definition of Metahuman systems. "Metahuman systems are a hybrid of humans and machines that learn, complementing and amplifying capabilities that potentially make such systems better at learning than either humans or machines separately" (Lyytinen, Nickerson, & King, 2021). Hence this definition, metahuman systems is systems where machines and humans cooperate and potentially work better than a standalone part would do.

2.3.2 Where are these systems used?

Early use of machine learning, usually was about machine learning being a tool for humans in tasks that were repetitive, needed a lot of computational power, etc. As early machine learning systems were used as a tool for a human. The machine learning agent helped by exploiting its raw data power to do repetitive and time-consuming tasks for the human. But as of the latest years, new configurations of machine learning systems have emerged. The new configurations of machine learning systems can be seen more as a colleague than a tool (Wiethof, Tavanapour, & Bittner, 2021). As seen from Table 2 from Baird and Maruping, there are four different configurations of these systems. First, there are reflexive ones, which is often seen in voice-based assistants. Secondly, there is anticipatory, which anticipates the needs of the user. Thirdly, we have prescriptive ones, which are seen in bots, autonomous vehicles, etc. These are known for acting. Lastly, it is the supervisory, which often is seen in decision-support. It is here the DATE-assistant can be classified, as it is helping the human user in making decisions of how to act adequately.

Table 1. Agentic IS Artifact Archetypes

(limited decision-making latitude) ↕ (expansive decision-making latitude)

Agentic Archetypes	Examples	What's Different?
Reflexive (i.e., reactive)	• Sensing and acting (or alerting) agents, e.g., rebalance a financial portfolio when specified allocations are out of balance • Virtual assistants that react to queries (e.g., voice-based assistants)	These agents act reflexively, in direct response to relevant stimuli. Decisions are limited to models that define how to respond to expected stimuli.
Supervisory (i.e., control system)	• Behavior modification systems (e.g., decision support, ambient intelligence, health behavior nudges, or financial trade suggestions) • Guidance systems such as those that observe human behaviors and remind them of process steps (e.g., visual cues, such as from smart lights, that guide how to put together furniture)	Supervisory agents evaluate deviations from the norm (or the status of goal progression) and seek to guide decision making or take actions that will help return to the norm or enhance probability of progression toward a specified goal.
Anticipatory (i.e., proactive)	• Social media content searching, filtering, and presentation • Digital content compilation (e.g., automatic video or album creation) • Wearable augmented reality agents that anticipate needs (e.g., provide names for people in the field of view)	Anticipatory agents proactively apply model-based "reasoning" to anticipate needs or wants (e.g., the artifact automatically generates media compilations).
Prescriptive (i.e., autonomous decision-making)	• Bots (e.g., chatbots, search bots, resume filtering bots, etc.) • Autonomous vehicles • Automated financial portfolio management • Legal agents (e.g., arbitration or even judicial decision prescription) • Medical agents (e.g., that make decisions during procedures)	Prescriptive agents act as substitutes for either behavior-based decision-making or outcome-based decision making by prescribing or taking actions.

Table 2: Baird and Marupings overview over Agentic IS Artifacts

As Baird and Maruping argues for a technology to be considered an agent, it must "possess a degree of intelligence that permits it to perform parts of its tasks autonomously and to interact with its environment in a useful manner" (Baird & Maruping, 2021). The DATE-assistant examined in my case, is considered an agent, as it consists of intelligence in several manners. First, the DATE-assistant has the possibility of answering external customers autonomously if it is confident in the answer. Secondly, the DATE-assistant helps the human agent serving information concerning the case and checks if there are similar cases existing in the database of already solved cases. I will advocate the accuracy of the machine learning systems studied being agents, as they fulfil this degree of intelligence. Therefore, I will consider the DATE-assistant as a machine learning agent throughout this thesis. At the same time as I consider the DATE-assistant as a machine learning agent, there must be a human,

further called human agent, in the configuration for it to be a metahuman system. Therefore, the metahuman system consists of a machine learning agent (read DATE-assistant) and a human agent (read case handler). *"Following from this, our term agentic IS artifact refers to rational software-based agents that have the ability to perceive and act, such as take on specific rights for task execution and responsibilities for preferred outcomes" (Baird & Maruping, 2021).* Hence, Baird and Maruping and their characteristics of an agent, their term "*agentic IS artifact"* is used for software-based agents. Drawing the line to the DATE-assistant, the DATE-assistant can be classified as that. Since the DATE-assistant is responsible for categorisation of cases and handing out relevant information about the case to the human agent.

2.4 Why do we need trust in new configurations of machine learning systems?

"The vast majority of IS use research, assumes that IS artifacts are tools that serve as a means to achieving a user`s ends." (Baird & Maruping, 2021). As Baird and Maruping suggest, the majority of IS use research assumes that IS artefacts are tools that serve the meaning of the user. This is something that can be seen in the field of artificial intelligence too, where the early stages of AI-systems were used in repetitive tasks that humans did, like fabric work. Even though this was the early days of artificial intelligence, the use of IT-systems has always been dependent on trust, trust regarding if the technology is acting as it is supposed to.

As technology has evolved so has artificial intelligence. Baird and Maruping (2021) argue, there are a new generation of technology, which has given rise to different forms of IS artefacts. With the new generation of technologies rising, it has given rise to "*IS artifacts that are agentic in nature"* (Baird & Maruping, 2021). For IS artefacts to be agentic, they must "*have the ability to perceive and act, such as take on specific rights for task execution and responsibilities for preferred outcomes" (Baird & Maruping, 2021).*

This new generation of IS artefacts is "*imbued with the capacity to learn, adapt, act autonomously, and be aware of the need to act without being prompted by users. "(Baird & Maruping, 2021)* and came because of the massive potential of human-artificial intelligence systems to outperform either the human agent or the machine learning agent alone. Where the potential in these systems come from the human - AI collaboration. Human-AI collaboration *"implies that AI systems work jointly with humans like teammates or partners to solve problems ".* Another characteristic in these systems is that there must be "*at least one human and at least one AI interacting with one another in a shared environment or task*" (Schelbe, Flathmann, Canonico, & Mcneese, 2021). Where the argumentation and ability behind the human-AI systems *"lies in leveraging either agent's strengths"* (Schelbe, Flathmann, Canonico, & Mcneese, 2021). We also need to see the collaboration as a process, a process which is evolving and interactive, where the human agent and the machine learning agent are engaged in joint activities to achieve one or more shared goals. It is in this joint environment the metahuman system is operating, and the goal of the development of a metahuman system is to be superior to what the agents would be alone.

With this new generation of AI-systems, the artefacts have new capabilities and can act without being prompted by users, it highlights that there is a paradigm shift in the form of relationship between the IS artefacts and the humans. This new form of relationship needs to take into consideration the collaboration between the two agents. Since the machine learning agents' tasks have changed from repetitive tasks, to helping in decision making in potentially critical situations.

2.5 Mechanisms

As this thesis is exploring how to facilitate trust in metahuman systems, where I point to trust being facilitated through mechanisms. These mechanisms are what my findings rely on. When using the term mechanism, I rely on Østerlie and Monteiro (Østerlie & Monteiro, 2020) definition, where "mechanism" is used to explain a process of action. In this thesis, the process of action is how DNV facilitates trust for the human agent in the configuration of the machine learning agent.

2.6 Chapter Summary

To facilitate trust in metahuman systems, the developers need to develop knowledge of the interaction between the machine learning agent and the human agent. Metahuman systems refers to machine learning systems and humans cooperating in accomplishing a shared task. Based on this, I extend existing literature by exploring how an organisation facilitates trust in metahuman systems. Table 3 summarises the most important terms used in this thesis.

Term	Definition
Metahuman system	"Metahuman systems are a hybrid of humans and machines that learn, complementing and amplifying capabilities that potentially make such systems better at learning than either humans or machines separately" (Lyytinen, Nickerson, & King, 2021).
Mechanism	Is used to explain a process of action.
Agent	"For a technology to be considered an agent, it must "possess a degree of intelligence that permits it to perform parts of its tasks autonomously and to interact with its environment in a useful manner" (Baird & Maruping, 2021).
Machine learning agent	Refers to a machine learning algorithm which cooperate with a human.
Human agent	Refers to a human which cooperate with a machine learning agent.
Agentic	*"Refers to rational software-based agents that have the ability to perceive and act, such as take on specific rights for task execution and responsibilities for preferred outcomes" (Baird & Maruping, 2021).*

Table 3: Important terms of the thesis

3. Research approach and method

In this chapter, I will present my choice of methodology and methods of inquiry, and within which philosophical paradigm research has been conducted. I will explain why I choose to conduct my research this way, followed by an elaboration of how my fieldwork was carried out, including ethical considerations.

3.1 Case description: DNV and DATE

This study is conducted in cooperation with Det Norske Veritas (DNV) centred at Høvik. Since 1864 DNV has been a world leading classification society, and a recognized advisor for the maritime industry. DNV delivers world-renowned testing, certification and technical advisory services to the energy sector including renewables, oil and gas, and energy management. They are leading certification bodies, helping businesses assure the performance of their organisation, products, people, facilities, and supply chain.

As part of being a leading certification company the customers' ships need to be inspected regularly. Before the inspection is done by a surveyor in DNV the customers usually contact DNV regarding the parts that need to be improved.

During DNVs digitalisation, DNV has developed a machine learning system named DATE. As of now the shipowners with class agreement, can use the DATE-assistant, which gives them direct access to more than five hundred domain experts all over the globe in all different fields, such as hull, machinery, etc. Here they will have access to experts 24/7 from Monday to Friday, but DNV also covers urgent cases on weekends and holidays. Since there are three different offices which are behind this system. The main office is at Høvik, and when they leave for work, the Houston office will look at the incoming cases, and when the Houston office leaves the Singapore office is in. Referring to DNVs advertisement of the service, when having access to the DATE-assistant the customers will have direct and easy support and

technical expertise (DNV, 2021). Additionally, the response time will be convenient. Thirdly, the customers will get official and written replies to the compliance needs. Fourthly, the expert knowledge exceeds the technical aspect of the shipping domain, by having knowledge on the rules and regulations as well. Lastly, the customers will have an easy overview of the company's ongoing and past cases that have been solved.

The DATE-assistant's responsibility is to receive a case from a customer. Further, the DATE-assistant must use knowledge gained through thorough training, to categorise the unsolved case. After the categorisation is done, the DATE-assistant sends the case to the department which handles cases with the given category. Additionally, the DATE-assistant is supporting the domain experts with solved cases that are similar the one received from DATE-assistant.

3.2 Philosophical paradigm

When conducting valid research, Myers (Myers, 1997) argues it is based on some underlying assumptions. These assumptions can be categorised in the positivist, critical and interpretive paradigms. When conducting research these “worldviews” provide some boundaries of the choice of methodology, methods of inquiry and analysis. It is crucial, since these lenses make the researcher look at the world in a different way.

Within the field of Information Systems (IS), the positivist paradigm has been dominant, which is characterised by the researchers’ assumption that “[...] reality is objectively given and can be described by measurable properties which are independent of the observer and his or her instruments” (Myers, 1997). Testing of theories is a typical approach within this paradigm, aiming to “[..] increase predictive understanding of the phenomena”. Critical research, on the other hand, assumes that “[...] social reality is historically constituted and that it is produced or reproduced by people” (Orlikowski & Baroudi, 1991). Critical researchers believe that though people can act to change their social and economic circumstances, they are limited by social, cultural, and political domination.

The researcher aims to reveal, explain and change “[...] the restrictive and alienating conditions of the status quo” (Myers & Klein, 2011). The research in this thesis has been conducted within the interpretive paradigm. The assumptions of an interpretative researcher are based on a belief that “access to reality (given or socially constructed) is only through social constructions, such as language, consciousness, and shared meanings” (Myers M. D., 1997). The aim of interpretative research is to “produce an understanding of the context of the information system, and the process whereby the information system influences and is influenced by the context” (Myers M. D., 1997). When conducting research, I believe that who I am as a person and my prior experiences will influence my choices of how I both gather and analyse empirical data.

To choose the research paradigm for this thesis I used the article “Inquiry when doing research and design: wearing two hats” by Verne and Bratteteig (Verne & Bratteteig, 2018) and asked myself the three following questions: "1) Who owns the problem?", "2) Whose meaning is represented?" and "3) Who delineates the fieldwork?". Using this framework, I got insight into what type of research paradigm I used, and consequently this directed the methods being used in the thesis.

When deciding the research paradigm, I was convinced the interpretative paradigm would be best for my thesis, as the interpretive research paradigm assumes that access can only come through social means, such as language (Myers M. D., 1997), so when I conduct an interpretive study - an access I will get this through interviews. In this way, I will have the opportunity to understand the phenomenon through intersubjective understanding, where the experience is developed through my fieldwork. This type of fieldwork is done through case studies, where I will get descriptions of the phenomenon in the home context, and it will be helpful to get different versions of events. The different versions of events are based on the unique view the different people provide.

Descriptive questions are primarily designed to describe what is going on or what exists in a context. There are two standard methodologies when using descriptive research

questions: case studies and ethnography. The case study methodology is bound in the study of the phenomenon in the phenomenon's real context. On the other hand, ethnography requires the researcher to spend a considerable amount of time in the field. The researcher must understand the subject and further describe and understand the phenomenon in the subject's social and cultural context. The significant difference between these methodologies is the boundary-setting conducted by the researcher, where in ethnographic studies, the researcher follows the "flow." There are usually the same methods for data collection in both methodologies, such as interviews and observation, so the data collection is of similar practice. I will use a case study approach, where I will conduct interviews and observations. I will use the interviews to get insight into the field and the persons working within the studied company. This way, I can get inter-subjective answers to my study. I will use this approach because I think it will be the best approach to answer the research question through interviews and observations. By using the approach of case study, I will have the possibility of getting the thoughts and opinions of the employees in DNV concerning my subject matter.

3.3 Choice of methodology

In addition to the philosophical paradigm, a researcher decides on a research methodology - a strategy for how to conduct research. The strategy I chose in this thesis is a case study, explained in the subchapter below.

3.3.1 Case study

Case studies “investigates a contemporary phenomenon within its real-life context” (Myers M. D., 1997). I chose to use case study as my research approach because I found studying people's thoughts and opinions around trust in Machine Learning systems the most valuable. Myers et al. argue that case studies are “particularly well suited to do IS research, since the object of our discipline is the study of information systems in organisations” (Myers M. D., 1997) - where the researcher's interest has changed from a perspective studying technical issues to a perspective studying organisational issues instead.

Over the course of my thesis, the problem formulation evolved towards a focus on how DNV has facilitated their employees' trust when cooperating with their machine learning systems. One viable path to answering this question was to further understand how new users were introduced to the system, which challenges are there for new users, and which arrangements are used, to develop knowledge to support the machine learning systems design of constant feedback. This focus leads back to the use of case studies, which is described as a detailed inquiry of a specific case with a focus on the activities, functions, and local meanings within this case (Stake, 2005). As the process of developing knowledge about how the employees and the system collaborated involved a variety of different members within DNV, the need for gathering empirical data on a broad set of users and evolved people remained relevant for the defined problem.

3.4 Methods for data gathering

My method for data gathering has been interviews. My focus for the investigations has been to gain an understanding of how human agents can trust their configurations of machine learning agents. By having the lens of the interpretive research paradigm, it has broadened my understanding of what I want in the data collection and the focus of the case study. The role as an interpretative researcher is to interpret the empirical data, which in my case will be a DNV employee’s interpretation of the machine learning agent.

3.4.1 Data gathering activities

In this section I will explain how I conducted the case study, including a detailed description of how I used the methods mentioned earlier. I will also describe how I got access to the case.

My interviews have been conducted online, which over the last years have become a somewhat natural working context for the employees over the years because of the pandemic. Online work is how the human agent, and the machine learning agent have collaborated to complete assigned tasks.

3.4.1.1 Gaining access

As Crang and Cook (Crang & Cook, 2007) argue, gaining access to participants and/or a field of study may be difficult. As I wanted to do my fieldwork with DNV, and they had a stay-at-home policy because of the ongoing COVID-19 pandemic, I was afraid it would be difficult to recruit participants. Beforehand, I had some contacts inside DNV, which gave me the contact information to the head of machine learning department. After the initial contact, we had a meeting where I informed the person about my intentions, which was approved, and later encouraged to pursue within DNV by my contact. Thus, I had gained access to the organisation.

3.4.1.2 My role

As I did not know in advance how my online interviews would unfold, and I did not have much knowledge on machine learning systems in the start, I chose to act as a novice (Randall, Harper, & Rouncefield, 2007). The role is described by Randall, Harper and Rouncefield to be valuable, as one is "[...] licensed to ask, naive, even stupid questions and, thus explore much of what is tacit" (Randall, Harper, & Rouncefield, 2007) to the participants. I found this useful because my participants tended to explain the technical and social aspects of the system in unfamiliar wording. To not miss anything, I focused on asking questions that may be redundant, in addition to my planned questions.

3.4.2 Interviews

Interviews as a data collection method is one of the most common techniques for understanding the context of people's everyday lives (Crang & Cook, 2007).
Crang and Cook argue that interviews cannot be treated as one specific method, as "all social research involves learning through conversation. Within interpretative research, interviews are described as" [...] a keyway of accessing the interpretations of informants in the field" (Walsham, 2006).

Interviews can be done in a variety of ways, with one or more interviewees, like focus groups or questionnaires. Interviews will also vary in structure, from highly structured, to semi-structured, and to relatively unstructured. Structured interviews resemble questionnaires, where the predetermined questions and the interview structure is followed, while unstructured interviews resemble a conversation where the researcher tries to steer the conversation through a few themes he wants to explore. Semi-structured interviews use elements from both, by having a set of predetermined questions, but keeping the possibility of follow-up questions or deviations open. The hybrid of elements made me chose semi-structured interviews for my data collection.

My first interview was with the leader of the machine learning department in DNV, where the person's main responsibility is to lead the department in the right direction, at the same time making revenue for the organisation. This interview gave insight to which processes were in place, the working practices, and the environment of the machine learning system. The person also gave some directions, including suggesting people that could be interesting to interview as part of the early stages of the study.

Additionally, I had three interviews with internal scientists with the responsibility of making long term strategies for internal use. These employees have been in DNV for a long time and provided the organisation with good insight and input for the technological strategy. This was good for the early exploratory part of the study, when I was exploring what type of machine learning system they had, and the decision making. By getting access to the internal researchers, I got introduced to the long-term strategies. Which made me understand the processes behind the different strategies chosen, and how they are developed.

Thus, two interviews with developers of machine learning systems were conducted, where one of them was directly involved in the implementation and maintenance of the DATE-system, and the other in the implementation of similar systems both internally and externally. Both interviews gave me the technical knowledge of the machine learning system, what type of training the employees get, knowledge concerning the training of the machine learning system, and how the developers are trying to understand the problems of the users.

Three everyday users of the system were also interviewed as they provided another perspective on how it is to collaborate with the machine learning system. I wanted to understand the challenges and views of the people using the machine learning system directly in their everyday work, as I wanted a bottom-up perspective as well as a top-down perspective. These interviews were specifically relevant in the later stages of the study, as they provided concrete insight related to the trust and collaboration with the machine learning system.

Nr.	Time of interview	Role in organisation	Input	Length of Interview
1	7.10.2021	Leader of Machine Learning Department	Processes in place, the working practices, and the environment of the machine learning agent.	1:05:21
2	15.10.2021	Data Scientist in Equinor	Insights into trust in machine learning systems.	18:51
3	4.11.2021	Internal scientist	Understanding of the processes behind the different strategies chosen, and how they are developed.	34:53
4	5.11.2021	Internal scientist	Understanding of the processes behind the different strategies chosen, and how they are developed.	40:20

5	15.11.2021	Internal scientist	Understanding of the processes behind the different strategies chosen, and how they are developed.	38:28
6	26.11.2021	Developer of DATE-system	Technical knowledge of the machine learning agent, what type of training the employees get, knowledge concerning the training of the machine learning agent, and how the developers are trying to understand the problems of the users.	1:45:57
7	18.1.2022	Developer of machine learning systems	Technical knowledge of the machine learning agent, what type of training the employees get, knowledge concerning the training of the	49:45

			machine learning agent, and how the developers are trying to understand the problems of the users.	
8	1.2.2022	End User of DATE	Perspective of interacting with the machine learning agent	35:25
9	3.2.2022	End User of DATE	Perspective of interacting with the machine learning agent	30:21
10	4.2.2022	End User of DATE	Perspective of the interacting with machine learning agent	30:18
11	10.3.2022	Developer of DATE-system	More insight into the machine learning agent, the users and how the developers continuously improve m the machine learning agent.	1:08:20

3.5 Methods for data analysis

The analysis of this thesis is a parallel process of engaging in both the literature and empirical data, allowing for an abductive process where my contribution was iteratively shaped by both, by using a thematic analysis of my empirical data gathered through interviews with people inside DNV.

The following paragraph is a presentation of the methods I have used for analysing the empirical data gathered, through interviews since they were enriching. I have analysed the empirical data collected using thematic analysis. I will explain thematic analysis further in the coming subchapter.

3.5.1 Thematic analysis

All the empirical data I have gathered for this thesis is qualitative, and the material I have analysed is transcribed interviews. For my analysis, I chose to use coding techniques from grounded theory, which often is used in case studies (Lazar, Feng, & Hochheiser, 2017).

Thematic analysis is characterised as a flexible method to identify, analyse, and develop patterns in empirical data. The researcher chooses to do the analysis with an inductive, or bottom-up approach, which leads to the themes being "[...] strongly linked to the data themselves", as opposed to the researcher aiming to "[...] fit the data into a pre-existing coding frame, or the researchers' analytic preconceptions" (Braun & Clarke, 2006), as done with a deductive approach. Where the latter, the researcher has a theoretical or analytical interest in the data.

As I did not want to ignore parts of my empirical data that at first sight did not relate to my research, I used an inductive approach in my coding process with themes identified in the empirical data. However, since my original interest in the themes was theoretical, some of the themes identified in the empirical data were naturally inspired by related literature and thus can be characterised as partly deductive. As Braun and Clarke emphasise; "[...] researchers cannot free themselves of their theoretical and epistemological commitments,

and data are not coded in an epistemological vacuum" (Braun & Clarke, 2006). This highlights how emergent themes are likely to be influenced by the researcher.

When conducting the thematic analysis, I started by categorising my empirical data in different themes. These themes are later mentioned as trust facilitating mechanisms. After having sorted out the mechanisms, I found the relevant data for the different mechanisms. The data is used to explain the themes and their role in mechanisms of facilitating trust in the configuration of the machine learning system.

The configuration analysed in the case with DNV is illustrated in Figure 1.Where the machine learning system is part of a configuration of a metahuman system. Where a metahuman system accordingly to Lyytinen, Nickerson, and King is *"Metahuman systems are a hybrid of humans and machines that learn, complementing and amplifying capabilities that potentially make such systems better at learning than either humans or machines separately"* (Lyytinen, Nickerson, & King, 2021). The metahuman system analysed consists of two human users, where the first one is the customer who pays DNV for the service. When the customer has a problem, it contacts DNV through the machine learning agent, named DATE-assistant, and creates an unsolved case. When the DATE-assistant receives the unsolved case, it will categorise it. The categorisation is based on knowledge the DATE-assistant has gained through training. After categorising the case, the DATE-assistant sends the case to the second human user, the case handler.

The case handler is a domain expert in the category given by the DATE-assistant. In my analysis, I found out at this stage, one of two actions can happen. Either the DATE-assistant has wrongfully categorised the unsolved case, and therefore the case handler must manually re-categorise the case to another case handler. If the case gets re-categorised, the DATE-assistant gets feedback on the fact that the category is wrong, and it should be what the case handler has manually done. Or, the categorisation done by the DATE-assistant is right. If so, the case handler solves the case with information provided by the DATE-assistant. Then, at the same time as the case handler closes the case and sends the solved case to the customer, the DATE-assistant gets feedback on the fact that the categorisation is correct. Where this will become a cycle of feedback to the DATE-assistant, and it will learn for every iteration done.

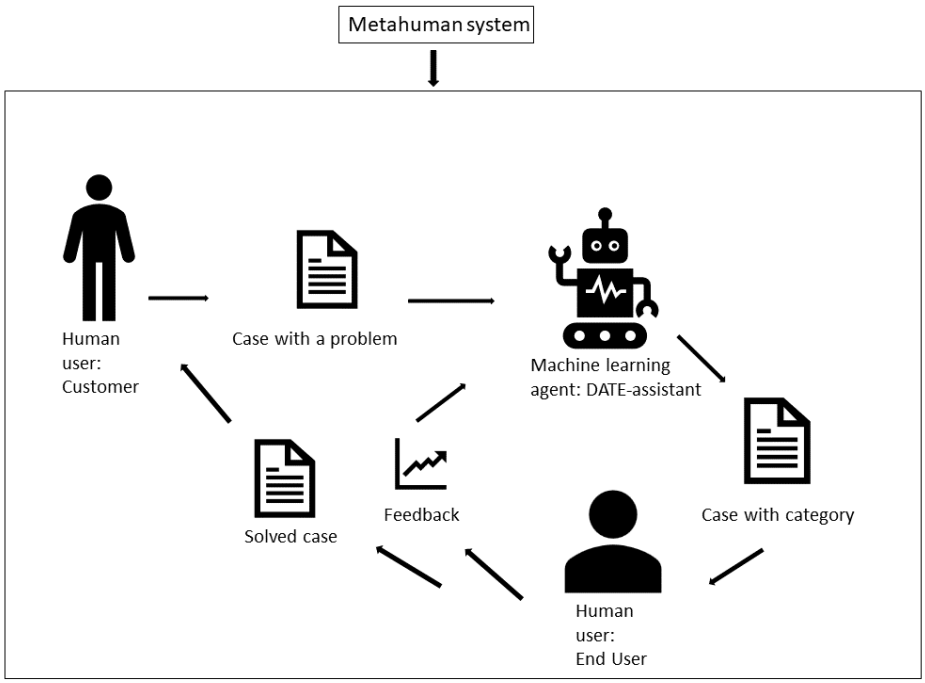


Figure 1: The Metahuman system

3.6 Ethical considerations

When doing research with people, the researcher usually gets access to a lot of information about them - in my case mainly insight into their thoughts and feelings regarding their work. It was important for me to specify that my only interest was about the systems and processes related to them, and not measure their work in any way. And this facilitated for a trustworthy relationship, where the participants trusted me with their information and therefore possibly shared more than if the relationship was untrustworthy.

3.6.1 Consent form

All the people I have been interviewing have signed a consent form (see Appendix C) specifying the object of my thesis and what their participation involves. The consent form also specifies that the data from audio recordings will not be used for other purposes than my

thesis, that participation is voluntary, and they have the right to withdraw their consent at any given time.

I also got some questions regarding the publication of the thesis, where one of the persons interviewed was uncertain about how much they were able to tell in the interview, if the thesis gets publicised, since it was a matter of organisational secrets. This made me talk to my supervisor about the possibilities of not publicising. Consequently, I approached my informants and established dialogue on the matter, where we concluded that if there is anything they could not say, I would take it out of the thesis. Therefore, I can publicise the thesis.

3.7 Chapter Summary

In summary, the empirical research in this thesis is based on an engaged research project in collaboration with DNV, where the goal has been to address a real-world problem, at the same time as contributing to academic literature. The engaged research project has been conducted in an iterative manner. The activities of investigating the topic in collaboration with DNV and formulating the research problem has been continuously revisited and evaluated over the evolution of the project. Based on the evolving understanding gained and investigation of feasibility to address the evolving research question, case study has stayed the form of inquiry. I have utilised interviews for data collection, focusing on gathering a broad set of perspectives with different actors within DNV. During the efforts of data collection, I have gradually built an understanding of the dynamic between the DATE-assistant and the case handler interacting with it. Additionally, the dynamic has made me understand how trust is facilitated in this configuration of metahuman systems. My case study can be identified as interpretative, since I attempt to understand the context of interest through intersubjective meanings, experiences and thoughts of the informants existing in the context. The analysis of the empirical data gathered has been a parallel process of engaging in both the empirical data and related literature, allowing for an abductive process where my contributions have been shaped iteratively by both. This has been done through several rounds of thematic analysis to describe the identified

phenomenon. The result of this process has been the contribution of five trust facilitating mechanisms.

4. Findings

In this chapter, I will start by describing the analysed configuration of the metahuman system. Next, I will explain how trust facilitating mechanisms are part of this configuration of the metahuman system.

4.1 Configuration of the metahuman system

The configuration of the machine learning system is illustrated in Figure 2. The configuration consists of two human users, where the first one of the human users is the customer. When the customer has a problem with a ship or something in that regard, they send their concerns to DNV. The customers must send their concerns through a specific email address, which is connected to the machine learning agent. As seen from Figure 2 the machine learning agent which handles the cases inside DNV is the DATE-assistant. The DATE-assistant then uses the knowledge it has gained through huge amounts of training, to categorise the case sent from the customer. After the categorisation, the DATE-assistant sends the case to the other human user in the configuration. This human user is the case handler, the case handlers are domain experts within different domains of the ship industry. When the case handler receives the case, one of two things happen. Either the case is wrongfully categorised by the DATE-assistant. If so, the case handler needs to look at the case, and decide which new category the case needs. After deciding which new category the case needs, the case handler manually address which domain it is related to and send it to the right department within DNVs organisation. When cases get manually re-categorised by the case handlers, the DATE-assistant gets feedback that the given category is not correct, but the category given by the case handler is correct. Or, if the case is rightfully categorised, the case handler is set to solve the case. To help the case handler in solving the case, the DATE-assistant provides the case handler with previously answered cases, which is similar. The DATE-assistant also gives the case handler a percentage amount of similarity between the current case, and previously

solved cases. The DATE-assistant does this to help the case handlers in the process of solving the case. When the case handlers get information of similar cases, they can give a thumbs up or thumbs down, which gives the DATE-assistant feedback on the similar cases provided to the case handler. After the case handler has solved the case, the case handler contacts the customers with the answer, and marks the case as closed. At the same time as the case handler marks the case as closed, the DATE-assistant gets feedback that the category given on the case was correct. From here the DATE-assistant gets trained on the new set of data and will gradually be more accurate.

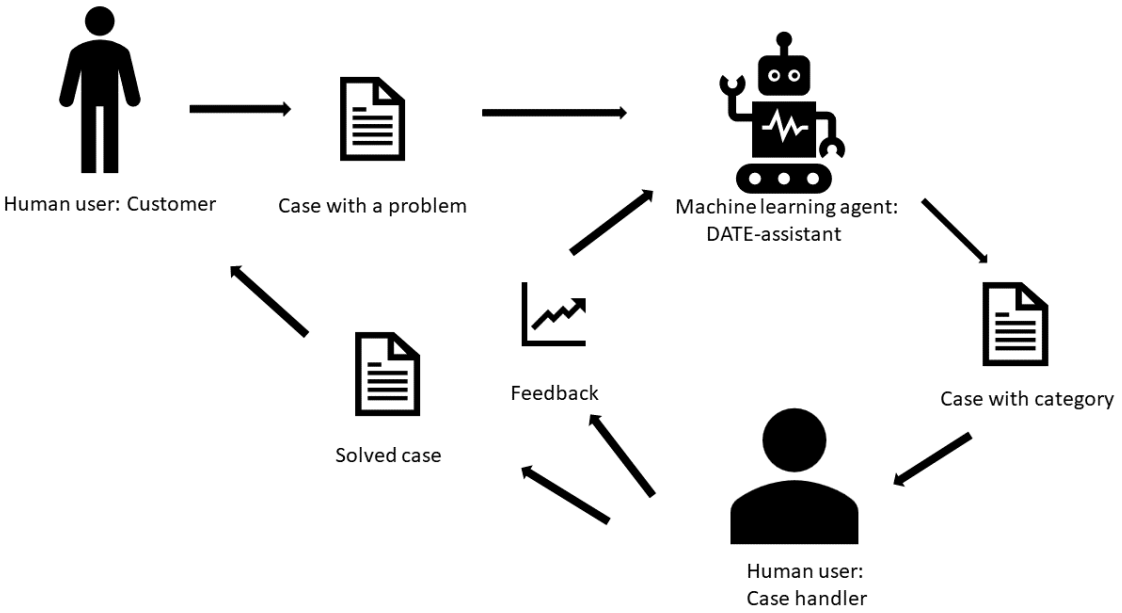


Figure 2: DNVs configuration of metahuman system

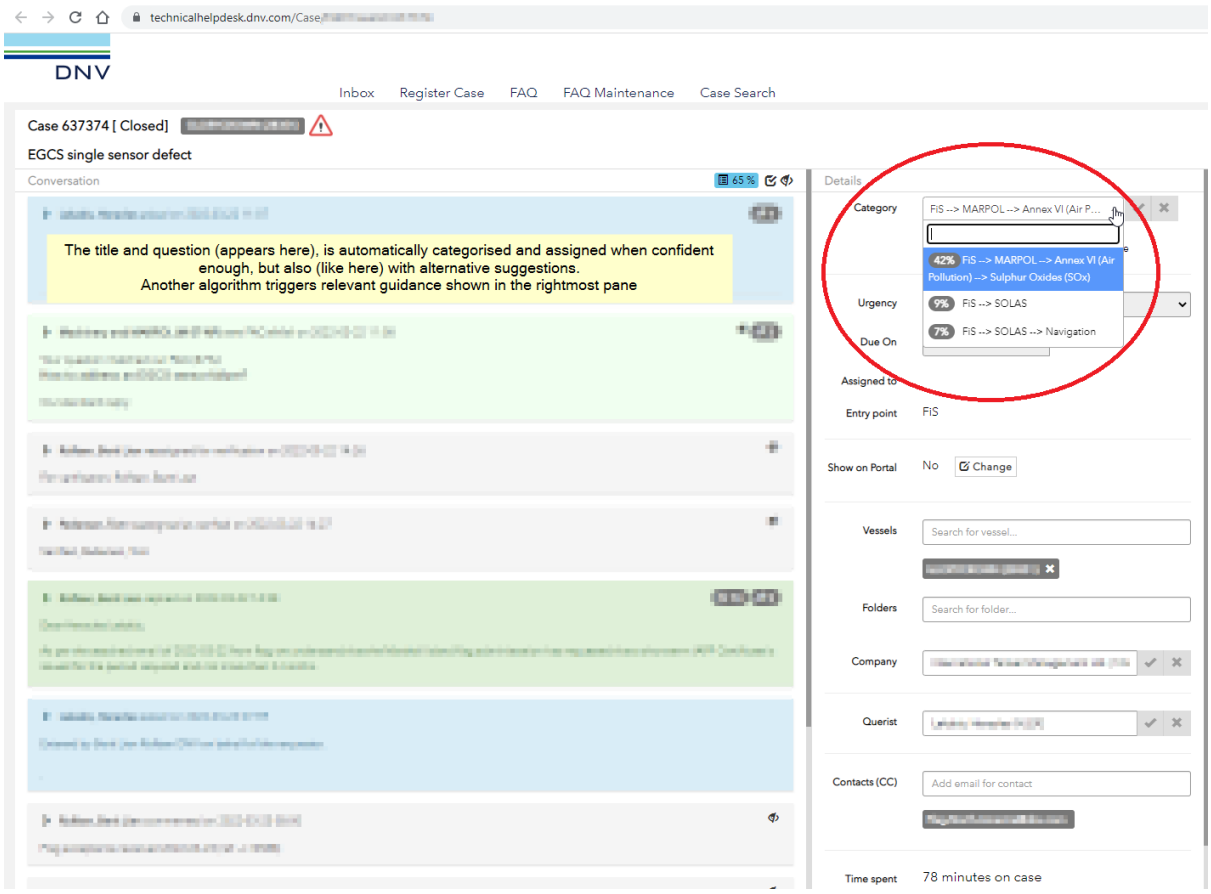
4.2 Mechanisms as part of the configuration

As part of the configuration, we can from Figure 2 see that the DATE-assistant is the machine learning agent in the analysed configuration where the trust facilitating mechanisms are a process of action in this configuration of the DATE-assistant.

In the following subchapters, I will present the five trust facilitating mechanisms that are part of the illustrated configuration.

4.2.1 Constantly providing feedback to the system

The goal of the DATE-assistant is to categorise between seven hundred different categories in real time and then send the case to the case handler, which is a domain expert, so the customer gets the help they need. For the customer to get the best help, the case needs to be sent to the right department in DNV. The DATE-assistant does this categorisation and then sends the unsolved case to a case handler, with a percentage of confidence. The case handler then needs to check the accuracy of the categorisation given by the DATE-assistant. If the DATE-assistant has done it correctly, the case handler will get served answers from similar cases, correspondingly marked with the percentage of similarity. When the case handler is served with answers from relatively similar cases, they can give feedback to the answers with a thumbs up/thumbs down function, seen in Picture 4.



Picture 3: Percentage of confidence in categorisation

The provided feedback is used in the training of the DATE-assistant, which will influence the cases shown. The feedback on the categorisation itself stems from the fact that

the case handler is doing the case or re-categorising it. When the case handler re-categorises cases, the DATE-system gets feedback on what category the case handler gives the case and will take the re-categorisation into consideration in a new iteration of training. The same will happen if the case handler solves the case in the category which the system has proposed, but then in a positive loop in training.

"On the category itself, you can change it and you can confirm." – (Informant 9)

Through the learning phase of the case handlers, the system experts try to get the case handler to understand that the more feedback they give to the DATE-assistant, whether it is right or wrong, the more accurate the DATE-assistant will become in the long term. This is something they are "pushing" on the case handlers as the DATE-assistant needs the input from the case handler to gradually improve and become more valuable. As seen from Figure 3, this will become a continuous cycle, which will help the DATE-assistant improve for every iteration completed.

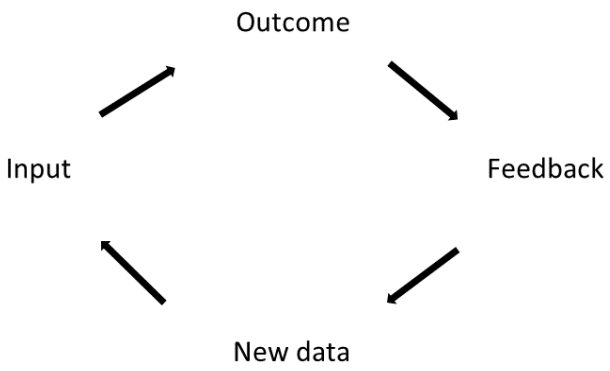


Figure 3: Feedback loop to the DATE-assistant

"[..] On the proposed replies, because we have some proposed FAQs, there are thumbs up and thumbs down, and kind of neutral. That's the way we teach the system, how we tell the system whether this was a good match or not. "– (Informant 9)

As said by informant 9, this is their way of updating the knowledge of the DATE-assistant. Which in turn will make the system more accurate as seen in Figure 3.

"Yeah, that's for the new ones. It's still this training phase where we use a lot of these thumbs up thumbs down or edit buttons or propose to the developers that we should change because it's not really correct. "– (Informant 9)

As informant 9 mentioned above, when new subjects or cases are introduced to the DATE-assistant. The case handler has responsibility in knowledge sharing between the case handler and the DATE-assistant. The DATE-assistant is then put in a training phase, in this phase the case handler will give feedback, thumbs up and down, to the proposed information from the DATE-assistant.

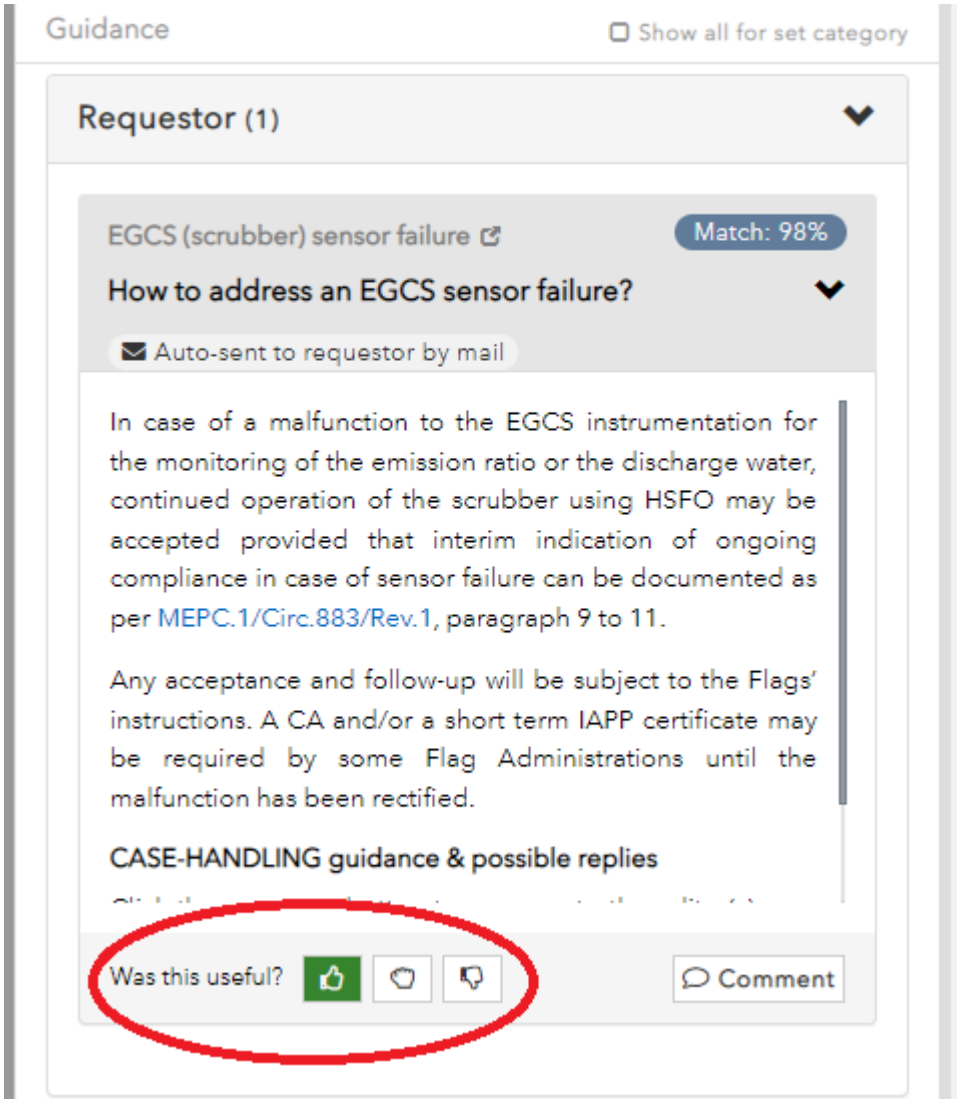
"She [co-worker of the informant] assigns and we are supposed to take it and start training the system on these FAQs and find the best matching proposals. This is for the new answer so it's kind of an iteration, so we go and look at how often the system gets the right answers. And if not, we can go in and change some of the parameters or do some more training. "-(Informant 9)

As shown in Figure 3, when the case handler constantly is providing the DATE-assistant with feedback on the given information, the DATE-assistant is evolving and getting better for each iteration of training. Consequently, when the DATE-assistant gets better and more accurate, the information provided by the DATE-assistant is more precise, which in turn will help the case handler trust the DATE-assistant, as the DATE-assistant is providing information as the case handlers perceive as correct.

"Yes also, if we take the case handler then, then they contribute, we try to get a system where they contribute to the knowledge as well. And here we are talking about a ML-

system where you want feedback on the predictions and what the DATE-assistant did. Because there it is, it is like a child, you say that the child should do something, then the child does something and then you correct along the way. So here we try to inspire then these case handlers to give thumbs up and thumbs down depending on whether that knowledge was useful or not, and if they then give thumbs down, then we run a new retraining, with new data, so that the prediction is a bit different day No. 2. and has taken into account the thumbs down and should give a little less chance that pops up that knowledge in a similar context in the future. "-(Informant 11)

As informant 11 emphasise, the case handlers are participating in the knowledge sharing to the DATE-assistant. Where the DATE-assistant is coming with proposals related to a subject and the case handler is telling the DATE-assistant which is the best match. This process of knowledge sharing between the case handler and the DATE-assistant is something DNV facilitates and encourages. As seen from Figure 3, feedback from the case handler to the DATE-assistant is essential in the configuration. Where the process of constantly giving feedback to the DATE-assistant contributes to the DATE-assistant being more precise and confident in future predictions. When the DATE-assistant gets better through the feedback provided by the case handler, it facilitates the case handlers trust in the DATE-assistant, since the case handlers perceive the information from the DATE-assistant as correct.



Picture 4: Feedback mechanism to the DATE-assistant

4.2.2 Users getting a greater understanding of the system by using the system

During the lifetime of the metahuman system, DNV wants the case handlers to learn the DATE-assistant by using it. Thus, they start exposing the case handlers for the system early in the learning phase of the case handlers. During this learning phase, the case handlers will be introduced to the DATE-assistant and have their first interaction with it, as well as they have the possibility to ask a system expert.

"I'm not sceptical anymore[...]". -(Informant 9)

As you can see from informant 9, when they started using the DATE-assistant the person did not trust the decisions made by the DATE-assistant. Nevertheless, by having this learning phase, where the case handlers of the DATE-assistant have the possibility to ask system experts. The case handlers get clear expectations of what the DATE-assistant can or not. Which over time has made informant 9 gaining trust in the decision making of the DATE-assistant.

The case handler would first try out the DATE-assistant based on curiosity, compelling need, or another reason. The case handler would then over time build his or her own experience around the DATE-assistant. As informant 7 exemplifies:

> *"An example which comes to my mind is Netflix or Amazon, these recommendation systems. And they navigate you Stian, when you first meet the system of Netflix, you are maybe hesitant, maybe curious and at the same time sceptical. And all of the sudden you find yourself trusting Netflix's recommendation more than your own gut feeling." -(Informant 7)*

But, the big difference between DNV and the big players in the consumer market such as Netflix, Amazon, etc. is that there is a habituation phase for the case handlers of the DATE-assistant. In this phase there is always an available system expert behind the application.

> *"They [the case handlers] will speak to a person in flesh and bones that will help them understand the problem and understand the limitations and interpret some of these results. This happened for instance some months ago. I had one of our case handlers, an internal case handler came back to us, and said I see these deviations on the results. I got this number from the model and then I got this other number, which is the real number and I see some differences. Can you help me figure it out? And then you know, I got to talk to this person through how machine learning is operating, works, and getting the user to understand that we are not talking about 100% perfection all the time. But there are some intuitions which are inevitable, but that doesn't necessarily mean that the model can't be used. The model is a guideline." -(Informant 7)*

This constant availability of a system expert helps the case handlers start using the DATE-assistant. If there are some challenges in the use, the case handler can ask one of the available system experts. After the respected help, the case handler may have problems interpreting the results from the DATE-assistant; it can have a wrong categorisation, or it can have some bizarre recommendations of FAQs. Consequently, the case handlers will have the option to ask a system expert.

> *"It is really getting them [case handlers] to understand the data, the model limitation and how to interpret the model results. That, you know, goes a long way in building trust in the application itself."-(Informant 7)*

This is how one of the system experts (informant 7) is defining trust; you need some sort of knowledge, for them to get an understanding of the mechanisms, limitations, and possibilities. When you have this understanding of the DATE-assistant or a similar application, it is easier for a case handler to trust the results from the DATE-assistant, as it provides an understanding of why this is the results of the given input. This facilitates the case handle to trust the DATE-assistant, since the case handler understands why the DATE-assistant derived the results given to the case handler. Consequently, by understanding the limitations of the DATE-assistants' capabilities, and how it comes to results, it can help facilitate trust for the case handlers to the DATE-assistant. This is also stated by informant 9:

> *"Yeah, exactly. The more I use the system [DATE-assistant] I think by giving feedback to the developers and the system [DATE-assistant], they are able to tune the accuracy, so we are also a case handler. We [case handlers] are getting more confident as time goes by. "-(Informant 9)*

> *"I'm not sceptical anymore, in the beginning when we started, I was a bit more, but the accuracy was not that good at that time." -(Informant 9)*

As seen from Figure 2, the trust facilitating mechanism of constantly giving feedback to the DATE-assistant is part of the configuration. When the case handler gets a richer

understanding of the DATE-assistant's capabilities. This makes the case handler know about how the DATE-assistant bases its decisions and how it reaches prediction. This facilitates trust for the case handler to the DATE-assistant since the case handler understands the DATE-assistant after using it over time. As informant 9 said during one of my interviews: "*I'm not sceptical anymore, in the beginning when we started, I was a bit more, but the accuracy was not that good at that time".* This states that the case handler trusts the DATE-assistant by using it over time.

4.2.3 Involving users

> *"[...] it's the first, second, third, the last thing which matters here is getting involved with the case handlers. And try to understand how they think about that problem, and how they see success." -(Informant 7)*

As part of the implementation of new metahuman systems in DNV, they will have case handlers involved in the process. In the involvement of case handlers, DNVs developers strive to get an understanding of the expectations the case handlers have for the new metahuman system. They have some questions they want to get answered as part of this acquirement:

- What do the case handlers expect the DATE-assistant can do?
- Does the case handler understand the DATE-assistant they get provided with?
- Does the case handler understand the estimates the DATE-assistant is giving?
- Do the case handlers understand that the DATE-assistant is not perfect?

As informant 7 describes it:

> *"We will have a chat, then the question would be "if you were to describe what you have in mind, and something that will make you happy with this project, please tell how this would look like". And by answering that question you get a lot of information about their expectations, their problems, how they vision of the problems, what they*

use it for. For it may be in the start that they say they want this, but in the end, they want something else, but thought that this was what they needed." -(Informant 7)

By asking these questions to the case handlers, the developers get an understanding of expectations, the case handlers' problems, how the case handlers think about the problem, and what they will use to solve the problem. As informant 7 emphasises, it is important that the case handler gets an understanding of how to interpret the estimates provided by the DATE-assistant and that the DATE-assistant is not perfect. Both factors are principal for the data scientist to communicate with the case handler.

"[..] so does the case handler understand the model that I am providing. Does he understand the, you know, the estimates the model is providing? Does he understand that this model isn't perfect?" -(Informant 7)

Hence, by involving the case handlers early, by asking the questions mentioned by informant 7, informant 7s team gets an understanding of case handlers' usage of the DATE-assistant, what type of struggles they have, and how they succeed. It is also important that the case handlers get the understanding of how to interpret the results given by the DATE-assistant. By involving the case handlers, the developers gain a greater understanding of the struggles, measures of success, what type of knowledge they already have or not have, which in turn can help in the facilitating of gaps of knowledge and unrealistic expectations. Furthermore, this process facilitates trust for the case handler to the DATE-assistant, since the case handler is involved in the design of the DATE-assistant, therefore the case handlers do not have unrealistic expectations of the DATE-assistant.

4.2.4 Accurate predictions on historically answered cases

Before becoming digital, DNV answered their cases without a digital interaction between the customer and the case handler in DNV. Some of the cases that arose at that time are still quite common. Because the shipping industry still has topics that were relevant in 1970, such as damage to the hull, lightning of the ship, fire on the ship, etc., and they still are today. These types of cases have been part of the shipping industry for decades. During the

decades DNV has solved numerous similar challenges and although the technology has changed, it is still relevant to look at how these cases have been handled in the past.

Since the DATE-assistant is trained on cases done by DNV, there are a lot of recurring cases over the decades. These cases have been part of the training data of the DATE-assistant since they released it, which makes the accuracy of these cases higher than newly introduced cases.

> *"...So, it's working well, definitely for subjects that we have been handling for some time. It's not always good for new, let's say new categories or new subjects. This still requires probably some more training, so it is trained properly. "-(Informant 9)*

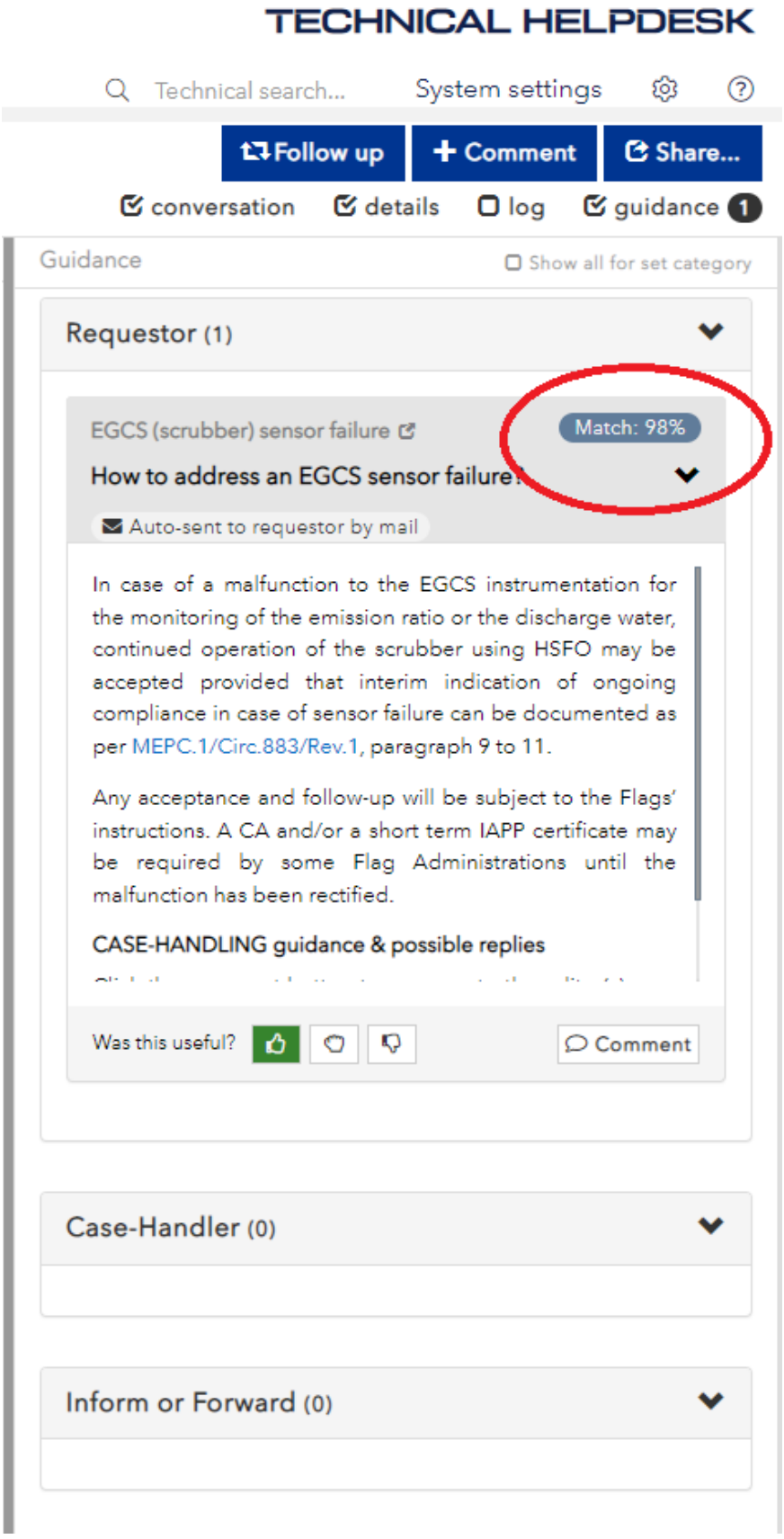
As informant 9 highlights, the DATE-assistant is not always categorising correctly when introduced to new cases or cases slightly different from previous cases. This is emphasised by informant 9, as the DATE-assistant is trained on a limited set of cases, so when a new case occurs, the DATE-assistant does not know how to categorise it. Here, the case handler has an important role in the improvement of the system, as the DATE-assistant does not know the right category. When this occurs the case handler must manually assign the case to another case handler. When the case handler has set the new category, the DATE-assistant will use the new data in training, with the intention of using the same category in similar cases in the future. This happens in more than one iteration, which will gradually make the DATE-assistant more accurate and confident with new cases.

As part of this, the case handler is aware of new subjects and cases introduced to the DATE-assistant. The case handler will then know which cases that may need some extra attention and later be manually re-categorised. Since the accuracy of new subjects, categories, or questions, is lower than the regular cases since the DATE-assistants have not trained on them yet. Informant 9 is having a hard time trusting the categorisation done by the DATE-assistant, concerning new categories. At the same time informant 9, trusts the DATE-assistant's categorisation of well-known cases.

"There is actually quite a lot in being such a person [case handler] and having this job. Also, the tool [DATE-assistant], both a, yes not true, is the means / tool you use to implement this, and the tool [DATE-assistant] helps you then implicitly through that this conversation then what you talk to the customer about is stored in one place, so it's retrievable. and therefore, you have to use that tool and not talk to people on regular mail."-(Informant 11)

Hence, by using the DATE-assistant when helping the customers, the answers provided to the customers are retrievable which makes them reusable for other case handlers in the future. As emphasised by informant 9, the accuracy of answers given to “known” cases, is high.

Seen in the configuration, Figure 3, the feedback loop from the case handler to the DATE-assistant is iterative and continuous. This will mean that for every case done, the DATE-assistant gets feedback on the proposed categorisation. Thus, when the DATE-assistant is put in front of a case which has been in the system for a long time, it knows what to do about it, as it has sufficient training on a huge number of similar cases. All the learning done by the DATE-assistant through immense amounts of similar cases and over a long time makes the accuracy on historically answered cases high. This helps the case handlers trust the DATE-assistant, since the DATE-assistant categorises these cases with a high accuracy, which is a result of all the feedback gained and the continuous training cycles done with the DATE-assistant.



Picture 5: Indication of similarity between cases

4.2.5 Producing service documents

Production of service documents is one of the trust facilitating mechanisms in this configuration of the metahuman system. When using the term service documents, training of the system, documentation, FAQs, and webinars are included in the definition. Although, one of my informants, which has a responsibility with producing service documents, mentioned that they [DNV] wished they did not need documentation and webinars, as they want the DATE-assistant to be as intuitive as possible. Albeit, they have made some service documents to help the case handlers of the DATE-system. I will present and describe the different types of service documents provided by DNV.

4.2.5.1 Training of the system

First part of the service documents is the initial training of the case handler in the configuration of the system. This is the first interaction with the DATE-assistant. During this training phase the case handlers have the option to ask questions directly to the system experts which has developed the DATE-assistant. The case handlers will also get help in how they [case handlers] should interact with the DATE-assistant, what they can expect from the DATE-assistant and getting an understanding of how they can cooperate with the DATE-assistant to solve different tasks.

> *"Yeah, I am not involved much myself, but I am supposed to get some of this training. I am training the system, which also means training myself to train the system." - (Informant 9)*

Nevertheless, the case handlers also do need to train themself to train the DATE-assistant. This is important as the DATE-assistant needs constant feedback. However, the feedback provided by the case handler must be correct, so the DATE-assistant will have positive learning from the feedback and become more accurate in the future predictions.

4.2.5.2 Documentation

Although the developers of the DATE-assistant aim for the DATE-assistant to be intuitive for the case handler, without need for documentation. There are different kinds of documentation regarding the DATE-system; there is some for the DATE-assistant itself, which will help a case handler handling the DATE-assistant. There is one for the case handlers, regarding how to use the information provided by the DATE-assistant. As informant 11 explains the different users of the DATE-system:

> *"Yes, there are many users, it is the main user, and we can talk about 4 user categories here. It is an external customer, who owns a ship or operates a ship. It is our surveyors who are not experts in everything, but they should be happy to be on the boat and check that it is in order and wondering about something and then they ask the case dealers. and they are experts in all these different disciplines we have to keep a boat going. So, there are 30 certificates they need and there are experts in various fields. So, it is the case handlers who get help from this system, so if not, the DATE-assistant was able to answer it to the others right away. Also, there are those who maintain the knowledge, who make sure that the DATE-assistant has some knowledge that should pop up to these people." -(Informant 11)*

But the effort put in to make the documentation is labour intensive and demands a lot of effort from the developers. As informant 11 states, it is quite challenging to produce documentation when they continuously improve the DATE-assistant. Since they [DNV] have the DATE-system in an intermediate situation, without knowing how the final state would look like, they have a challenge with producing updated documentation.

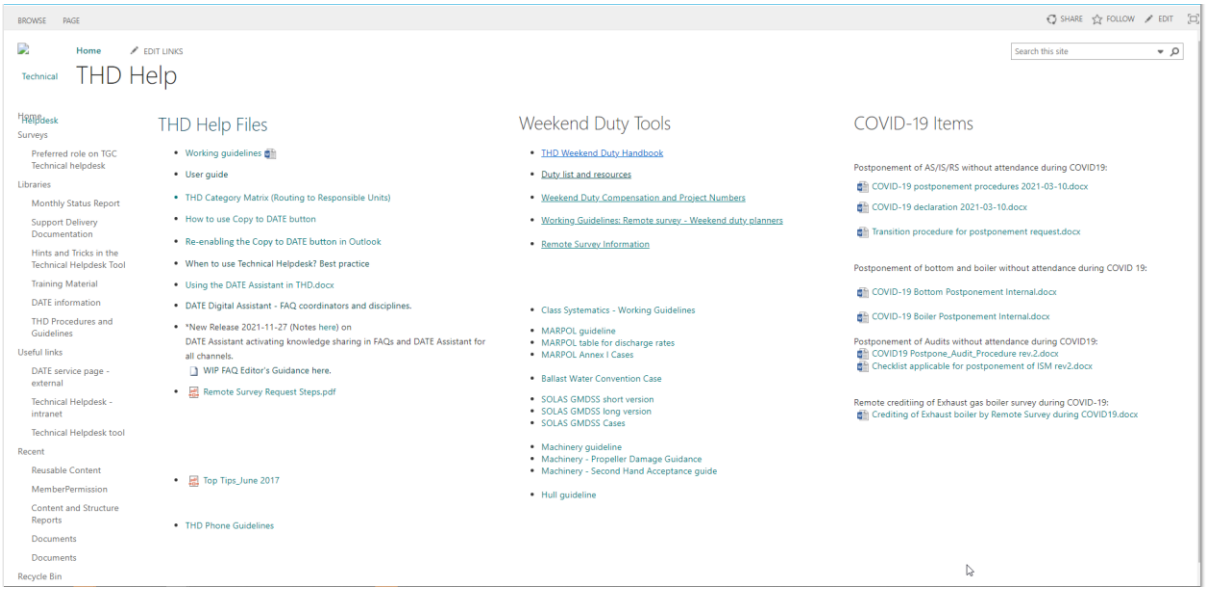
> *"There I can say that we have had, it is challenging when you drive agile development and continuous improvement over time, because you are constantly on the move somewhere, and you do not quite know what the system is going to be. So, it has been absolutely awful with documentation along the way, because we have made a half system to begin with, which works somewhere, but since it does not work everywhere*

then there are workarounds and stories and extra things you have to do along the way simply because you are in an intermediate situation. "-(Informant 11)

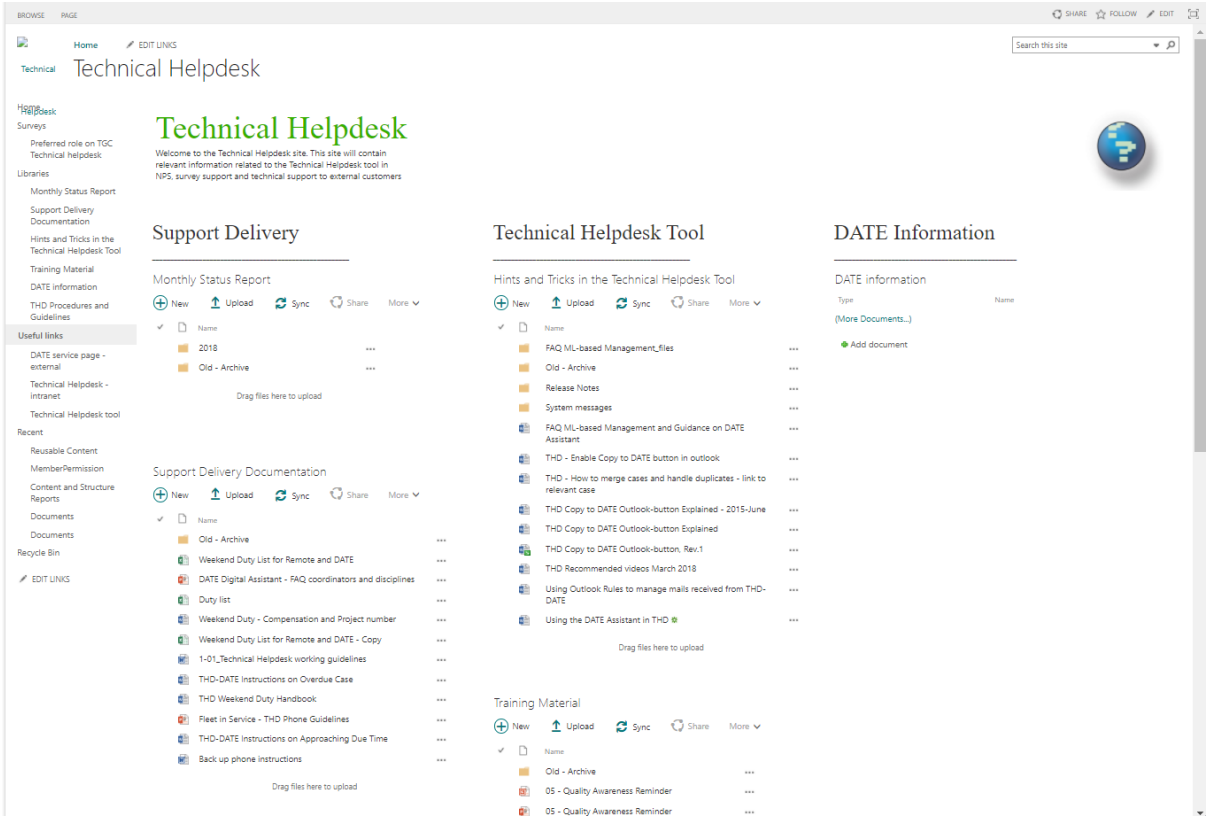
However, the developers want the DATE-assistant to be less complex, so the case handler does not need to know the DATE-assistant in-depth to use it. As of now, the case handler should have extensive knowledge to use and collaborate with the DATE-assistant.

"So now we are trying to reduce the complexity so that things will go more by themselves and happen without people having to know very much." -(Informant 11)

Seen from Picture 6 and Picture 7, there is a lot of documentation produced. Ranging from help files to documentation about the DATE-system.



Picture 6: Documentation produced by DNV



Picture 7: Documentation for the DATE-assistant

4.2.5.3 Frequently asked questions (FAQs)

Frequently asked questions (FAQs) are cases which are stored in the DATE-assistants database over completed cases. These completed cases will appear on the screen of the case handler from the DATE-assistant. Thus, the DATE-assistant is providing knowledge to the case handler, by providing similar cases. At the same time the case handler will also be supplied with a percentage of similarity of the case. Then the case handler can use the old case to answer the customer.

> *"We are trying to make a system [DATE-assistant] where the case handlers are contributing to the knowledge. Since we are talking about a machine learning system where we wish feedback of the predictions and the knowledge the DATE-assistant provided. It [the DATE-assistant] is like a child, you tell it to do something, then it does something, and then correct it along the way. So, we are trying to get the case handlers to give a thumbs up or thumbs down depending on if the information provided was useful or not. Then we do a retraining of the system, with new data, so the predictions*

will be a little different the next day and we have taken consideration of the thumbs down, so the system will have less chance of giving the same information in a similar context in the future." – (Informant 11)

As we can see, the frequently asked questions can be reused if the earlier given feedback is precise and good. Then a case handler with a similar case can use the answers provided in the past. As seen from the informants, the feedback given to the DATE-assistant will make an impact on future predictions. By giving good feedback, the case handler is part of knowledge sharing with the DATE-assistant.

4.2.5.4 Webinars

The webinars for machine learning systems in DNV, usually comes after a deployment of a new configuration of a machine learning system. After the release of the new system, the system experts hold a webinar with the case handlers of the system. During the webinar the system experts demonstrate how the DATE-assistant works, as well as what the case handlers can do and not do in the system.

"When we are deploying these types of systems, do we usually have a webinar with the case handlers, where we are demonstrating the DATE-assistant. During the webinar we are looking into what's happening and what people [the case handlers] don't understand. So, in the time after the webinar, we can go out with more information about the struggles."- (Informant 11)

As informant 11 emphasises, it is not only the demonstration part from the experts that are important during the webinars, but also how the case handlers are using the DATE-assistant and what type of struggles the case handlers have when using the DATE-assistant. This makes the experts of the system aware of what kind of struggles the case handlers have and have the possibility to use this information to push out information regarding the struggles.

"The insight we have is that you have to look at what you can do, to make stuff happen. And to make stuff happen you have to pay attention to the case handlers, and what's happening and not happening."- (Informant 11)

Hence, by paying attention to the case handlers and their behaviour with the DATE-assistant, the developers acquire valuable knowledge about how the case handlers interact with the DATE-assistant. Then, they proceed to use this knowledge in making things happen, so be fixing the user experience (UX) or something else that can improve the DATE-assistant. Following is an example of insight the DNV developers got after they merged with Germanischer Lloyd (GL). The insight gained was that you must pay attention to the case handlers and their interaction with the DATE-assistant.

"Ehh, I don't know if I should use the word proactive about this. It is all about paying attention to the things you have designed and seeing if it's working. If you have put a system out there and then you have a webinar. I did this when we merged our company with GL. At that time, we had two case handler systems [DATE-assistants], and all the people from GL in Germany were supposed to use our system. The German version had a mail system which worked, and the employees were measured in customer satisfaction. So, we migrated the German tool into our tool, and gave them webinars and said "Ok, let's go". The chief of the Germans said everything was good, everything worked, and everybody was satisfied. But when we checked what was happening, none of the German employees used the new system."- (Informant 11)

First, by having all these service documents available whenever, the case handlers have the opportunity to check them if they are uncertain about anything which will make them more secure when interacting with the DATE-assistant. But it is not only the case handler which benefits from these, the DATE-assistant also benefits from these through the FAQs. In this interaction DNV tries to facilitate for a behaviour and mindset as seen under, so both parties in the configuration benefits from the interaction:

"Yes also, if we take the case handlers then, then they contribute, we try to get a system where they contribute to the knowledge as well. And here we are talking about

a ML-system where you want feedback on the predictions and what the DATE-assistant did. Because there it is, it's like a child, you say that the child should do something, then the child does something and then you correct along the way. So here we try to inspire then these case dealers to give thumbs up and thumbs down depending on whether that knowledge was useful or not, and if they then give thumbs down, then we run a new retraining, with new data, so that the prediction is a bit different day No. 2. and has taken into account the thumbs down and should give a little less chance that pops up that knowledge in a similar context in the future. "– (Informant 11)

The last trust facilitating mechanism in the configuration is the production of service documents. The production of service documents consists of training of the DATE-assistant, where the case handlers are in a training phase. In this training phase, the case handlers have the opportunity to directly contact the system experts of the DATE-assistant. Furthermore, the trust facilitating mechanism consists of documentation. There are different kinds of documentation; one regarding the DATE-assistant itself, this helps the case handler handle the DATE-assistant; another concerning the case handlers use the information provided by the DATE-assistant. Webinars are the last part of this trust facilitating mechanism. Here, the system expert holds information webinars for case handlers about the DATE-assistant. All parts of this trust facilitating mechanism helps the case handlers to understand the DATE-assistant, use the DATE-assistant, how to interpret categorisations done by the DATE-assistant, and fill in information gaps when needed. This facilitates trust by helping the case handlers from the first interaction with the DATE-assistant, throughout the iterations of interaction with the DATE-assistant.

Picture 8: Advertisement of the DATE service

4.3 Chapter Summary

In this chapter, I have presented what I have called trust facilitating mechanisms. The trust facilitating mechanisms refers to a process of action done to facilitate trust for the case handler to the DATE-assistant and are my empirical findings. The five trust facilitating mechanisms identified are constantly providing feedback to the system, users getting greater understanding of the system by using the system, involving users, accurate predictions of historically answered cases and producing service documents.

5. Discussion

In this thesis I set out to address the question; *"Which mechanisms facilitate trust in metahuman systems?"*

When discussing an implementation or maintenance of machine learning systems, challenges with configuration of the machine learning systems can occur. In the past, machine

learning systems have executed repeatable tasks, but due to technological development they can now be a part of decision making, if not *making* decisions of urgency. These systems can be seen as metahuman systems, a configuration between the machine learning system and at least one human, which collaborate in a given space to answer specific tasks. This new configuration of systems provides a need to think differently about trusting the systems. Earlier, the most important was the rightness of input data, to ensure the machine learning system could do the exact same as the human agent. Whereas now, the trust between the human agent, seen as the end user in Figure 2, and the machine learning agent's decision making are of importance, combined with the feedback given by the end user to the machine learning agent. Throughout the study I have found five trust-facilitating mechanisms which I will present and discuss.

5.1 Five trust facilitating mechanisms in metahuman systems

In the following section, I will present and discuss the five mechanisms identified. Table 4 gives an overview of the five trust facilitating mechanisms identified in the metahuman systems, a description of how they facilitate trust and examples from the empirical data.

Trust facilitating mechanism	Description	Example from empirical findings
Constantly providing feedback to the system	Feedback loop seen in the configuration of the machine learning agent, where the machnine learning agent gets feedback for every case. Additionally, the machine learning agent gets feedback of proposed similar cases.	End users setting cases as closed or manually giving a case a new category provides feedback to the system. End users giving thumbs up and thumbs down on suggested similar cases.
End users getting a greater understanding of the system by using the system	End users gets an greater understanding of the machine learning agent by interacting with it over time.	End users understanding more of the machine learning agent through interaction with the machine learning agent.
Involving end users	Involving end users in the development of new metahuman systems, where the end user will use it after implementation.	End users involved in the design of new metahuman systems, by answering questions from the developers.
Accurate predictions on historically answered cases	Cases that has been around for awhile and been exposed to the machine learning agent so it has high accuracy on these.	Cases the machine learning agent has been exposed to for a long time, where the accuracy of the categorisation is high.
Producing service documentation	Producing service documents so the end users gets an understanding of the machine learning agent at first. The end users also have the possbility to check it later.	Produced service documentation in DNV, consists of webinars, FAQs, documentnation and training of the system.

Table 4: Summary of the trust facilitating mechanisms

5.1.1 Constantly providing feedback to the system

The first mechanism I identify is *constantly providing feedback to the system*. When facilitating trust in metahuman systems, the developers should consider how one could constantly provide feedback to the machine learning agent. By doing so the developers design for a feedback loop, where the machine learning agent, called DATE-assistant in the case, gets gradually better for each iteration of training as continuous feedback is forwarded from the case handler. These new configurations of metahuman systems are expected to have a positive impact on organisations, due to either superior performance or improved efficiency (Brynjolfsson & McAfee 2014, Davenport & Kirby 2016). But as emphasised in the existing literature, getting to the stage where the metahuman systems outperforms humans and machine learning agents alone, requires a conception of the technology as a teammate rather than a tool (Wiethof, Tavanapour, & Bittner, 2021). When humans have a teammate-approach, it is vital that they "make corrections and improve the agent '' so the machine learning agent is provided with feedback, and can be trained when doing so (Wiethof, Tavanapour, & Bittner, 2021). This training is happening over time and through iterations. After being through huge amounts of training the machine learning agent will use the feedback to become sufficient in the categorisation.

5.1.2 Users getting a greater understanding of the system by using the system

The second mechanism I identify is the *users getting a greater understanding of the system by using the system*. When developing configurations of metahuman systems, the trust facilitating mechanism of users getting a greater understanding of the system by using the system should be present. A challenge identified both in literature and in my findings relates to the difficulty to understand the reasoning behind an outcome when the reasoning is hidden from view (Flyverbom, Leonardi, Stohl, & Stohl, 2016). Without the knowledge of how the agent reasoned from A to B, the human agent can be critical and do not trust the outcome as the element of trust is lacking. This lack of trust can be established through

using the agent and getting familiar with how the machine learning agent has worked in the past. Having good experiences facilitates further development of trust. An important consideration is also trust facilitating trust, as if one human agent outspokenly has trust to the machine learning agent, another human agent can trust the validation and judgement of the machine learning agent without much experience using it if they trust the other human agent. One good experience can be enough to validate the already existing positive bias, without a comprehensive trust building process by using the system. And vice versa; one human agent can outspokenly not have trust to the judgement the machine learning agent provides, which can affect other agents, as their trust towards the human agent is greater than the trust towards the machine learning agent. One bad experience with the machine learning agent can therefore be enough to delay, or even ruin, the trust building process between human and machine.

Another struggle which is seen in both the literature and this case, regards the development of trust between a human and AI. From my empirical findings, one of my informants' states that he was sceptical when introduced to the DATE-assistant. However, over time, this scepticism has changed into trust. Glikson and Wolley also emphasise this trust-transformation in their article; the trust trajectory between a human and robotic AI is similar to human relationships – building trust takes time. Whereas the trust might start out low, time will provide hands-on experience, and consequently trust increases (Glikson & Wolley, 2020).

5.1.3 Involving users

The third mechanism identified is *involving the users.* The trust facilitating mechanism of involving the end users in the design of the system, is historically done in user centred design (UCD) and participatory design (PD). In PD the researcher "*explores conditions for user participation in the design and introduction of computer-based systems at work".* Hence, the approach of PD, includes users both in the design and the introduction of the system developed. Whereas in UCD, the whole design process of an artefact or system is evolved around the users demands and needs.

UCD is a design approach, initially introduced by Norman and Draper (1986). The essence of this approach emphasises how the user of the software and the needs of the user must be considered when developing new technology (Norman & Draper, 1986). This underlines the notion that the "*purpose of the system is to serve the user"* (Norman & Draper, 1986), which is central to UCD. As seen from Norman and Draper about UCD, the software is perceived as a tool for the end user, rather than a contributor. Where Wiethof, Tavanapour, & Bittner argues that to achieve the synergy of working together with the machine learning agent, the end user must accept the machine learning agent. This is distinct from the UCD approach, since metahuman systems are made for a synergy of work, and the end user and the machine learning agent should complete each other. Karat adds to the general understanding of UCD;" *For me, UCD defines an iterative process whose goal is the development of usable systems*" (Karat, 1997). This is in line with the involvement of users in the study. The involvement of users is to make them understand the machine learning agent, the limitations of the agent, and how to interpret the results from the machine learning agent. The purpose of this is to make the machine learning agent usable for the end user, which in the case with DNV is the case handlers.

Another point mentioned in the literature and seen through the analysis of my case, is the trust trajectory for virtual embedded AI. This trust trajectory usually starts out high, but then drops as a result of experience with the machine learning agent (Glikson & Wolley, 2020). Since this struggle is seen in the literature, the developers try to avoid this type of trust trajectory, by involving the users in the development, whereas the involvement of the users is to give them realistic expectations of the machine learning agents capabilities.

5.1.4 Accurate predictions of historically answered questions

The fourth mechanism I identify, is *the accuracy of historically answered questions.* This mechanism is dependent on the learning of the machine learning agent in the metahuman system. From Davenport and Kirby, we can get an understanding of the importance of learning as a feature concerning metahuman systems, as "machines that learn as parts of wider systems where *both* humans *and* machines learn jointly" (Davenport & Kirby, 2016).

From my empirical findings, I discovered there are some cases existing for a long time. These cases are well documented in the database, and the DATE-assistant has thorough training on these cases. During this phase of learning, the DATE-assistant has received feedback in the form of trial and error from the case handler. This simplified definition of learning is mentioned by Thorndike: "[..] learning can be first simplified to the observation that any agent's learning involves some sort of trial and error" (Thorndike, 1932). Lyytinen, Nickerson, and King add to this: "As it moves through this process it acquires new capabilities that make it better fit operating in that environment" (Lyytinen, Nickerson, & King, 2021). Emphasised in the related literature, the machine learning agent is dependent on the collaborating human agent, as the learning done through the trial-and-error phase will make the machine learning agent acquire new capabilities, which will improve it further to better fit in the task solving environment (Wiethof, Tavanapour, & Bittner, 2021, Lyytinen, Nickerson, & King, 2021, Baird & Maruping, 2021).

Although the machine learning agent is improving, the rules and regulations in the industries are dynamically changing. This makes continuous feedback and learning essential, as the machine learning agent's perception of right and wrong is affected by the changes of rules and regulations.

5.1.5 Producing service documents

The fifth mechanism I identify is *producing service documents.* When facilitating trust in metahuman systems, organisations should consider how one can produce service documents to serve end users. A persistent challenge I found in my empirical findings, was that the developers of metahuman systems in DNV struggled to produce service documentation which is up to date. This challenge is also found in the literature, where Glikson and Woolley describes the lack of explanations to why a decision is made are of concern within the ML tools (Glikson & Wolley, 2020). The same concerns are highlighted by Leonardi and Treem, where the output presented to users is given minimal transparency into how the output is generated (Leonardi & Treem, 2020). As seen from the case with DNV, the end goal is an intuitive DATE-assistant. Hence, it is possible for the case handler to interact

with the DATE-assistant in ease. At the same time the need for production of service documents is present.

This illustrates that the production of service documents is desirable, and concerning the empirical findings, the case handlers have the possibility of using the service documents when confronted with an output with minimal transparency. Even though the service documents can enlighten parts of the “black-box” of the DATE-assistant's decision making, Leonardi and Treem argue further that even if all data and logic is made accessible there is no guarantee that the output generated is transparent (Leonardi & Treem, 2020).

5.2 Extending current knowledge on trust facilitating mechanisms in metahuman systems

As seen from the analysis, these trust facilitating mechanisms can be overlapping. One example is the production of service documents and the users getting a greater understanding of the system by using it, where the case handlers always have the possibility of getting to know the system by taking advantage of the already produced service documents.

Yet an example is the involvement of the users and constantly giving feedback to the system, where the users are involved in the design of the metahuman system, additionally to contribute during the lifecycle of the metahuman system by constantly providing feedback to the machine learning agent. This overlap is shown in Figure 4.

The trust facilitating mechanism of involving users complements the trust facilitating mechanism of constantly providing the machine learning agent with feedback, by the users being involved in the knowledge sharing, and improvement of the machine learning agent. At the same time, the developers would have time to get an understanding of the needs and context of use of the users. The involvement of the users is historically done within the UCD approach, as the user's needs are important. Additionally, it is important that the user changes their view of the machine learning agent, transforming from a tool to a teammate (Wiethof, Tavanapour, & Bittner, 2021). The literature suggests that when the view of the

humans' changes, the machine learning agent is dependent on the human user, since the feedback the machine learning agent gets from the human is invaluable (Davenport & Kirby, 2016).

Another example of the trust facilitating mechanisms that complement each other, is the trust facilitating mechanism of producing service documents and the users getting a greater understanding of the system by using the system. As my analysis illustrates, the informants start out with low trust in the interaction with the DATE-assistant. But as the informant gets hands-on experience with the DATE-assistant, the informants trust trajectory increases (Glikson & Wolley, 2020). To contribute to the experience gained over time, the developers of the DATE-assistant produce service documents. Hence, the users have documentation to support them during the interaction, when there are knowledge gaps present.

Lastly, the trust facilitating mechanism of accurate predictions of historically answered cases, is a result of DATE-assistant constantly getting feedback from the users. Thus, the predictions of similar cases to those that have been answered a lot in a historical perspective, is perceived as high accuracy. This is a direct result of the feedback given by the users, as it is possible to see the importance of the continuous process of feedback given to the DATE-assistant.

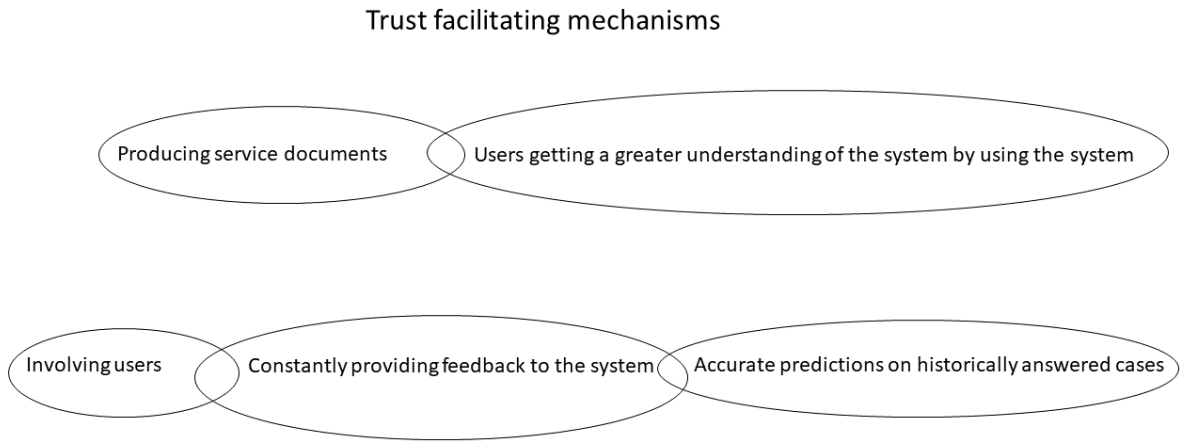


Figure 4: Illustration of overlap between the trust facilitating mechanisms

5.3 Contribution

The contribution of this thesis is both practical and theoretical. I relate the theoretical contribution to the literature concerning trust in metahuman systems. In my chapter of related research, I noted that prior literature has highlighted a series of characteristics of metahuman systems. Lyytinen, Nickerson, and King (2021) introduced the concept of metahuman system as a system where humans and machines amplify their capabilities and potentially perform better than they would do separately. Where Lyytinen, Nickerson, and King have highlighted the learning aspect of metahuman systems, I have investigated the trust aspect of metahuman systems. Foremost, I offer a practical contribution to the developers in DNV, addressing the trust facilitating mechanisms for the case handler to the DATE-assistant detected through an iterative process of problem formulation from my engaged research project. By identifying the trust facilitating mechanisms within their metahuman system, I provide DNV with insight on how each of the mechanisms contribute to facilitating trust in the metahuman system.

Secondly, emphasised by Lyytinen, Nickerson, and King, it is typical to observe metahuman systems regarding the learning fragment of the system (Lyytinen, Nickerson, &

King, 2021). This concerns the model used in implementation of the machine learning systems, for example if it is used linear regression instead of supervised learning, as these systems are getting better and are set to do more complicated tasks as seen earlier. As Baird and Maruping (2021) suggest through their framework, the systems can be reflexive, which means they are react on the input given by the user. For example, they can work as a virtual assistant of some type, such as a voice-based assistant. Another configuration is supervisory systems, which is a control system, where the system serves as a behaviour modification, and therefore is a decision support for the user. This supervisory system can for example help the user with trade suggestions.

The third suggestion mentioned by Baird and Maruping, concerns that machine learning systems can be anticipatory systems, which means being proactive systems that "proactively applies models based "reasoning" to anticipate needs or wants". E.g., being a smart watch, which anticipates the need for sleep of the user. The last one is prescriptive systems. These systems or agents act as substitutes for behaviour-based decision-making or outcome-based decision-making by prescribing or acting. Common for all these configurations of metahuman systems are the interaction and cooperation with humans, where humans and machine learning agents work hand-in-hand to accomplish tasks. Hence Baird and Marupings framework, there are a lot of everyday systems that have a machine learning agent and a human who cooperates. Complemented by Wiethof, Tavanapour, and Bittner (2021) findings, there are a lot of users considering these machine learning agents as tools. This is present in my findings and the literature, where the complex dynamics of this cooperation continuously challenge the users in how they relate to the machine learning agent. Wiethof, Tavanapour, and Bittner (2021) further argue that to achieve a synergy where the human agent and the machine learning agent cooperate, the human agent is required to accept the machine learning agent as a teammate more than a tool. This is a shift in the acknowledgement of technology. I argue that this shift in mindset and acknowledgement of the machine learning agent conceptualise the best capabilities to why metahuman systems are developed.

As seen in Figure 4, to facilitate trust between the human agent and the machine learning agent, the human agent needs to get to know the system by collaborating with it. By using the system, the end user will get a greater understanding of the machine learning agent.

This is represented in Wiethof et als. article, where they write, *"Next, researchers consider the aspect of transparency fostering the understanding of an agent, its behavior and purpose to accept it as a teammate. This allows its human teammates to still criticize and improve it, which eventually ensures a certain feeling of control as well as an enhancement of the group process and its outcomes "(*Wiethof, Tavanapour, & Bittner, 2021). Additionally, when the human agent gets a greater understanding of the machine learning agent and what to expect from the machine learning agent, they will also get the capability of improving the machine learning agent. This is to ensure a human agent with a feeling of control. This is represented in the thesis where the human agent always has the possibility of giving feedback to the machine learning agent. This is also argued in the literature by Wiethof et al., achieving a synergy consisting of cooperation and learning between the human agent and the machine learning agent, the human agent needs to accept the machine learning agent and be willing to learn from it at the same time making corrections and improving the agent (Wiethof, Tavanapour, & Bittner, 2021). I identified this through my analysis of the empirical findings, where we can see that the end user collaborates with the DATE-system, and the end user has the possibility to give the DATE-assistant constant feedback, which will make the machine learning agent train on the new data, and then gradually improve. But as one of my interview subjects established, this is something they do more sporadically. This can lead to the opacity of the outcome from the machine learning agent, where *"Opacity refers to the difficulty to understand the reasoning behind a given outcome when such reasoning is obscured or hidden from view"* (Flyverbom, Leonardi, Stohl, & Stohl, 2016). When the end user struggles to understand the reasoning behind the prediction from the machine learning agent, it will be difficult for the end user to collaborate with the machine learning agent. As shown by Lebovitz, Lifshitz-Assaf and Levina, *"[...] whereby human experts and AI technologies work together to accomplish a task. The word augmentation is defined as a process of enlargement or making something grander or more superior*" (Lebovitz, Lifshitz-Assaf, & Levina, 2022). The meaning of these configurations of metahuman systems is to make the system superior to the human agent and the machine learning agent alone.

Table 5 summarises my contribution concerning the challenges experienced by DNVs developers in the mechanisms of facilitating trust in metahuman systems, as well as challenges highlighted in related literature.

Practical and theoretical challenge		Trust facilitating mechanisms	How it facilitates trust in metahuman systems
Challenge from existing litterature	To achieve the synergy of work between humans and ML-algorithms, the needs to corrections and imprive the algorithm (Wiethof, Tavanapour, & Bittner, 2021).	**Mechanism 1:** Constantly providing feedback to the system	The DATE-assistant gets better through the feedback provided by the case handler, it facilitates the case handlers trust in the DATE-assistant, since the case handlers perceive the information from the DATE-assistant as correct.
Challenge in DNV	DNVs developers struggle to get the case handlers to provide the DATE-assistant with feedback in every interaction.		
Challenge from existing litterature	The trust trajectory is similar to human relationships, where it starts out low and increases following hands-on experience (Glikson & Wolley, 2020).	**Mechanism 2:** Users getting a greater understanding of the system by using the system	The case handlers gets an understanding of the DATE-assistant by using it over time, which makes the case handlers trust it.
Challenge in DNV	DNVs developers struggle to get the case handlers to understand the DATE-assistant in the start of using it.		
Challenge from existing litterature	The center of this UCD is the user of the software, and the needs of the user must be considered when developing new technology (Norman & Draper, 1986).	**Mechanism 3:** Involving users	Involving case handlers in the design of new configurations of metahuman systems, where the developers try to understand the needs and expectations of the case handlers.
Challenge in DNV	DNVs developers struggle to get the case handlers to have realistic expectations of what the DATE-assistant can do.		
Challenge from existing litterature	Learning is an important characteristic in metahuman systems, where «machines that learn as parts of wider systems where *both* humans *and* machines learn jointly» (Davenport & Kirby).	**Mechanism 4:** Accurate predictions on historically answered cases	The DATE-assistant handles historically answered cases accurate, which helps the case handler trust it.
Challenge in DNV	The case handlers struggle to trust the DATE-assistants categorisation when introduced with new cases.		
Challenge from existing litterature	Today, the predictions made by a ML-algorithm is presented with minimal transparency into how the ML-algorithm reached it (Leonardi & Treem, 2020).	**Mechanism 5:** Producing service documents	Helping the case handlers from their first interaction with the DATE-assistant throughout the iterations of interaction by filling in knowledge gaps.
Challenge in DNV	The DNVs developers struggle to produce updated service documents.		

Table 5: Summary of contributions

5.4 Limitations

During my attempt to understand this social phenomenon through intersubjective meanings, thoughts, and experiences of my informants, I could not reach a complete subjectivity, as my biases, cultural- and social discourses, and general subjectivity will affect my cognitive work during the process of collecting and analysing empirical data. I must also acknowledge that my study is difficult to replicate, as my access to DNV, generating data, and my findings based on interpretations is limited to my subjectivity. However, in line with the tradition of interpretative research, this is not the goal either. As noted by Flyvbjerg, the view that one cannot generalise and make a valuable contribution based on a single case study is common. However, critiquing this view as narrow, he continues: *"That knowledge cannot be formally generalized does not mean that it cannot enter into the collective process of knowledge accumulation in a given field or in a society".* (Flyvbjerg, 2006). I argue for the relevance of the mechanisms beyond the case with DNV, since the mechanisms are identified through a combined effort of analysing my empirical findings, as well as discussions with cases and topics from relevant literature. To establish credibility of the results, I have strived to provide a rich description of the methods for data collection and analysis, as well as my findings to enable other researchers to follow the arguments that have led me to my conclusion and contribution.

Other limitations may be found in how I conducted my research project. I have utilised various forms of data gathering activities, where the primary source of information has been interviews with DNV employees. I have enriched my understanding with documents analysis. Still, my study could benefit from having conducted observations of the interaction between the case handler, the DATE-assistant, and the customers. One example of this is how the case handler is cooperating with the DATE-assistant in helping the customers. Conducting observations of this interaction could provide me with a richer understanding of the trust facilitating mechanisms in DNVs metahuman systems.

5.5 Further research

Based on the findings of my study I suggest some avenues for further research. First, my study is limited to examining the trust facilitating mechanisms of one organisation within a configuration of metahuman systems. Conducting a single case-study like this, the characteristics of DNVs employers has shaped the findings. I have shown how one vendor has developed a configuration of a metahuman system, where I have identified five trust facilitating mechanisms. One of the avenues for further research would be to investigate the relevance and applicability of the five trust facilitating mechanisms beyond the case with DNV.

Beyond this, it would be interesting to further investigate the dynamic between the case handler and the DATE-assistant where the focus would be on how the case handler is interacting and collaborating with it.

6. Conclusion

This thesis has explored which mechanisms facilitate trust in configurations of metahuman systems. This was explored through a case study done in collaboration with DNV. Through a yearlong case research project, I collaborated with employees in DNV by virtual means, gaining a richer understanding and insight into their practices as well as challenges to trusting the machine learning agent studied in the configuration of metahuman systems.

The five trust facilitating mechanisms were identified through analysing my empirical findings, therefore the main theoretical contribution of this thesis is the five trust facilitating mechanisms in a configuration of a metahuman system. These trust facilitating mechanisms are:

- Constantly providing feedback to the system
- Users getting a greater understanding of the system by using the system
- Involving users
- Accurate predictions of historically answered cases
- Producing service documents

These five trusts facilitating mechanisms can be leveraged by other vendors in a setting where a metahuman system is identified or developed, giving some guidance as to how to construct trust between a human agent and a collaborating machine learning agent.

Bibliography

Arrieta, A. B., Natalia, D.-R., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion*, 82-115.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. *Journal of the Association for Information Systems*, 325-352.

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation To and From Agentic IS Artifacts. *MIS Quarterly 45 (1)*, 315-341.

Baptista, J., Stein, M.-K., Klein, S., Watson-Manheim, M. B., & Lee, J. (2020). Digital work and organisational transformation: Emergent Digital/Human work configurations in modern organisations. *Journal of Strategic Information Systems*, 1-10.

Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Artificial Intelligence in Organizations: Current State and Future Opportunities. *MIS Quarterly Executive: 19 (4)*, 1-15.

Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems 22 (2)*, 281-303.

Berente, N., Gu, B., Recker, J., & Santanom, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, 3-41.

Blois, K. J. (1999). Trust in Business to Business Relationships: An Evalutaion of its Status. *Journal of Management Studies*, 198-215.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 77-101.

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* W.W. Norton & Company.

Christensen, T., & Lægreid, P. (2020). ICT Use in Central Government: Scope, Predictors and Effects on Coordination Quality. *International Journal of Public Administration*, 273-286.

Crang, M., & Cook, I. (2007). *Doing Ethnographies.* SAGE Publications Ltd.

Davenport, T. H., & Kirby, J. (2016). Just How Smart Are Smart Machines. *MIT Sloan Management Review*, 21-25.

Derrick, D. C., Seeber, I., Elson, J. S., & Waizenegger, L. (2021). Collaboration with Intelligent Systems: Machines as Teammates. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 358-359.

Dignum, V. (2019). *Responsible Artificial Intelligence.* Springer Verlag.

DNV. (2021, 06). *DATE-Direct Access to Technical Experts*. Retrieved from dnv.com: https://www.dnv.com/maritime/date/index.html

Du, M., Liu, N., & Hu, X. (2020). Techniques for Interpretable Machine Learning. *Communications of the ACM 63 (1)*, 68-77.

European Commission. (2019, April 08). *Ethics guidelines for trustworthy AI*. Retrieved from European Union: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and Organizing in the Age of the Learning Algorithm. *Information and Organization 28 (1)*, 62-70.

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities & Social Sciences Communications*, 1-9.

Flyvbjerg, B. (2006). Five Misunderstandings About Case-Study Research. *Qualitative Inquiry*, 219-245.

Flyverbom, M., Leonardi, P. M., Stohl, C., & Stohl, M. (2016). The Management of Visbilities in the Digital Age. *International Journal of Communication*, 98-109.

Friman, M., Gärling, T., Millett, B., Mattsson, J., & Johnston, R. (2002). An analysis of international business-to-business relationships based on the Commitment-Trust theory. *Industrial Marketing Management*, 403-409.

Glikson, E., & Wolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals Vol. 14 No. 2*, 627-660.

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, 1-16.

Hinings, B., Gegenhuber, T., & Greenwood, R. (2018). Digital innovation and transformation: An institutional perspective. *Information and Organizations*, 52-61.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., . . . Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *Conference on Fairness, Accountability, and Transparancy*, 560-575.

Huysman, M. (2020). Information systems research on artificial intelligence and work: A commentary on "Robo-Apocalypse cancelled? Reframing the automation and future of work debate". *Journal of Information Technology 35 (4)*, 307-309.

Jarvenpaa, S. L., Möhlmann, M., & Blomqvist, K. (2021). Advances in Trust Research: Artificial Intelligence in Organizations Mini Track. *Conference on System Sciences (HICSS)*, (pp. 5459-5461). Hawaii.

Jonsson, K., Mathiassen, L., & Holmström, J. (2018). Representation and mediation in digitalized work: evidence from maintenance of mining machinery. *Journal of Information Technology*, 216-232.

Karat, J. (1997). Evolving the scope of user-centered design. *Communications of the ACM 40 (7)*, 33-38.

Kensing, F., & Blomberg, J. (1998). Participatory Design: Issues and Concern. *Computer Supported Cooperative Work*, 167-185.

Kirsten, M. (2018). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* , 835-850.

Lai, Y., Kankanhalli, A., & Ong, D. C. (2021). Human-AI Collaboration in Healthcare: A Review and Research Agenda. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 390-399.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction.* Morgan Kaufmann.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science 33*, 1-23.

Leonardi, P. M., & Treem, J. W. (2020). Behavioral Visbility: A New Paradigm for Organization Studies in the Age of Digitization, Digitalization, and Datafication. *Organization Studies*, 1601-1625.

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. *Proceedings of the 54th Hawaii International Conference on System Sciences* , 63-72.

Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems = humans + machines that learn. *Journal of Information Technology 36 (4)*, 427-445.

Mikalsen, M., & Monteiro, E. (2021). Acting with Inherently Uncertain Data: Practices of Data-Centric Knowing. *Journal of the Association for Information Systems*, 1715-1735.

Møller, N. H., Shklovski, I., & Hildebrandt, T. T. (2020). Shifting Concepts of Value: Designing Algorithmic Decision-Support Systems for Public Services . *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (p. 12). Tallin: Association for Computing Machinery.

Myers, M. D. (1997). Qualitative Research in Information Systems. *MIS Quarterly (21:2)*, 241-242.

Myers, M. D., & Klein, H. K. (2011). A set of princinples for conducting critical research in information systems. *MIS quarterly*, 17-36.

Nissen, A., & Jahn, K. (2021). Between Anthropomorphism, Trust, and the Uncanny Valley: a Dual-Processing Perspective on Perceived Trustworthiness and Its Mediating Effects on Use Intentions of Social Robots. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 360-369.

Norman, D. A., & Draper, S. W. (1986). *User Centered System Design New Perspectives on Human-computer Interaction.* CRC Press.

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research*, 1-28.

Østerlie, T., & Monteiro, E. (2020). Digital Sand: The Becoming of Digital Representations. 1-41.

Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2020). Make way for the algorithms: Symbolic Actions and Change in a Regime of Knowing. *Organization Science 32(1)*, 18-41.

Petersen, A. C., Christensen, L. R., & Hildebrandt, T. T. (2020). The Role of Discretion in the Age of Automation. *Computer Supported Cooperative Work (CSCW) 29*, 303-333.

Randall, D., Harper, R., & Rouncefield, M. (2007). Ethnography and How to Do It. *Fieldwork for design: Theory and Practice*, 169-197.

Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 1-12.

Schelbe, B., Flathmann, C., Canonico, L.-B., & Mcneese, N. (2021). Understanding Human-AI Cooperation Through Game-Theory and Reinforcement Learning Models. *Proceedings of the 54th Hawaii International Conference on System Sciences* (p. 10). Hawaii: HICSS.

Stake, R. E. (2005). Qualitative Case Studies. In N. K. Denzin, & Y. S. Lincoln, *The SAGE Handbook of Qualitative Research* (pp. 443-466). Sage Publications.

Thorndike, E. L. (1932). *The fundamentals of learning.* Teachers College Bureau of Publications.

Verne, G., & Bratteteig, T. (2018). Inquiry when doing research and design: wearing two hats. 89-106.

Walsham, G. (2006). Doing interpretive research. *European journal of information systems*, 320-330.

Wiethof, C., Tavanapour, N., & Bittner, E. A. (2021). Implementing an Intelligent Collaborative Agent as Teammate in Collaborative Writing: toward a Synergy of Humans and AI. *Proceedings of the 54th Hawaii International Conference on System Sciences* , 400-409.

Willcocks, L. (2020). Robo-Apocalypse cancelled? Reframing the automation and future of work debate. *Journal of Information Technology 35 (4)*, 286-302.

Willcocks, L. (2021). Robo-Apocalypse? Response and outlook on the post-COVID-19 future of work. *Journal of Information Technology 36 (2)*, 188-194.

Zhang, Z., Yoo, Y., & Lyytinen, K. L. (2021). The Unknowability of Autonomous Tools and the Liminal Experience of Their Use. *Forthcoming in Information Systems Research*, 1192-1213.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology 30*, 75-89.

Appendix A

Intervjuguide Andreas (forsker innen AI/ML og risiko) og Frank

Oppvarming:

Bakgrunn for intervjuet aka forklare masteroppgaven og fokuset mitt innen AI/ML
Dele ut samtykkeskjema, fortelle om lydopptak og pseudonym

Innledning:

Introduksjon av intervjuobjekt, hvem er du?
Hvilken bakgrunn har du?
Hva jobber du med nå?
Hva er typiske arbeidsoppgaver for deg?

Hoveddel:

Du har nevnt at du jobber med AI/ML og risiko, hva slags risikoer er det snakk om?
Hvordan kan man eventuelt unngå disse risikoene?
Hva kan skje hvis man ikke tar hånd om disse risikoene?
Er det kun teknologiske risikoer eller er det menneskelige/samfunnskritiske?
Hvilke type AI/ML systemer vil du si at du jobber med?

- -Er det responsible AI?
- -

Hva tenker du er de største fallgruvene/risikoene når man lager systemer med AI/ML?

Har du noen strategier for å unngå risiko i AI?
Hvordan har dere kommet frem til disse strategiene?

Hvis du skulle tilegnet deg tillit til ditt AI/ML system, hvordan hadde du gått frem? Hvordan hadde prosessen sett ut?

Kan man bruke disse strategiene til å tilegne seg tillit hos en bruker/kunde?

Avrunding

Er det noe du tenker vi bør snakke om eller noe som du ønsker å legge til?

Er det noe du tenker jeg har glemt å spørre om, som kan være relevant for temaet?

Er det noen du tenker det kunne vært lurt å snakke med?

Avslutning

Tusen takk for at du ble med på intervjuet :D

Appendix B

Mini Conference in DNV

Time	Title	Speaker
0800	Welcome and introduction	Pierre C Sames, DNV
0810	Trustworthy industrial AI-systems and the ADA framework	Asun St.Clair, DNV
	Developing AI-enabled systems	
0830	Validating AI-systems for service enhancements	Michael Chen, DNV
0850	AI-based innovations in marine and industrial sectors: functional potentials and certification perspectives	Prof Rudolf Mester, NTNU
0920	Discussion	
0930	Break	
	Approving and commissioning AI-enabled systems	
0940	Assuring autonomous ship technology	Øystein Engelhardtsen, DNV
1000	AI testing: A novel Bowtie risk management and assurance framework for machine learning systems	Prof Yanwei Fu, Fudan University
1030	Discussion	
1040	Break	
	Operating AI-enabled systems	
1050	Assurance during operations: Combining physics and data-driven models	Simen Eldevik, DNV
1110	Robustness and Sensitivity of AI systems - Two Sides of a Coin	Prof Yin Shen, NTNU
1140	Discussion	
1150	Conclusions	Pierre C Sames
1200	end	

Splitting the definitions on regard of the model used

Governance of AI: Trust and ethics:

- EUs Trustworthy AI Guidelines
- ISO/IEC JTC 1/SC 42

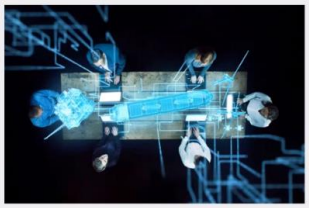
Ethics and trust: Cannot be an afterthought

Ethics in design:

- Anticipating consequences

Ethics by design:

- Behaviour of AI systems



Ethics for Design(ers):

- Integrity of all actors in research and implementation processes

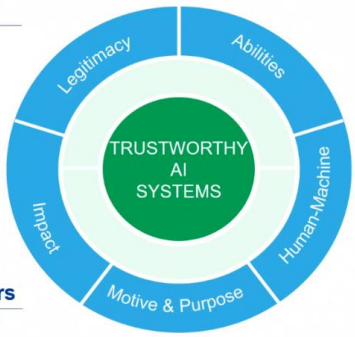
Characteristics of Trustworthy AI Systems

Legitimacy

- Data quality
- Suitability
- Performance
- Risk management & Safety

Transparent impact on stakeholders

- Impact assessment
- Responsibility ascription
- Continuous monitoring



Clearly defined purpose

- Disclosure of goals
- Benefits to stakeholders
- Corporate accountability

Ability to perform and capacity to verify

- Design quality
- Data preparation & Model training
- Testing & Simulation
- Explainability
- Evidence

Understanding contexts:
Human-machine interdependency

- Agents in the AI system
- Roles
- Communication
- Machine-to-machine Interfaces

Appendix C

Vil du delta i forskningsprosjektet

” Which processes are needed for customers to trust AI/ML systems”?

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å finne ut hvilke strategier/prosesser som må til for at kunder skal kunne stole på AI/ML systemer. I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål
Formålet med prosjektet er et studentprosjekt(masteroppgave) der studenten skal finne ut hvordan man kan implementere AI/ML systemer som kunder og brukere stoler på.

Hvem er ansvarlig for forskningsprosjektet?
Universitetet i Oslo er ansvarlig for prosjektet.

Hvorfor får du spørsmål om å delta?
Utvalget kommer fra DNVs egen ML avdeling, samt personer som er blitt pekt på som interessante etter kontakt med andre deltakende personer i prosjektet.

Hva innebærer det for deg å delta?
I prosjektet kommer jeg til å gjennomføre intervjuer, der opplysningene som samles inn handler om dine arbeidsoppgaver med AI/ML systemer er fokus. Opplysningene som blir gitt til meg under intervjuet vil bli tatt opp gjennom et lydopptak og deretter transkribert over til et dokument, der personen som deltar vil bli anonymisert.

- Hvis du velger å delta i prosjektet, innebærer det at du blir med på et intervju. Intervjuet vil vare i ca 30-45 minutter. Intervjuet inneholder spørsmål om hvordan kunder av dere(DNV) kan ha tillit til deres AI/ML systemer. Hvilke prosesser dere har i bunn når dere utvikler tillitsfulle AI/ML systemer.

Det er frivillig å delta
Det er frivillig å delta i prosjektet. Hvis du velger å delta, kan du når som helst trekke samtykket tilbake uten å oppgi noen grunn. Alle dine personopplysninger vil da bli slettet. Det vil ikke ha noen negative konsekvenser for deg hvis du ikke vil delta eller senere velger å trekke deg.

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger

Vi vil bare bruke opplysningene om deg til formålene vi har fortalt om i dette skrivet. Vi behandler opplysningene konfidensielt og i samsvar med personvernregelverket.

- Den som vil ha tilgang ved UiO er Stian Grimsrud, som er studenten som utfører undersøkelsen
- For å best sikre at uvedkommende ikke får tak i personopplysninger vil f.eks. navnet ditt bli byttet ut med en kode/pseudonym som lagres på en egen navneliste adskilt fra øvrig data, samt lagre data på sikker server.

Hva skjer med opplysningene dine når vi avslutter forskningsprosjektet?

Opplysningene anonymiseres når prosjektet avsluttes/oppgaven er godkjent, noe som etter planen er rundt slutten av juni 2022. Etter endt prosjekt, vil personopplysninger og lydopptak slettes.

Hva gir oss rett til å behandle personopplysninger om deg?

Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra Universitetet i Oslo har NSD – Norsk senter for forskningsdata AS vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

Dine rettigheter

Så lenge du kan identifiseres i datamaterialet, har du rett til:

- innsyn i hvilke opplysninger vi behandler om deg, og å få utlevert en kopi av opplysningene
- å få rettet opplysninger om deg som er feil eller misvisende
- å få slettet personopplysninger om deg
- å sende klage til Datatilsynet om behandlingen av dine personopplysninger

Hvis du har spørsmål til studien, eller ønsker å vite mer om eller benytte deg av dine rettigheter, ta kontakt med:

- Universitetet i Oslo ved Stian Grimsrud(tlf: +47 97480220, epost: stiangri@uio.no) eller veileder Alexander Kempton (epost: alexansk@ifi.uio.no).
- Vårt personvernombud: Roger Markgraf-Bye kan nås på epost (personvernombud@uio.no)

Hvis du har spørsmål knyttet til NSD sin vurdering av prosjektet, kan du ta kontakt med:

- NSD – Norsk senter for forskningsdata AS på epost (personverntjenester@nsd.no) eller på telefon: 55 58 21 17.

Med vennlig hilsen

Alexander Kempton
(Forsker/veileder)

Stian Grimsrud
(student)

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet «Which processes are needed to get customers to trust AI/ML systems?», og har fått anledning til å stille spørsmål. Jeg samtykker til:

- ☐ å delta i intervju

Jeg samtykker til at mine opplysninger behandles frem til prosjektet er avsluttet

---

(Signert av prosjektdeltaker, dato)