“We have to trust it, or else we can just throw it away”: The use of decision support systems during extreme weather events

Mats Ekre Wang

Master’s thesis at the Department of Psychology

UNIVERSITY OF OSLO

15.05.2018

2018

“We have to trust it, or else we can just throw it away”: The use of decision support systems during extreme weather events

Mats Ekre Wang

http://www.duo.uio.no

Abstract

Global climate changes increase the risk of extreme weather events, posing a threat to our society. To reduce the impact of such events, efficient emergency management is crucial. Extreme weather events are characterized by a myriad of information from various sources which emergency managers must systemize and understand to make the right decisions. Automaton can be utilized to ease the demands on emergency managers by developing a decision support system. However, automation of prior human tasks is challenging and trust in automation is essential to deal with these challenges.

The aim of this thesis is therefore to explore (1) emergency managers' reflections about a decision support system in the context of extreme weather and (2) how learned trust in the model from K. A. Hoff and Bashir (2015) works in this context. As there are different ways to develop decision support systems, the differences between a system based on machine learning and a traditional system has been investigated within these two aims. Ten participants tested a prototype of a decision support system, half tested a traditional system while the other half tested a system based on machine learning. To investigate the first aim, they were interviewed in a semi-structured approach using the SWOT format which was analyzed inductively. The second aim was investigated using a specific interview guide to capture learned trust, which was analyzed deductively.

The themes from the inductive analyses were (1) aspects and characteristics of the system, (2) users of the system, (3) operational context, (4) interaction with the system and (5) decision making. The results provide valuable insight for further development of decision support systems. Moreover, the deductive analysis indicates that learned trust from the model of trust by K. A. Hoff and Bashir (2015) is relevant in this context, with some suggested adjustments that should be further addressed. There were no major differences between the participants testing the two different systems. This thesis is an early, explorative approach based on a limited sample. Nevertheless, it contributes with insights to the field of new and complex automated decision support systems based on machine learning.

Keywords: automation, decision support systems, emergency management, extreme weather incidents.

Acknowledgement

Firstly, I want to thank the participants who took their time to participate in this study and sharing insightful knowledge and opinions. Further, I want to thank SINTEF for letting me be part of the project and including me in their open and exciting research environment. I want to thank my initial supervisor at SINTEF Atle Refsdahl, which lead me on to decision support systems. My second supervisor at SINTEF, Marita Skjuve. You have gone above and beyond, making the last year inspiring, challenging and fun.

Thanks to my supervisor Cato Bjørkli for support and guidance throughout this thesis. Furthermore, I want to thank Bjørkli for his enthusiasm for Human Factors which has given me great inspiration.

Thanks to my class for making the last two years fun and interesting.

Thanks to friends and family, especially my brother, for all the support.

Thea, thank you for putting up with me, helping me and keeping my motivation up to the end.

Oslo, May 2018

Mats Ekre Wang

Table of contents

Extreme weather events are costly and can potentially be fatal. Globally, economic losses are reaching an yearly average of 250 to 300 billion USD due to flooding, cyclones, tsunamis and earthquakes (Field et al., 2012). In US alone, an average of 500 people die each year due to weather and climate related disasters and the frequency of extreme weather is expected to increase in the following years (Lubchenco & Karl, 2012; Meld. St. 10 2016-2017). Therefore, it is imperative that society at large is able to adapt to these challenges.

In Norway, emergency management is dependent on emergency managers' decision-making. These decisions are often based on limited information from weather and flood casts, experience from earlier incidents and physical observations from units in the field. It is therefore apparent that several aspects can influence the decision-making process.

One way to support emergency managers is by developing decision support systems (DSS). Automation can be used to make DSS that expands human capacity through assisting the emergency managers in acquiring and categorizing large quantities of information and give them advice of different actions (Nof, 2009; Parasuraman, Sheridan, & Wickens, 2000). However, automation is not at simple subtractive/additive task where a task formerly performed by humans is now performed by a machine (Dekker & Woods, 2002). Implementation of new technology alters the work situation and can introduce new practices and work methods (Flores, Graves, Hartfield, & Winograd, 1988). Humans interact socially with technology (John D. Lee & See, 2004), and the quality of the relationship between automation and the users can influence *if, how* and to what *extent* the technology is being used (Parasuraman & Riley, 1997). Thus, to what extent the potential of a DSS is being utilized in the context of extreme weather depends on the quality of this relationship. The notion of trust is further important in this context. As in any relationship, trust is an essential aspect of the human - machine interaction (John D. Lee & See, 2004). A system that is not trusted will not be used and appropriate levels of trust can guide users on when to follow the system's recommendations and when not to (Parasuraman & Riley, 1997; Schaefer, Chen, Szalma, & Hancock, 2016).

With technological development, systems based on artificial intelligence, such as machine learning (ML), is no longer an element of the future. Consequently, understanding how humans interact with such sophisticated autonomous systems and their reflections on using them in the context of extreme weather is important to utilize the full potential of new technology.

As the use of DSS is new in the context of emergency management during extreme weather events, this thesis seeks to (1) explore how emergency managers reflect on the use of a DSS in the context of extreme weather and (2) explore an excerpt of a model of trust in automation by K. A. Hoff and Bashir (2015).

The thesis is structured as follows: Firstly, extreme weather and emergency management in Norway is described and automation and DSS are defined. Secondly, challenges related to automation, the importance of trust and a perspective on human centered automation is presented. Thirdly, the methodology of this thesis is described, and finally the results are presented and discussed.

Extreme weather and emergency management

Extreme weather

While "extreme weather" is a concept most people are familiar with, no agreed upon definition exists resulting in several approaches and understandings of the concept. As Stephenson (2008) put it; extreme weather is "easy to recognize but difficult to define" (p.12). The weather can be considered to be extreme when a variable is "*above (or below) a threshold value near the upper (or lower) end (´tails´) of the range of observed values of the variable*" (Seneviratne et al., 2012, p. 111). One problem with this definition, however, is that an event not deemed extreme by this definition can still have extreme impacts. Stephenson (2008) therefore suggests a taxonomy for extreme weather events based on three criterions. These are *rarity*, *severity* and *longevity*. For an event to be rare it must be rare in the context where it happens. This means that snow in Norway is not extreme, but can be in Barcelona. The severity of an event refers to the impact an event may have. Longevity refers to if an event is acute or constant, such as storms compared to drought. This means that the extremity of a weather event depends on the context, severity and longevity.

The most common extreme weather events in Norway are flood, avalanches, storms and storm surges. While avalanches are the most fatal, it is seldom that weather events have casualties Norway (Meld. St. 10 2016-2017). Still, there are major costs associated with extreme weather. Between 2000 and 2017 approximately 10,8 billion Norwegian kroners where payed out by insurance companies related to damages from nature events (FinansNorge, 2018).

As demonstrated by both the definition and prior events; extreme weather event are both complex and dangerous. These events can have catastrophic consequences and it is essential that they are managed in order to minimize costs and fatalities. This puts an increased pressure on emergency managers who are responsible for the decision-making during these incidents. In addition to the complexity of the weather, the way the emergency management or organized may also influence the emergency manager's ability to make decisions. It is therefore important to take a closer look at how the emergency management team is structured in Norway to gain a better understanding of the role of a DSS during extreme weather.

Emergency management in Norway

Norway is divided into state, counties and municipalities. The counties are administratively responsible for the municipalities, and the county governor is the supervisor of each county (Berg, 2015). The county governors are responsible for information sharing and coordination throughout extreme weather events and across municipalities (*Fylkesmannens samfunnssikkerhetsinstruks*, 2015). Each municipality is responsible for the crisis management in an ongoing extreme weather event within the municipality (Sivilbeskyttelsesloven, 2010).

The Norwegian Meteorological Institute monitors and alerts any events regarding strong winds, heavy rainfalls and storm surges and the Norwegian Water Resources and Energy Directorate (NVE) monitors and alerts regarding flood and avalanches (Meld. St. 10 2016-2017, p. 79). In advance of an extreme weather event, the Norwegian Meteorological Institute or NVE notifies the county governor. The county governor in turn notifies the municipalities and advice a rise in the emergency preparedness level. Further, information from both The Norwegian Meteorological Institute and NVE are available from the online services; yr.no and varsom.no.

In case of an extreme weather event, different actors such as volunteer organizations (Red Cross, Norwegian Rescue Dogs, Norwegian People’s Aid), public services (police, health, fire, public roads, Water departments, civil defense) and private/partially private actors (power companies, machine contractors) have different roles (Meld. St. 10 2016-2017). These actors will often have their own responsibilities even though the municipalities have the overarching responsibility (Furevik, 2012). However, it is often procedure to form emergency management teams within the municipality when an extreme weather is approaching.

The teams consist of leaders from the organizations mentioned above, and are responsible in coordinating and supporting the work of the organizations (Furevik, 2012). Emergency management teams will be used as a collective term for these teams throughout this thesis. Emergency managers are responsible personnel working with emergency management. This can be managers working in either operation centers or out in the field. Emergency managers can also be the different members of an emergency management team. Throughout this thesis, the term emergency manager will be used for any responsible personnel that have a role related to emergency management during extreme weather events.

Although there is an efficient system for emergency management on an organizational level, there can be challenges related to that this system is dependent on the decision making from emergency managers as the next section will display.

Challenges within emergency management

While long-term and regional extreme-weather events such as flooding and storms gives time to prepare compared to a terrorist attack or a fire (Lubchenco & Karl, 2012), there are still challenges during such events such as practical challenges in terms of blocked roads posing a threat to transportation and evacuation during flooding or avalanches (Bründl et al., 2004; Meyer, Rowan, Savonis, & Choate, 2012).

Adding to this, there are often problems related to the available information. While emergency managers receive notification about extreme weather from the municipality, they have to seek information from different sources to get specific information for a given area. There are two main sources of official information about extreme weather events in Norway; Yr and Varsom. Both sources can overestimate or underestimate flood levels (DSB Rapport, 2015; Nedre Eiker Kommune, 2013) and emergency managers often need to validate this information with other information sources such as (1) other national and international meteorological institutes and weather forecasters, and/or (2) observations through reports from units in the field or they go out in the field themselves. Moreover, the information can often be limited. During flooding, for instance, lack of information about water levels in rivers, can results in time spend on estimating an measuring water levels to get the needed information before making decisions (Fylkesmannen i Oppland, 2014). In addition to the different sources of information the emergency managers must keep in mind the existing risk analyses and contingency plans for the municipality (Meld. St. 10 2016-2017).

In this myriad of information sources and considerations emergency managers must make decisions such as to initiate flood preventive measures, evacuate an area or to close roads. These decisions are made by the emergency management teams in collaboration with units out in the field or by the emergency managers themselves, such as the police because they are the formal leaders of any incident and can make independent decisions (Furevik, 2012).

To make these decisions the emergency managers rely on the information described above, in addition to local knowledge, experience and general knowledge that they themselves or collaborators have. The multitude of information can lead to information overload (Bawden & Robinson, 2009), perception of high workload (Hart & Staveland, 1988) and loss of situational awareness (Endsley, 1995). Furthermore, it can cause stress, decreased working memory and more narrow attention (Wickens, Lee, Liu, & Gordon-Becker, 2013). These factors can ultimately lead to decreased performance and faulty decisions.

By developing DSS that can give the emergency managers information and recommendations during an extreme weather event it is possible to assist the them in their decision-making process. DSS can be seen as a form of automation, where a former human task are done by a system (Parasuraman et al., 2000). To fully understand this process, it is therefore important to define automation and the relation to DSS as the next section will cover.

Automation

The following section will present the definition of automation, including various tasks and levels of automation, before describing automation in terms of decision support systems in the context of extreme weather and the development of such systems.

Defining automation

Parasuraman et al. (2000) defines automation as "a device or system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator". The authors distinguish between four different types of tasks that can be automated: (1) acquisition of information, (2) analysis and manipulation of information, (3) decision and action selection, and (4) action implementation.

A system can do several of these tasks, and in the context of extreme weather events a DSS could collect information from different sources such as forecasts and sensors, analyze this information and suggest a set of actions. Hence, provide decision support.

Automation and Decision Support Systems

DSS are "any interactive system that is specifically designed to improve the decision making of its user by extending the user's cognitive decision-making abilities" (Zachary, 1988, p. 997). This definition emphasizes the notion of DSS as an extender of the human decision-making capabilities. Such systems do not choose for the user, but can provide suggestions and recommendations (Zachary, 1988). DSS can thus be seen as automation of one or several tasks.

Parasuraman et al. (2000) also separates between different levels of automation. Meaning that automation can range from humans having full control to the automation having full control over the system. DSS can be placed around level 3 which "*narrows the selection down to few*" and level 4 which "*suggests one alternative*" (Parasuraman et al., 2000, p. 287). Unlike other types of automation, DSS do not replace the human and it is still the human that have to make the final decision (Cohen, Parasuraman, & Freeman, 1998).

Using the framework from Parasuraman et al. (2000) a DSS is the automation of different tasks and at a lower level of automation.

Developing Decision Support Systems for Extreme Weather Events

Technological systems are already used to provide better predictions of heavy rainfalls, effects extreme weather will have on power grids, weather effects in relation to food production and solar radiation (Aybar-Ruiz et al., 2016; Biffis & Chavez, 2017; Eskandarpour & Khodaei, 2017; Zhang, Zwiers, Li, Wan, & Cannon, 2017). With the continuous development within technology it is possible to make advanced systems that can assist emergency manager in tasks related to extreme weather events.

From a technological perspective it is a vast array of different approaches to develop DSS (Nof, 2009; Russell & Norvig, 2016). In this thesis two approaches will be examined. These are a qualitative approach to multiple criteria decision making technique (MCDM) and machine learning (ML). Within these methods there are more nuances, but this is described in a more general sense in this thesis.

The term MCDM is used about different approaches that seek to assist in decision making where there are different factors that influence a main decision (Mardani et al., 2015). Simplified, this is done by breaking a problem, such as whether to close a bridge during flooding or not, into sub-problems (Mardani et al., 2015). This is done by letting experts in the field identify import factors that contribute to the main decision (Ho, 2008). If the case is whether to close a bridge, experts list different factors such as wind, waves and rain, and weigh the importance of the different aspects. The weighted factors are put in a hierarchical structure where to close the bridge or not is at the top and the different sub-problems are subordinate. This is then imported into a software that can, based on the hierarchical structure that the experts have contributed to, analyze real-time data and give advice on recommended actions (Erdogan, Refsdal, Nygård, Randeberg, & Rosland, 2017).

The other approach is by using ML, a type of artificial intelligence (AI). The field of AI is focused on developing intelligent behavior for machines such as reasoning, planning, voice recognition and learning (Nau, 2009) and a learning machine seeks to provide better advices based on prior observations of the environment (Russell & Norvig, 2016, p. 693). By letting a machine study earlier weather data and what measures that where taken at the time, it can develop different advices based on historical and current data. This is done by firstly developing different scenarios for when there is a need to close the bridge. These scenarios are then applied to the historical data to when the bridge where closed and not. By studying a vast amount of historical data the algorithm behind the system is strengthened. When extreme weather is forecasted, the system will analyze the incoming data and recommend closing the bridge or not, based on the system algorithm that is again based on historical data.

The two approaches are developed and function in different ways. MCDM is developed more manually and must be updated by developers if the system does not perform properly. ML, on the other hand, is self-learning based on historical data and further evolves when it is used in facing new extreme weather events. Thus, this sets a higher demands on the quality of the data as ML is no better than the data it is based upon (DeBrusk, 2018).

The development of such systems is often intended to make systems safer and to increase human performance (John D. Lee & Seppelt, 2009). This is not always true, as the next sections will show.

Challenges with automation

As described, automation is not an additive task where the former function of the human are automated without altering the situation (Dekker & Woods, 2002; Flores et al., 1988).

A DSS in the context of extreme weather is automation of complex cognitive tasks. The system becomes an active agent as it percepts weather data from the environment and actuates in forming advice, based on these perceptions (Russell & Norvig, 2016, p. 34). Users have to relate to this agent, and users interact socially with such complex technology (John D. Lee & See, 2004). The relationship between the user and the system becomes more important compared to simpler technology, as the responsibility for the situation are shared between the user and the system (Madhavan & Wiegmann, 2007).

In comparison to a flashlight that has one task (e.g., to illuminate), DSS have a far more complex task of enlightening emergency managers in complex situations with a high degree of uncertainty. It is obvious to the user when the flashlight stops working, and the users can rely on the flashlight to fulfill its task as long as it has enough battery power. However, using a DSS this is more complex and can thus make new pathways that leads to errors that are unforeseen (K. A. Hoff & Bashir, 2015; Perrow, 1999).

There are many examples of failed automation both where there is a technical error and where the error can be attributed to the interaction between the operator of a system and the system. One such example of the later is the cruise ship Royal Majesty, which followed a GPS route and ran aground where the National Transportation Safety Board (1997) concluded that the crew on Royal Majesty had too much reliance in the GPS system.

When automation becomes more complex, humans forms relations with the technology (Muir, 1987) and in any relationship trust is inevitable (John D. Lee & See, 2004). The next section will elaborate on why trust is important in automation.

Trust in automation

Defining trust

Trust have been studied in many contexts and several definitions exists (John D. Lee & See, 2004). In this thesis, trust is defined as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (John D. Lee & See, 2004, p. 51). This entails that trust is an attitude towards the DSS that it will assist in the decision making process. So, if the emergency manager has trust in the DSS they have an attitude towards the system that it will help achieving their goals.

Trust and automation

Trust is especially important in a critical context such as emergency management or military settings where the operators own or other’s life is dependent on the interaction between the operator and the technology (Hancock et al., 2011). Specifically, users can have too much (overtrust) or too little trust (undertrust) in the automation (Parasuraman & Riley, 1997), which in turn can result in accidents. Hence, it is important to strive for the appropriate level of trust. An automated system is seldom failure free and “user’s must be calibrated to the decision aid, so that he neither consistently underestimates nor overestimates its capabilities” (Muir, 1987, p. 527).

Models of trust

There are have been different approaches to model trust in automation. Hancock et al. did a meta-analysis in 2011 on trust in robots and suggested a model. The author argued that there are similarities between general automation and robotics and building on this work a new meta-analysis where done by Schaefer, Chen, Szalma and Hancock in 2016. This lead to a new model of trust in automation. Both model separated between human factors, factors associated with the system and environmental factors.

Prior to this, John D. Lee and See (2004) made a qualitative model of trust in automation that incorporated some additional views on trust. Based on this work, K. A. Hoff and Bashir (2015) did a qualitative review of 101 papers with 127 studies from 2002 to 2013 that lead to a model of trust in automation that this thesis will use. This model is a vast model and this study will focus on an excerpt of this.

Model from Hoff and Bashir. The model consists of three layers of trust from Marsh and Dibben (2003); dispositional trust, situational trust and learned trust.

Dispositional trust represents an overall tendency to trust the system, independent of the context and system. Dispositional is divided into trust age, gender, culture and personality trait. *Situational trust* represents the dependency of the situation the interaction happens in. Situational trust is divided into external and internal variability. Internal variability is the individual's mood, attention capacity, self-confidence and expertise within the field that the system operates in. This differs from the dispositional trust as they may vary depending on the context. External variability refers to type of system and the complexity of the system that are in use, the workload and the difficulty and framing of the task, the organizational setting and the perceived risks and benefits of using the system. Lastly, *learned trust* is about the evaluations a user has about the system depending on earlier experience with, and knowledge about, similar systems. This thesis will further explore the layer of learned trust. See figure 1 for factors of learned trust.

Learned trust is divided between initial learned trust and dynamic learned trust. Initial learned trust are the expectations prior to an interaction with a system. This can be based on attitudes and/or expectations towards technology, if there is a reputation and how a system's/or similar system's reputation is, earlier experience with the system or similar systems and how the users understand the system. Dynamic learned trust is about the interaction the user has with the system. This depends on the system performance related to reliability, validity, predictability, dependability, error and usefulness. Furthermore, how the performance is perceived can be affected by design features.

As DSS in the context of emergency management during extreme weather events are new to the potential users the layer of learned trust especially interesting. This entails both the perception of systems performance and the preexisting knowledge prior to an interaction making these two aspects interesting in relation to DSS in the context of emergency management during extreme weather events.

Design features, the interdependent relationship between reliance on system, system performance and dynamic learned and how the initial learned trust affects the initial reliance strategy will not be further addressed.

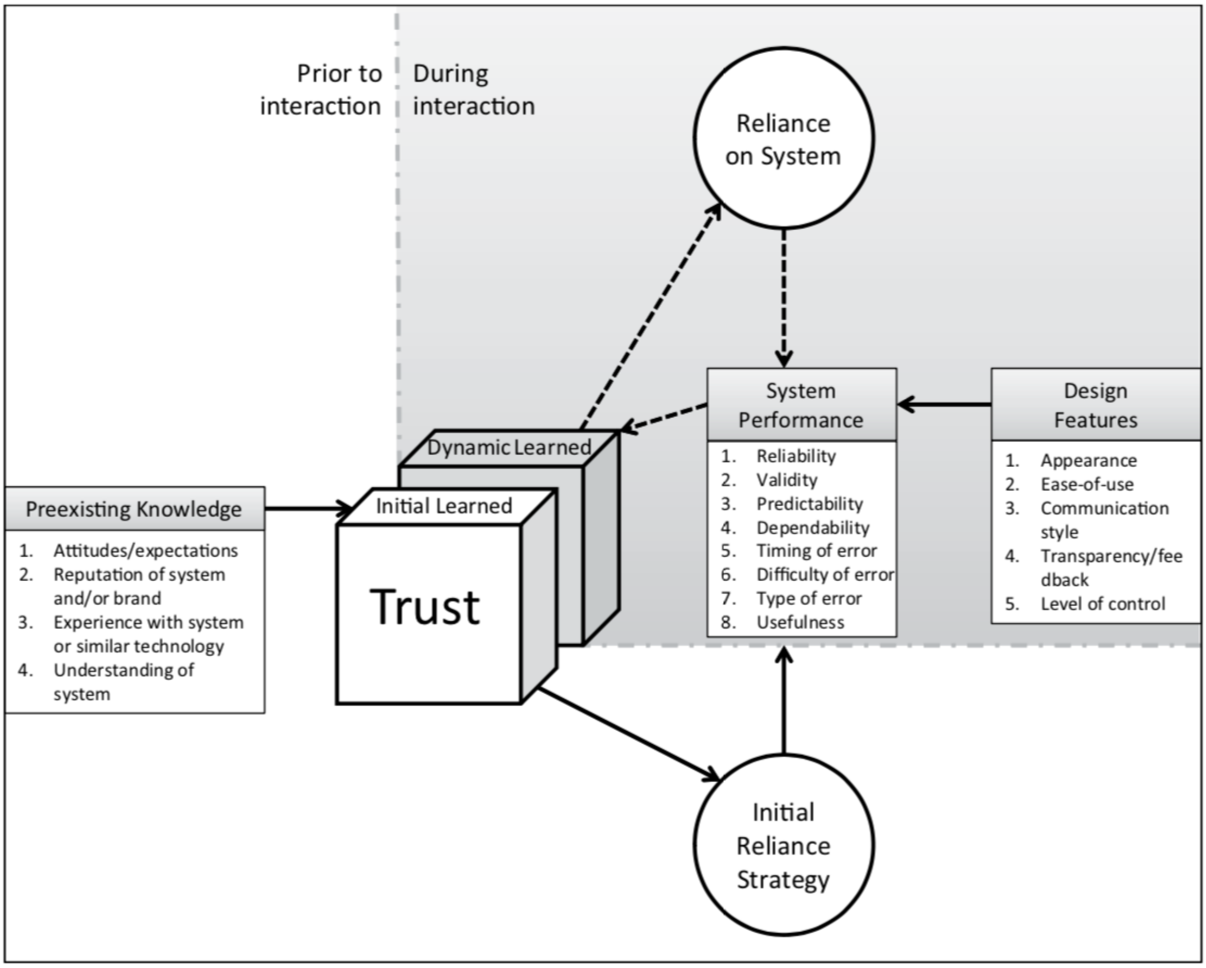


Figure 1. Factors that influence learned trust. From "Trust in automation: Integrating empirical evidence on factors that influence trust." Hoff, K. A., & Bashir, M. (2015). *Human Factors, 57*(3), 407-434. Copyright © 2014, Human Factors and Ergonomics Society.

Knowing that there can be pitfalls in automation, how we design DSS are of great importance in order mitigate some of the challenges and to build for trust.

Human centered automation

According to John D. Lee and Seppelt (2009) there are two perspectives on why the cruise ship Royal Majesty ran aground. One is to say that this was a failure of the system, or a human error in some way. Another is that the design of the system failed to assist the crew in their new task of navigating with assistance.

As previously noted, the implementation of new systems and automation alters the current situation and is not a straightforward task (Dekker & Woods, 2002; Flores et al., 1988). Systems must be based on a human needs, and not what is possible to automate, and it is therefore a need to design systems that take the operator in to account (Tedre, 2008).

Through designing a DSS that incorporate psychological perspectives it is possible to design for a system that will be accepted, used and used properly (Maguire, 2001) in order to achieve harmony between the human and the automation (John D. Lee & See, 2004; Parasuraman & Riley, 1997; Wickens et al., 2013).

Aims of this study

The aim of this thesis is twofold. Firstly, the thesis explores the initial reflections that emergency managers have regarding the use of DSS. Secondly, it explores how the model of trust from K. A. Hoff and Bashir (2015) works in this context. In addition, potential differences between systems based on MCDM and ML are examined in relation to both aims; initial reflections and trust.

In order to design a DSS that mitigates the challenges of automation, feedback from the potential users is imperative. How do the emergency managers view the use of a DSS in the context of extreme weather? What are their reflections on this topic?

Further, trust is important both to ensure intention to use and correct use of a system (Parasuraman & Riley, 1997). K. A. Hoff and Bashir (2015) suggest that their model can guide design of automation and explain trust, and as DSS are new in this context learned trust is especially interesting. Is learned trust within the model of K. A. Hoff and Bashir (2015) applicable in this context? Are there considerations that needs to be taken for the model to be relevant in this context?

Lastly, there are different approaches to develop DSS. To interact with complex systems based on ML, instead of traditional systems, can become a reality for emergency managers. How does the type of system affect the users? Does it affect trust?

As there has been little research within the field of DSS in this context and on DSS based on ML, this thesis takes an explorative and qualitative approach seeking to provide insight on design of such systems and to generate new hypotheses within the field. This is done by examining the following four research questions:

1. *What characterize emergency managers' reflections about a decision support system in the context of extreme weather?*
2. *Do the reflections differ between emergency managers using systems based on ML compared to systems based on MCDM? If so, how?*
3. *How applicable is the layer of learned trust within the context of decision support systems for emergency managers during extreme weather events?*
4. *Does learned trust differ between emergency managers using systems based on ML compared to systems based on MCDM? If so, how?*

The research questions were examined by having five emergency managers test a prototype of a DSS based on ML, while another five tested a prototype of a DSS based on MCDM. This test was followed by interviews.

Method

The following section will describe the sample and sampling methods, preparation for the data collection and the study procedure.

Affiliations

This study is conducted in collaboration with SINTEF and the EU project ANYWHERE[1]. The purpose of ANYWHERE is to develop technology that assists emergency managers during extreme weather events (ANYWHERE, 2017). The data for this thesis was collected in collaboration with a researcher at SINTEF and will, in addition to this thesis, be used as input in the development of the ANYWHERE system.

Sampling method

A purposive recruitment was done to get participants that would provide rich information about the topic (Etikan, Musa, & Alkassim, 2016; Patton, 1990). The participants needed to (1) have experience from work with emergency management and (2) be a central decision maker during extreme weather events to be included in the study.

A total of 16 participants was contacted between October and December 2017, and the final sample consisted of 10 participants. The sampling was done by recruiting participants through SINTEF's existing network, from a college that educates in emergency management, where typical students are experienced employees that seeks further education, and by directly contacting relevant participants. The participants recruited from the college (two participants) were given a gift card of 500 NOK for their participation.

Participants

The participants had background from emergency management in either aviation, police or fire departments, municipality work or public roads. Three of the participants worked in the western part of Norway and the remaining worked in the eastern part. This extended the scope of the sample as weather related challenges differ between the eastern and western part of Norway. Two of the participants were female and eight were male. Five of the participants had used DSS before, three had heard of, but never used it and two had neither heard of or used a DSS. CIM, which is a tool for managing risk, safety and incidents

[1] Project number: Grant Agreement No. 700099, (http://anywhere-h2020.eu/), stated in 2016.

developed by One Voice[2], was the DSS that five of the participants had prior experience with. Two of the participants had tried other DSS in addition to CIM.

The participants were randomly assigned into either the MCDM or the ML group. There was one female in each group and the amount of earlier experience with other DSS were similar between the groups. The average age for the participants was 44 (r = 34-59) in the MCDM group and 45 (r = 30-58) for the ML group. The average years of experience in current position was 6 (r = 0,5-13) in the MCDM group and 5 (r = 2-10) for the ML group. The average years of experience within emergency management was 36 (r = 4-38) in the MCDM group and 21 (r =6-32) for the ML group. The participants in the MCDM group had participated in an average of 36 extreme weather events (r = 8-125) and the participants in the ML group had participated in an average of 9 weather events (r =0-30).

Preparation before the data collection

The preparation for the data collection included development of a prototype of a DSS and a case to test the prototype in. The following section will describe these elements in more detail.

Development of the case. An adapted version of a role-playing game called ANYCaRE (ANYWHERE Crisis and Risk Experiment), developed by project partners in the ANYWHERE project, was used to present the participants with a platform in which to test the DSS. The original objective of this game was to provide potential users of ANYWHERE with a common platform to test different tools (Terti et al., 2017), and it was thus deemed appropriate for this study.

The game consists of instructions, a map over a town and weather data. The role of the participants is to be emergency managers for the town and to make decisions for three areas that are affected by extreme weather. See appendix A for the map.

Four adaptations were carried out to tailor the game to the purpose of this study: *First*, the instructions were modified and translated into Norwegian. *Second*, there were made some adjustments to the map to make the case more appropriate for Norwegian emergency managers. Snow was added on the mountain peaks and area A was moved. *Third*, some changes were made to the weather data. The weather data were taken from a weather incident called Synne that occurred in Norway in 2015, instead of the weather data used in the original case (Ekstremværrapport Hendelse: Synne, 2015). This was to make the case more realistic for Norwegian emergency managers. The case and the adjustments were validated by a panel

[2] https://onevoice.no/cim

of four domain experts that works within the field of extreme weather management. *Lastly*, a time limitation. This specific time limitation was established after testing the case with emergency managers who validated that this restriction gave the case a realistic feel as the workload, stress and complexity increased with shorter decision making time.

Development of the prototype. A prototype of the DSS tool was developed by the author in collaboration with researchers and developers at SINTEF. The prototype was made more realistic and complex compared to how prototypes usually are designed. The objective was to get an operative system that the users could interact with and experience as a working system. The prototype was made with the tool Justinmind[3], and were finalized after several iterations and seven pilot tests.

Manipulation of differences between ML and MCDM. A purpose of this study was to examine the differences between systems based on MCDM or ML. To develop two fully functional systems would be expensive and ineffective at this early stage of the product development. It was therefore decided to develop one prototype for both groups and that the difference between the two groups would be how the systems were presented two the users and some minor differences in what the participants of the two prototypes had access to in the presentation of input data on the display. As the aim for this thesis is to collect initial reflections about the two different systems at an early stage this was deemed appropriate.

The ethical implications of misleading where discussed, and the conclusion was to convey to the participants that the system “would be based on ML/MCDM”.

The participants were explained that they were testing a prototype of a new system that is being developed. They were further told that this system was based on either MCDM or ML. The explanations about ML and MCDM were made by the author of this thesis in close collaboration with one expert in MCDM and one expert in ML at SINTEF. This was to ensure that the explanations were both correct and in a similar format. The explanations where assisted with illustrations and read to the participants in a similar way (see appendix B for illustrations).

One limitation of this is that the difference is in the explanation of the system (with the exception of presentation of input data), and not the actual system. However, based on the limited choices of actions and weather data available in the case, the actual experienced difference between the two systems would be minimal

[3] https://www.justinmind.com/

Interface and differences between the ML and the MCDM. The interface consisted of five tabs and was made simple to navigate to keep the focus on the experience of using DSS and not the design of the display in particular. See appendix C for examples from the prototype for each tab. On the *first* tab, there was an interactive map over the town where the participants could click on each area. This lead to the *second* tab and the part of the system that gave the decision support, called "risikobilde" (Norwegian for risk picture or risk overview). Both group were presented with the same recommended advice. However, the input variables that the participants could access varied according to study conditions. The users of MCDM got a list of input factors (rain, snow melt, water flow in river etc.) and an estimated risk to infrastructure and health and the users of the ML version only got the list of the input factors and not the risk to infrastructure and health. This is due to the existing limitations regarding transparency in the way the algorithm in the ML system is developed, not revealing the logic behind in a comprehensible way (Grunning, 2018) and was therefore included as a difference between the two systems in addition to the different explanations. See appendix C for the differences under "Advice tab MCDM" and "Advice tab machine learning". The *third* tab contained weather data for the three locations in the town. These were made to look like the Norwegian weather forecast service called Yr. The weather information could be shown for each area or for the areas combined. On the *fourth* tab the participants had access to flood forecast. This was made after the same template and warning levels as the Norwegian Water Resources and Energy Directorate provides through the webpage Varsom.no. Finally, on the *fifth* tab showed a combination of weather and flood forecast.

Decision support. Apart from the graphical interface, the system needed to work as a decision support system for the users providing credible advices. To achieve this, a model for decision support was developed by a researcher at SINTEF, an expert in risk based decision support at SINTEF and the author of this thesis. This was the logic, or algorithm behind the recommendations and did not vary across the two groups.

The MCDM approach were chosen as it would be costly to train a ML algorithm and that the data provided in the case were limited. This was developed using the DEXi software developed by Bohanec (2017). The input values (weather and amount of people in the areas) were used to estimate risk values for infrastructure and health. Based on this, a recommendation was made from the model. As for the ML condition these risk values were excluded in the presentation of the advice, but both ML and MCDM were based on the same logic. During the development, there were continuous dialogue with a panel of four experts within emergency management and flood, as recommended in developing DSS using MCDM (Erdogan et al., 2017).

Procedure

Prior to the case and interview. The author had in advance participated in a 35-hour course on how to conduct interviews. The focus of this course was the PEACE model (Clarke & Milne, 2001; McGurk, Carr, & McGurk, 1993). Seven pilots were conducted prior to the data collection to ensure that the case, system and interview guide performed properly.

In advance of the interviews the participants were sent the informed consent (see appendix D). This information was repeated to the participants prior to the interviews and the participants had to sign the informed consent.

Case. Each testing session was initiated by giving the participants information about how to solve the case and about the DSS. The participants were told that the DSS was developed by SINTEF, and that it was based on either ML or MCDM depending on which group the participant were in. Further, they got an explanation of how either MCDM or ML systems work (depending on assigned group).

The DSS was presented on a tablet with an 9,7-inch screen which were deemed as a suitable choice during the pilots, and the same tablet was used every time.

The gameplay consisted of three rounds of four minutes. During each round the participant were to make choices for the three areas on the map to minimize damage on infrastructure and loss of lives. The choices were either to (1) await the situation, (2) start flood preventing measures or to (3) evacuate the area. In addition, there were two specific choices for area C regarding cancelling a festival and closing the bridge. To effectuate their choice of action the participants filled out a form for each round (see appendix E).

The interviewers were in the room while the participants solved the case, and the participants were told not to ask any questions during the test sessions. Keeping the interviewer in the same room was chosen as the rounds where so short that leaving the room would create more disturbance than to quietly observe. The interviewers where conscious not to sit straight in front of the participants to give the participants space while solving the case.

Interviews. Six of the interviews were conducted by a researcher from SINTEF and the author of this thesis together, two interviews were done by the author and two were done by the researcher at SINTEF. The first six interviews where done by both together. This was to ensure that the two interviewers adopted the same interview style and used the same type of follow-up questions.

Six of the interviews were conducted at the participants' own workplace and four were conducted at SINTEF. The interviews took place in December 2017 and January 2018. The interviews were recorded and the mean time of the interviews was 35 minutes with a range between 26 minutes and 46 minutes.

Transcription. The interviews were transcribed by the author. The author played back the recording afterwards and crosschecked with the transcript and errors where corrected. There were five places where the recording was unclear. This was short words or half sentences that were not of significant value with a total loss of sound for twelve seconds.

Measures

Because there has been little research on the use of DSS within the context of extreme weather and less on the use of ML, an explorative qualitative approach using interviews was deemed appropriate. This can contribute with rich understanding of the users' experiences and reflections contributing (Polit & Tatano Beck, 2010; Pratt, 2009)

The measures in this study were interviews and demographic variables. The interview was divided into two parts that reflected the research questions. The first part of the interview was aimed at the first two research questions related to the reflections of the emergency managers and the differences between ML and MCDM. The second part of the interview was aimed at the second two research questions related to what the layer of learned trust from the model from K. A. Hoff and Bashir (2015).

Interview guide. The interviews and the following analyses where separated in two parts. See appendix F for the interview full guide.

The *first* part of the interview was a semi structured interview based on the SWOT format (Dyson, 2004) that covers strength, weaknesses, opportunities and threats. This format assists the users to think about a topic in a positive-negative dimension and a time dimension and does not demand specific types of answers (Thomas. Hoff, 2009; Lone, Riege, Bjørklund, Hoff, & Bjørkli, 2017). This is beneficial for capturing the emergency managers' reflections around the use of DSS in this context without any prior assumptions.

For the *second* part of the interview a guide was created to cover the layer of dynamic learned trust from K. A. Hoff and Bashir (2015). Preexisting knowledge and system performance from the model from K. A. Hoff and Bashir (2015) were chosen as design of the display and usability of the system in particular were not the specific aim of this study. Two aspects of system performance were excluded from the guide: (1) reliability was excluded, as the reliability of the advices would not be possible to investigate in the case, and (2) error was excluded as the aim were initial and general reflections from the emergency managers and the system would not fail during the case.

The question formulations were based on papers that included measurement scales referred to by K. A. Hoff and Bashir (2015) and existing measures of trust found in (Chien, Semnani-Azad, Lewis, & Sycara, 2014; *Guidelines for Trust in Future ATM systems: Measures*, 2003; Jian, Bisantz, & Drury, 2000; Madsen & Gregor, 2000; Mcknight, Carter, Thatcher, & Clay, 2011).

Demographics. Demographic information where collected prior to the interview (Appendix G). These were sex, age, employer, work title, work experience in current position and in total within emergency management, education, how many weather-related incidents they had participated in and experience with DSS.

Analysis

A thematic analysis following the framework by Braun and Clarke (2006) was carried out. This is a method that have a long tradition within qualitative research (Boyatiz, 1998) and is suitable for identifying patterns of information across large datasets (Braun & Clarke, 2006). The data analysis software Nvivo 11 was used to assist in the analysis.

The *first* part of the interview was in a SWOT format to capture the participants' reflections on the DSS and were analyzed inductively. The *second* part of the interview was based on an interview guide that captured preexisting knowledge and system performance from the model of K. A. Hoff and Bashir (2015), and the data from the second part of the interview was analyzed deductively. The residual codes from the second part of the interview were then analyzed inductively. These codes fell within the themes already identified from the first part of the interview. Therefore, the same themes and sub-themes were used for the first part of the interview analyzed inductively and for the residuals of the second part of the interview that were analyzed deductively.

The data was analyzed by their semantic level, which means that the explicit meaning was coded, and that there were made no interpretation of an underlying meaning or assumptions (Braun & Clarke, 2006).

Inductive analysis. The analysis was done within the framework described by Braun and Clarke (2006). The five phases are described below:

Getting familiar with the data. The interview transcriptions were read through to get an overview and first impression of the data before coding the material.

Generating initial codes. This is the first level of coding were each meaningful unit in the transcript is coded. The transcripts were coded using the definition from Thomas Hoff et al. (2009, p. 7).; “the smallest meaningful unit that reflects the informant’s experience and understanding of the topic of interest”. Depending on the meaning, this could vary from parts of a sentence to several sentences (Thomas Hoff et al., 2009).

The material was reviewed and recoded several times as more codes where made, and a codebook was established to organize thoughts and avoid the use of similar codes. This originally consisted of 110 codes and definitions. After several cycles the final codebook contained 87 codes.

Identifying themes. The codes were reviewed, compared and combined into themes and sub-themes. The first round resulted in 10 themes and 33 sub-themes.

Evaluating the themes. The initial themes and sub-themes were evaluated by going back to the transcripts to assess if they reflected the content in the transcripts. After reviewing the themes with the supervisor of this thesis, 5 themes, one residual theme and 25 sub-themes were established.

Naming the themes. Each theme was given a name that reflected the content.

Deductive analysis. The second part of the interview was coded using a deductive approach based on the model from K. A. Hoff and Bashir (2015). The second part of the interview were aimed at preexisting knowledge and system performance, but the entire model was used during the analysis. This was because the participants talked about topics outside of the preexisting knowledge and system performance, but still within the model from K. A. Hoff and Bashir (2015).

The remaining codes that did not fit within the model were analyzed inductively using the same approach as described for the inductive analysis. Some new codes were made, but all the codes fell within the existing themes and sub-themes from the first part of the interviews.

Reliability. Themes do not emerge from the data, but the researcher plays an active role (Braun & Clarke, 2006). This means that another researcher might have interpreted the data differently. To mitigate some of this bias, the analysis was discussed with both the researcher at SINTEF and the supervisor of this thesis continuously.

To further strengthen the analysis a measure of the inter-rater reliability could have been included as recommended by Mays and Pope (1995). They argue that by a comparison of another coders’ independent analysis the quality of the analysis is strengthened. Armstrong, Gosling, Weinman, and Marteau (1997), however, argue that qualitative research is subjective and that reliability tested through an independent analysis of the transcriptions does not strengthen the quality of the research as the transcriptions are already colored by the researcher. The author of this thesis decided not to do an inter-rater reliability test as this study was not a validation of a model but an explorative design.

Ethical considerations

The participants were given the informed consent per e-mail in advance of the study and the information was repeated before the demonstration and interview. They were informed that they were anonymous, that their personal data only were available for the author and one employee at SINTEF, that the data was stored on a safe server and that they could withdraw from the study at any time.

The information that the participants contribute with was not sensitive or related to health issues. Further, the topic of this thesis was not deemed as sensitive and there were not expected a negative impact of participating. However, extreme weather events are not ordinary so if a participant were to experience any issues or unpleasantness during, or after participating in the study SINTEF would follow up on the participant in accordance with their guidelines.

This study is approved and conducted in line with the guidelines of the Norwegian Social Science Data Services (Project number 50677).

Results

The aim of this thesis is twofold. Firstly, the thesis explores the initial reflections that emergency managers have regarding the use of DSS in the extreme weather context. This was done in the first part of the interviews using the SWOT format. The data from this part of the interview was analyzed inductively.

Secondly, the thesis explores how the model from K. A. Hoff and Bashir (2015) works in this context. This was done in the second part of the interviews using an interview guide that captures preexisting knowledge and system performance. The data from this part of the interview was analyzed deductively. The statements that were not covered by the model from K. A. Hoff and Bashir (2015) were analyzed inductively.

Additionally, the differences between the MCDM and ML group were examined in relation to both the reflections and to the model from K. A. Hoff and Bashir (2015). This was done by analyzing each interview separately and comparing the results from the two groups.

The following section will present the results from the analysis. The results from the inductive analysis based on the first part of the interviews is presented, followed by the results from the deductive analysis. The differences between the ML and MCDM group are reported continuously throughout both sections.

Inductive analysis

The inductive analysis resulted in 5 themes and a residual theme. These were (1) *aspects and characteristics of system*, (2) *decision making,* (3) *interaction with system,* (4) *operating context,* (5) *users* and (6) *residuals*. The themes, sub-themes and the frequency, defined as the number of participants that mentioned each theme, is illustrated in Table 1. The themes and the group differences will be described in more detail below.

Table 1

Overview of frequencies for themes and sub-themes

Themes and sub-themes	**Groups**	
	MCDM	ML
Aspects and characteristics of system	4	5
Information from system	4	5
Design and transparency	2	4
Logic and functioning of system	3	3
Advice from system	2	3
Dependability and security	2	1
Decision making	5	5
Context and considerations	5	4
Types of decision	4	3
Explanations and stories	3	3
Certainty/uncertainty	2	1
System in decision making process	4	5
Interaction with system	5	5
How to treat output from system	4	5
General use of the system	3	4
Error	1	2
Dependency of system	1	2
Operating context	5	4
Current situation and systems	4	3
Technological development	3	2
Workplace considerations	2	2
The context of weather	2	1
Users	4	4
Work-related experience	3	4
Knowledge	4	3
Residual	5	5
Case-related	3	4
About ANYWHERE system	5	5

Note. Bold = main theme and italic = sub-theme. Frequencies indicate number of participants that mentioned each theme

Aspects and characteristics. This theme encompasses attributes of the system and DSS in general. The participants liked that the system conjoined *information* from two different sources (Varsom and Yr), which made it easy to get an overview of the available information. Some of the participants also stated that they liked that the information were projective so that they could start preparing early. The participants also stated that the system was easy to use due to the way the information was presented, as in the following extract:

> "*It is all the information you need about the situation… it is a simple tool. It is right there. There is nothing that you need to look up or search for. You have it right in front of you.*"

The participants also talked about how the system *functions and the logic* behind it. Half of the participants emphasized the importance of input data being up to date in order to achieve correct output.

> *"The system can miss out on that there were thousand tourists from a cruise ship down at a pier (...) So the consequences can be more severe, or one can decide to evacuate a building that were evacuated yesterday"*

Several of the participants specifically talked about the *advices* that the system gave. Two participants in the MCDM group thought there was a lack of explanation of what they advices were based on. They felt they had to try to find out for themselves how the system reached its conclusions. This made the unsure on whether the system was correct or not. Some of the participants also described the advices as being general as a positive attribute, while others thought the advices where to assertive:

> *"I am always skeptical of emergency management tools that become almost commanding or ordering*"

As for *dependability and security* some of the emergency managers talked about system failure as a challenge with all technological systems and two suggested that the system could be manipulated or hacked, and underlined the importance of being aware that systems can fail.

Group differences. There were some differences between the groups related to the focus on the use of experts in the development of the system. Specifically, one participant in the MCDM group found it easier to trust the system because experts had developed it. There were also some differences in how they talked about the logic and functioning of the system. Two emergency mangers in the ML group talked about how the machine learning worked and were curious about the functioning of ML. One from the MCDM talked about the algorithm behind the system specifically, which none of the participants in the ML group focused on. Two participants in the MCDM group talked about the lack of transparency and that they struggled to understand the logic behind the advices. Further, one participant from the MCDM group thought the system had clear limitations in that the system had to project every event, which would be time and resource consuming for large areas. No one in the ML group mentioned issues related to the limitations in relation to how the system functions and are developed.

Decision making. This theme was about decision making, both within the case and in general and entails the *context* of decision making, what *considerations* the emergency managers had to take and *types* of decisions. The participants talked about how time pressure could make it difficult to assess all options before deciding. They reflected on the consequences of making decisions such as the cost of evacuation, and the implication of taking measures when it is necessary – even though it might be expensive, or taking measures when it is not necessary. The participants also talked about decisions they had made in the past, and explained that you need to be able defend or vouch for a decision at a later time, and how this might look if they followed a recommendation from the system that was wrong, as highlighted in the following extract:

> "*(...) and then some Councilman, who has gotten himself a new program, "no, the program said that we must cancel (the festival)" (...) Should like to see him stand there and defend that, when nothing happened"*

Moreover, the participants talked about the use of *DSS in decision making*. They discussed about how the system supported and eased the decision making process, how it could give a higher consistency in decisions and that it was timesaving. Two participants, however, stated the system oversimplified a complex task as decision making in extreme weather events is, and were concerned that the system did not consider all the different factors and that this could lead to faulty decisions, as exemplified in the following statement:

"If you only base your decision on this (the system) and you cancel the festival because of a higher water level, it just, it becomes too simple"

Group differences. There were few noticeable differences related to this theme. However, one participant from the MCDM group emphasized the fact that the system had been developed by experts in extreme weather was a strength in the decision making process, because the experts would know more about flooding, rains and what measures that are important than the participant.

Interaction with system. This theme contained topics related to the interaction with the system.

The main issue the participants stressed related to this theme was how to *treat output from the system.* This entailed concerns about users blindly follow the advice. One participant reflected on the system's status in the organization and that it had to be expressed clearly whether the users must follow the advice or not. They also talked about trust and the need for emergency mangers to evaluate the advice from the system themselves, and not merely follow the advice from the DSS:

"It (the system) is made to tell you all about weather and what measures you should take (...) so if it works five times in a row, what is the reason not to trust it next time? (...) In the end you don't take your own decisions because you have a tool that does it for you"

Further, the participants talked about blaming the system. While most agreed that in the end it is the emergency manager's responsibility, one participant also stated that blaming the system was possible, even though it is not good practice.

"If an expert has said that it is not likely (likelihood for flooding), so okay, then you at least have someone to blame"

The participants also talked about cross-checking the systems' output with other information or with physical observation of the weather, to validate the available information as described by the participant below:

"*I start to look for things that can substantiate the advice, so I struggle with believing that the system is giving me the right answer"*

They further reflected about *general use of the system*. Some mentioned that it was important to understand how to use the system, while others talked about how experience with using the system could be beneficial to understand how it worked in different settings. Three participants also reflected on user *errors*, such as users pushing the wrong button. Finally, they were also concerned that emergency managers could get too *dependent* on the system and not be able to operate without the system in case the DSS failed.

Group differences. There were no notable differences between the groups, except one participant from the MCDM group who said that one could blame a wrong decision on the experts that had developed the system, even though this was not of good practice to do. This reflection was not brought up by the participants in the ML group.

Operating context. The participants talked about the context that the system will be used in.

Here the participants reflected on the *current situation and systems* regarding existing. Two participants were concerned that false alarms regarding extreme weather reports occured more often now than before. Furthermore, the emergency managers talked about *technological development*, and explained that they experienced that technological advancements are coming to their sector, as it is in the rest of the society. They also reflected om *workplace considerations* such as issues related to implementation of new systems and proper training, as exemplified in the following extract:

"*It is not straightforward to implement new technology… The users must know why they are going to use the system. We see this here, changes – they are scary, terribly scary. So, I think training is very important*".

Finally, the participants talked about the *weather*, and explained that the complexity of the weather makes it difficult to predict future states, and that this complexity makes a DSS especially useful, but also challenging to develop.

Group differences. There were no differences between the two groups.

Users. This theme captured the participant's reflections on the meaning of different potential users of the DSS.

The participants discussed how *work-related experience* and *knowledge* could influence how a user would interact with the system. Experience was related to the familiarity with, and ability to assess the weather situation and the appropriate measures. Some were concerned that users with less experience would be more inclined to follow the advice from the DSS, as more experienced emergency managers can use their experience as a basis for decision making.

> *"… if you are totally inexperienced or uninterested, and the "the system told me to do it" and then you do it. Without thinking of the consequences at all."*

Related to knowledge participants argued that it was important to have knowledge in order to know how to assess a situation and make decisions, regardless of the presence of a DSS and that access to a DSS could not substitute knowledge and experience.

Group differences. There were no noticeable differences between the two groups.

Residual. The last theme was mainly about the ANYWHERE tool and the case the participants solved. The participants' feedback on the case will be addressed in the section about limitations. Some of the participants explicitly talked about the ANYWHERE *tool* and emphasized the opportunities for further development of the system, such as to use the system to communicate preparedness levels and forecasts to civilians.

Group differences. The participants from the ML groups were in general more positive towards the possibilities of the tool than the participants in the MCDM group.

Deductive analysis

The deductive analysis was based on data from the second part of the interviews. The interview guide for this part was aimed at capturing system performance and preexisting knowledge from the layer learned trust based on the model of K. A. Hoff and Bashir (2015). Specifically, the questions were aimed at attitudes, expectations, reputation, understanding of the DSS, experience with DSS within preexisting knowledge and validity of the advice, dependability, predictability and usefulness within system performance.

Table 2 gives an overview of the frequencies defined as the number of participants that mentioned each theme of preexisting knowledge and system performance. While the whole model was used during the analysis, only results for system performance and preexisting knowledge will be reported as this was the aim for this thesis. The coding for the whole model is in appendix H.

Topics that the participants talked about that were not captured by the model from K. A. Hoff and Bashir (2015), were coded, categorized and themed with the same approach as for the inductive analysis as described above. However, there was no additional information related to trust, therefore these results are not reported. Further, these statements did not add any additional information to the reflections from the inductive analysis so they were excluded from this study.

Table 2

Overview over deductive analysis results on layer of learned trust from model of K. A. Hoff and Bashir (2015)

Layers, factors and sources of variability	**Groups**	
	MCDM	ML
Learned trust	5	5
Preexisting knowledge	5	5
Attitudes	4	4
Expectations	5	5
Reputation	4	5
Experience	4	5
Understanding of system	4	5
System performance	5	5
Validity	5	5
Predictability	3	4
Dependability	3	4
Usefulness	5	5

Note. Bold = layers, bold italic = factors, italic = sources of variability. Frequencies indicate number of participants that mentioned each theme

Preexisting knowledge. Regarding *attitudes*, the participants were all positive towards the use of the DSS. While the participants had inquiries about the use and the pitfalls of the system, some of which are described in the previous section, their general attitude towards technology in this context was positive.

The participants had overall few *expectations* about the systems. They had not reflected particularly on it, but they were open, curious and excited. One participant, however, expected the system to be more primitive, messy and less user friendly:

> *"That you (developers) have crammed so much in there to make it as impressive and saleable as possible. Because then you can say "we can do this and we can measure this, and we can do that". People often misunderstand, quantity over quality and adds lots of functions and information (...) I want the correct information, I want the relevant information."*

Regarding *reputation*, the same participant thought other might think the same. The rest of the participants did not think that there would be reputations about such systems. One emergency manager guessed that the reputation would be mostly positive, if there were any. The participants also had little *experience* with similar systems and were not too concerned about the need for experience with the DSS or similar systems. One participant explained how lack of experience with such systems could be positive as one would be more open and not expect the system to be in one way or another. They were, however, more concerned about the need of general technological competency. They expressed that general experience with the use of technological systems was important to understand both how to use the specific DSS in the case and to use similar systems in a real working environment.

Regarding their *understanding* of the system, all the participants expressed an understanding of how the system worked based on the information they got before solving the case. Two participants thought that it was easier to make decisions when having such understanding because it helped them understand the

Group differences. While there was no major difference between the groups in relation to understanding, two participants emphasized that the basic principle of ML was easy to understand and that they liked that the system could learn. They expressed that their understanding of ML made it easier to trust the output because they knew the origin and process behind it.

System performance. Reliability was not a part of the study and the participants did not talk about reliability specifically. However, one participant sums up *reliability* and *validity*;

> "*If the system has shown to be correct many times, the user will develop trust. But if the system has been wrong many times, the user will develop mistrust. So, it is all or nothing for the system"*

Further, the participants found it difficult to say if the system gave correct advice as they did not know the outcome in the case. Some agreed with the system while others disagreed.

Most of the users found the system in large to be *predictable* – if they pushed one tab the correct information was displayed. Related to the advices, however, some found them surprising whilst other found them predictable, as reflected in the different opinions on the validity.

The participants thought the system was *dependabl*e in the case, or at least that they had no reason to think otherwise. They also had opinions about the use of technology outside of the context of the case:

> *"We have to be very aware of the vulnerability of using technology. You are dependent of power, Internet and all these things. What if that doesn't work?"*.

Overall, the participants perceived the system as *useful*. Some of the reasons were that the system provided decision support, gave higher consistency in decisions, removed feelings of uncertainty in decision making and emphasized that the system is needed as more extreme weather is expected.

Group differences. Few differences where present, but two participants in the MCDM group did not find the system useful and thought that the system had constraints outside of the case. This was amongst other because the system was perceived as inflexible. The participants argued that having to plan and adjust for different scenarios makes it impossible to use on a large scale where it is impossible to foresee all outcomes. Their impression was that the system only conjoined Yr and Varsom and that the advices were not useful.

Discussion

The aim of this thesis is twofold. Firstly, the thesis explores the initial reflections that the emergency managers have regarding the use of DSS. Secondly, it explores how the model from K. A. Hoff and Bashir (2015) works in this context.

The *first* part of the interviews was an open approach using the SWOT format, and the data material was analyzed inductively. The data gave an overview of the reflections of the emergency managers regarding the use of DSS. The *second* part of the interviews was based on the model from K. A. Hoff and Bashir (2015). The interview guide was aimed at two elements in the model; preexisting knowledge and system performance. The data material was analyzed deductively using the entire model, but only the results related to the research questions were reported. The results showed that the model is relevant in this context.

Lastly, the results show that there were some differences between the ML and the MCDM group in both the inductive and deductive analysis. The results will be summed up in the following section.

Summary of results

This thesis seeks to gain a better understanding of what characterizes the reflections of emergency managers and the differences between the groups testing ML and MCDM. The emergent themes in the analysis were aspects and characteristics of the system, interaction with the system, decision making, the operational context and the users of the system. There were some differences in the reflections of the emergency managers that can be attributed to the differences between the two systems. The differences were related to the way MCDM is developed using experts, the lack of transparency in the MCDM system and perception of limitations. Furthermore, there were more focus on how the system worked in the ML group compared to the MCDM group.

The themes indicate that the emergency managers have some common thoughts about the use of DSS in the context of extreme weather. The themes *aspects and characteristics of the system*, *users of the system* and *operational context* correlates with prior distinctions related to system, the situation and the users (Hancock et al., 2011; Schaefer et al., 2016). The theme *interaction with the system* was mainly about how to treat the output of the system. This is particularly important as the purpose of the system is to aid decisions. Lastly, the theme *decision making* covers the task of the emergency managers, namely to make decisions. It is important to incorporate these concerns in coming development of DSS in this context, and the topic will be further addressed below.

Further, this thesis explores learned trust and the differences between the two groups. These research questions were examined by a deductive approach. The model from K.A. Hoff and Bashir’s (2015) captured all the statements related to trust. There were however some issues in connection to placement of some of the statements related to preexisting knowledge. Further, there were some differences related to understanding of the system, perception of limitations and the role of the experts in developing MCDM systems.

The model from K.A. Hoff and Bashir’s (2015) is applicable in the context of DSS in the context of extreme weather. The issues related to the placement within preexisting knowledge will be addressed in the section about trust.

Reflections on the use of DSS

This section will discuss the main aspects of the themes from the inductive analysis. Due to the limitations of this thesis the author has chosen to elaborate on the most relevant findings within the different themes. The differences between the emergency managers' reflections in the two groups will be addressed in the end of this section.

Aspects and characteristics: This theme demonstrates the importance of three distinct characteristics of the DSS; (1) the information accessible in the system and the information the system bases its advices on, (2) how the advices are conveyed and (3) the consequences of using a system that can fail.

Firstly, the participants liked that the information originated from different sources (Yr and Varsom). These sources of information are known to the emergency managers and could have made it easier to understand and navigate in the DSS. Familiarity can also be linked to increased use of technology (Gefen, 2000), that certainly is preferable when implementing a DSS. Additionally, the participants also stated that presenting several sources of information at once made them more efficient as they could cross-check the information presented in the "risikobilde" with the weather and flood data from Yr and Varsom directly. This can indicate a need to assess the status for the system as the expressed that if was to check if the advices were correct. This finding can be linked to the absence of the prototypes confidence levels.

Antifakos, Kern, Schiele, and Schwaninger (2005) found that users with accesses to confidence levels did less verifying behaviour when the confidence levels were high. Conveying such confidence levels to the DSS could save the emergency managers the time of cross-checking with the raw-data form Yr and Varsom. However, it might lead users to stop verifying the system's advices and if the confidence levels are wrong, decisions can be made based solely on the system's advice. This, can in turn, result in blind trust in the system, as noted by some of the participants.

The emergency managers expressed that they were more effective in having access to more information directly in the system than if they had to look up other sources as they are used to. This finding is particularly interesting, seen as previous research tend to debate how much information to present at once and the implication for mental workload (Wickens et al., 2013). One explanation to this is that the amount of information was adjusted properly to their needs, and that the same amount of information ins real extreme weather event might be perceived as overwhelming. There is a trade off in having the right amount of information to let the emergency managers cross-check the advices, without having information overload.

Secondly, some of the participants stated that the recommendations where formulated in a way they perceived as too commanding. They were concerned that commanding recommendation could cause difficulties breaching with the recommended advice resulting in users following the advises even when they are not sure about their correctness.

While few studies have looked into implications of how recommendations are formulated (K. A. Hoff & Bashir, 2015), Nass, Moon, Fogg, Reeves, and Dryer (1995) found that the effect can vary according to the user's personality. That is, individuals with a more submissive personality were more inclined to follow advices from a computer formulated in a commanding way than individuals with a more dominant personality. To develop a DSS that are adapted to the user's personality would not be beneficial, however, this demonstrates the importance of being aware of the possible effect of different types of formulations when deciding on how to communicate recommendations in DSS.

Decision making: In this section, concerns about that automation oversimplifies the complex task of emergency management will be discussed.

Some of the participants indicated that extreme-weather events are highly complex situations that are difficult to automate. They felt that the DSS oversimplified the decision making process. One participant expressed that automation is more appropriate in an operative setting, such as in a cockpit where the environment is less ambiguous and there is a more direct relationship between a stimulus and a response. While this finding might indicate that the participants find the DSS less suitable in complex context such as extreme weather, it could be that this finding is related to this notion is the effect of familiarity. That is, automation in cockpits is common today, but automation in aviation was controversial earlies (Wiener, 1988). This might indicate that attitudes towards automation within extreme weather events might change over time.

This is, however, not to say that automation in extreme weather is not complex, nor that such concerns should not be addressed and such concerns underlines the notion that automation is not a simple additive task and that the design of DSS must be adapted to the users and their needs (Dekker & Woods, 2002). There is at the same time a balancing act of listening to the users and to develop systems that will be advantageous once in use as developing human centered automation does not mean that the users are absolute experts of every aspect of the system (Mica R. Endsley, 2015).

Interaction with the system: Within this theme there were some particularly interesting aspects related to (1) trust, (2) what measures the system includes and (3) dependency of the system.

Firstly, the participants were concerned about how to treat the output from the system. They emphasized the importance of performing personal assessments of the advices, and worried about the possibility of users blindly following the advice from the system. This indicates that trust is important in relation to DSS.

Secondly, in relation to how to treat the output, a participant expressed concerns as to what measures the system took. What if the festival in the case was important for the economy of the town, and that cancelling it would have consequences that the system did not weigh in? According to Friedman, Kahn, Borning, and Huldtgren (2013) human values should be implemented throughout the design process. Such values should not simply be defined as a monetary, but as elements that are important for persons or a group (Friedman et al., 2013), such as a festival.

The DSS used in the case recommended actions based on calculations of risk to infrastructure and health, but there might be other perspectives that should be included, such as social or cultural values. Previous studies have found that participants expected machines to make more utilitarian choices than humans, and the machines were thus judged more harsh when not making such choices (Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015). In contrast, another study found that humans blamed a human driver more for an accident compared to an autonomous driving system in accidents where both had equal blame (Awad et al., 2018). Both studies indicate that humans have some expectations to what machines should and should not do. This study adds to this notion and demonstrates that the users' values might influence their relation to the system. If the DSS is designed in a way that is not in line with these values, it may affect how users perceive and trust the advice that the DSS generates.

It can, however be difficult to automate value-based decisions and as Russell and Norvig (2016) puts it “a rational agent is one that does the right thing… but what is the right thing?” (p.36). Nevertheless, with the development of more complex and autonomous systems these value perspectives are important incorporate to make users comfortable with the systems.

Thirdly, the emergency managers were concerned that they could get dependent on the system. The participants were concerned that emergency managers, having used the DSS, would not be able to make decisions if the system were to fail. Bannon (2011) argue that humans are adaptive creatures that can recover, improve, improvise and compensate for a failing system. However, to ensure that they are fit to do this the emergency managers should understand the current situation. This is linked to the presence of raw-data, as mentioned earlier. That is, a benefit of providing the emergency managers the opportunity to look into of the raw-data is that it increase their understanding of the current situation, and not being out of the informational loop (Endsley & Kiris, 1995). By designing a system that opens for a broader understanding of the situation than to just present the recommendations there could be easier for the emergency managers to operate without the system if this were to fail.

Operating context: Three important topics connected to the context the DSS are to operate in were training and implementation will be discussed.

The participants talked about implementation and training. Training was underlined as important for users to understand how the system worked, how to use it and to help implementation. The emphasis on training is in line with previous studies, and training on how to use the system can be beneficial to understand the possibilities and limitations of automation to ensure the correct use (Mica R Endsley, 2015).

One of the participants expressed that any new system was met with resistance, but that if the emergency managers understood the purpose and potential of the system it was accepted. In relation to this, Harper and Utley (2001) emphasize that implementation should incorporate perspective on the organizational culture and to give employees decisions freedom in how to use the system. However, in the context of extreme weather, decisional freedom in how and when to use the system can be problematic. As mentioned by some of the participants, there could be difficult to know when to follow the system or not. Seen as the consequence of following vs. not following the DSS in this context can have large consequences, having guidelines that specify how to, and when to use the system is imperative.

Users*:* Related to the users of DSS, the emergency managers talked about the importance of experience and competence.

The participants emphasized the importance of the users' experience and knowledge about such events. They were concerned that users of the system with less prior experience with extreme weather events could be more prone to follow the advices from the system. Sanchez, Rogers, Fisk, and Rovira (2014) found that users of a collision avoidance system in agricultural machines who had experience within the domain had less reliance in the system compared to the inexperienced users. The authors argue that this might be a result of experienced users having a better understanding of the costs of crashing such machines than the inexperienced. Furthermore, Klein, Calderwood, and Clinton-Cirocco (1986) found that experience made it easier for fire ground commanders to recognize a situation and to act upon the cues they perceived. Bringing these findings into the context of extreme weather, one could argue that the experienced emergency managers weigh in more factors and have a better understand the consequences of different outcomes. Thus, a novice would not have the same basis of experience, making it more difficult to assess the situation.

This finding demonstrates that the use of DSS in relation to experience can be double-edged. Both experienced and inexperienced expressed that they got support in their decision making process, however both stated concerns related to how less experience could lead to overtrust in the system. To further underline this, they also expressed that the inexperienced might be less competent of assessing the recommendations from the system as they had less experience with handling extreme weather events.

On the other hand, research within expert predictions have shown experienced decision makers do not always make better predictions (Hardman, 2009). To develop an intuition for correct decisions, such as the fire ground commanders, one must work within an environment characterized by a stable relationship between an event and the outcome, in combination with appropriate feedback and practice (Kahneman & Klein, 2009). In many cases, extreme weather events have a stable relationship between an event and an outcome. But there can also exist outcomes that are first discovered in the evaluations after such incident, hence failing to give the emergency managers the appropriate feedback necessary for them to develop expertise.

As follows, it is more to decision making in this context than experience, and both experienced and inexperienced emergency managers might benefit from having a DSS if these concerns are considered and that the DSS is not used to as a substitute for experience and knowledge.

Differences in reflections between the MCDM and ML group. Overall, there were not major differences in the reflections between the two groups, but some minor once where present. This section will discuss differences related to (1) perceptions of limitations and (2) interest in and perceptions of logic and functioning of the systems and (3) general comments on the differences.

Firstly, there were some differences in the perception of limitations. Two participants from the MCDM group struggled to see the applicability of the system. One participant explained that even though the system handled the case well, the DSS ability to handle the number of factors that is present in a real-life situation was debatable. While MCDM systems are developed in close collaboration with experts, different factors are weighed in and the advices are fitted to the specific area in which the system is to operate. This can have been perceived as inflexible and with some obvious limitations. The ML group on the other side were more positive to the possibilities of the system. This demonstrates how the flexibility of ML might be viewed as preferable in this context.

Secondly, there were some differences in both focus on and understanding of the logic and functions of the two systems. One participant from the MCDM group emphasized that the system was easier to trust and felt safer to use because it had been developed by experts. This is in line with previous literature. Rasmussen, Pejterson, & Goodstein (1994) *as referred in John D. Lee and See (2004, p. 66)* contends that since technology has no intentionality it is the intentionality of the designers that are presented in a system and Parasuraman and Riley (1997) argues that trust in automation is trust in the designers of the system. This is an interesting view in relation to the use of ML were the developers are less engaged in a direct fashion. This could cause a lower degree of trust when using a system based on ML, as designers have had less intervention in the design process. However, no one from the ML group mentioned lack of trust or increased burden of responsibility. In general, the participants in both groups meant that the final responsibility was on the emergency managers and not the system. Furthermore, the participant that emphasized the way MCDM are developed had less experience within emergency management, indicating that it could simply have been an expression of lack of experience rather than a difference between the two systems.

Overall, there were not major differences. This might indicate that whether the system is based on MCDM or ML does not affect how the emergency managers perceive the system. Both group indicated that an understanding of the system was important, but there were little differences. One explanation for this finding can be that the users tested the same system that acted in the exact same way and there was no experienced difference or that whether the system is based in ML or MCDM are secondary to the users, and that performance, the possibilities and how the systems worked is seen as detrimental for their reflections and opinions about the DSS.

Summing up, this section has shown the participants’ reflections and discussed them in connection to existing research. The emergent themes show that the emergency managers have some common concerns, inputs and reflections, which demonstrates that automation comes with certain challenges. The next section will elaborate on the model of trust by K. A. Hoff and Bashir (2015).

Learned trust

This section will discuss the aspects of pre-existing knowledge and system performance based on the second part of the interviews which were analyzed deductively.

This will be done by comparing the results from the interviews to the model. Due to the constraints of this thesis the author has chosen to elaborate on some of the relevant topics. First, each aspect of pre-existing knowledge and system performance will be discussed following by the differences between the two groups.

Preexisting knowledge. Preexisting knowledge from the model of K. A. Hoff and Bashir (2015) consists of attitudes and expectations, reputation of the system or similar systems, experience with the system or similar systems and understanding of the system.

***Understanding of the system*.** According to K. A. Hoff and Bashir (2015), an understanding of how the system works can be beneficial for trust in automation. However, in the article the authors focus more on training to increase understanding of the system, than how understanding itself can be related to trust. During the interviews, understanding was more related to how the system worked.

The participants focused directly on understanding how the system functioned. They meant that having knowledge about how the system worked made it easier to understand how the system had concluded, and hence making it easier to trust the advices. In difference to K. A. Hoff and Bashir (2015) that emphasize understanding as in to understand the limitations and possibilities of the system, the participants talked about understanding how the system worked and the logic behind. These two types of understandings can be somewhat similar, but one is about understanding when the system works and not, and the other is understanding how the system works. This might indicate that for this context, understanding of how the system works is as important to trust as understanding under what circumstances the systems works and not.

The model from K. A. Hoff and Bashir (2015) have however two other sources of variability that can be related to the understanding of the system. These are system complexity and type of system within the layer of situational trust. Situational trust is related to the situation that the interaction happens in, according to the authors. However, K. A. Hoff and Bashir (2015) do not go into detail as to how these two affects trust other than that the type of system and how complex the automation is can affect trust. Further, as trust formation happens within the users, the understanding of how the system works can be understood as understanding of the logic, the type and the complexity of the automation. Indicating that understanding of trust in this context might be more complex than what K. A. Hoff and Bashir (2015) suggests related to an understanding of when the system works and not.

On the other hand, the participants received quite detailed information about how the different systems worked, prior to the case. This can have affected the participants, making them more focused on this topic.

Understanding how the system works can also have pitfalls. One participant expressed that if the logic behind the system is followed, the advice should be correct. This perception of how the system works might lead to overtrust in the system and not an appropriate amount of trust. Which underlines the importance of training to understand the limitations and possibilities of the system, as argued by K. A. Hoff and Bashir (2015).

Attitudes, expectations and reputation. The model indicates that attitudes, expectations and reputation can affect trust.

The participants had some inquires, as discussed previously, but in general they had a positive *attitude* towards the use of technology in the context of extreme weather. There seems to be limited research on attitudes in relation to trust in automation. One study indicates that positive attitudes can contribute to increased use (Workman, 2005). One participant mentioned positive attitude in relation to implementation, but no one talked about attitude directly in relation to use or trust in the DSS.

Attitudes can however be complex. One study examined the difference in explicit and implicit attitudes towards automation (Merritt, Heimbaugh, LaChapell, & Lee, 2013). The participants in the study tested an automation where the errors of the system were either unambiguous or ambiguous. The study showed that implicit and explicit attitudes did not correlate and Merritt et al. (2013) suggested that implicit trust is the trust that is used when error in automation is ambiguous. In a real operative setting errors from the DSS in an extreme weather context could be ambiguous suggesting that in this context the model from K. A. Hoff and Bashir (2015) might simplify this aspect.

The participants had few *expectations* prior to the test of the DSS. One participant expected the system to be complex with various functions and that it was difficult to navigate and understand. Further, several were curious, but overall there were not many expectations.

Mayer, Sanchez, Fisk, and Rogers (2006) found that people with higher initial expectations had higher initial reliance in the system. In this context, this might indicate that the emergency managers have less initial trust, which again be connected to the context where the decisions are critical.

When asked about the *reputation* of such systems the participants did not think there were any rumors or reputation, except for the same participant that expected the system to be complex and difficult to use indicated that this might be rumor.

In the model attitude and expectations are conjoined, but there might be more beneficial to separate these as these were separate topics in this study. Further, there might that perception of reputation and expectations are connected.

Experience with the system or similar systems. K. A. Hoff and Bashir (2015) emphasize that prior experience with the system or similar technology can affect trust. During the interviews the participants separated between the specific experience with the DSS they tested and the experience with similar systems. The relationship between the system, similar systems and general technological systems is somewhat unclear in the article of K. A. Hoff and Bashir (2015) and will be addressed here.

Firstly, the participants indicated that experience with similar technology had little effect on their use. The participants indicated there were little knowledge or experience from other systems that they could transfer to the DSS in this study.

One possible explanation can be that the systems they had prior knowledge of are qualitatively different. As discussed previously, Parasuraman et al. (2000) separates between four tasks to automate. These are (1) acquisition of information, (2) analysis and manipulation of information, (3) decision and action selection, and (4) action implementation. The systems the participants had prior experience with were computer systems that combine information, but they had little experience in interaction with systems that recommend action selections. Hence, they do not see how prior experience is relevant. Another explanation might be that experience with similar technology is less related to trust, as the participants were more concerned about general experience in using technological systems.

They emphasized that general experience in using technology was important to understand how to use the system and not be afraid of using it. This might indicate that the users were more concerned about technological competence, which is not included in the model from K. A. Hoff and Bashir (2015). However, self-confidence affects trust (John D. Lee & See, 2004) and is a part of the situational trust layer in the model from K. A. Hoff and Bashir (2015). It might be that beliefs about general competence in using technology can affect the confidence a user has in the interaction with a complex DSS.

One explanation as to why the participants were more concerned about general technological competence might be that users with higher technological skills are more experienced in interacting with complex technology. Therefore, it is not the type of technology that is the concern, but whether the user can manage technology in general. However, it might be that this view is affected by the complexity of the system. The DSS in this study was not very complex, and if the participants were to test a full version of a DSS more concerns about experience with similar complex systems might arise.

Related to experience with the specific DSS, the participants talked about experiences with the performance of the system. If the system gave wrong advice they would develop mistrust and if the system delivered correct advice they would develop trust. This finding is line with the result that performance affects trust (Schaefer et al., 2016) and will be further addressed below.

System performance. System performance consists of reliability, validity, predictability, dependability, usefulness and timing, difficulty and type of error. For this study error and reliability were excluded.

Validity. According to K. A. Hoff and Bashir (2015) validity is related to trust, which is supported in the literature. All the participants were clear on that whether one trusted the system or not depended on the validity of the advices.

In this study however, there was not consensus between the participants as to whether the advices from the DSS was correct or valid. This can indicate that emergency managers, based on the same information, can make different decisions and supporting that a DSS can be valuable in this context. On the other hand, the case provided limited information which may have lead the participants to fill in the details with their prior experience and knowledge, making 20mm rain much for one participant and little for another. Further, it can have been difficult for the participants to assess the validity of the advices because they were never exposed to the consequences of the extreme weather in the case. This made it difficult to determine if the actions they selected were appropriate.

Regardless, all the participants were clear on that the validity of the advice was in close relation to trust.

Usefulness. According to the model, usefulness is related to trust. K. A. Hoff and Bashir (2015) explains, however, that there has been little research on this specific relationship. The authors argue that if the users see the usefulness they are more likely to trust the system.

Mostly, the participants found the system useful. Especially because it conjoined information from different sources, offered support in the decision making process system, could give higher consistency in decisions, could remove feelings of uncertainty in decision making and that the system is needed as more extreme weather is expected. As previously discussed related to perceptions of limitations within MCDM, however, there were two participants that struggled to see the value of the DSS. When asked about trust the participants that struggled to see the usefulness were more concerned about the validity of the advices than usefulness.

This might indicate that usefulness is more related to other aspects than trust. Legris, Ingham, and Collerette (2003) found that usefulness was related to acceptance of technology. On the other hand, another study showed that if users perceive a DSS as useful, they were more inclined to rely on it (Parkes, 2009).

The link between usefulness and trust, and whether usefulness is more suitable in terms of technology acceptance, needs to be further addressed in this context.

Dependability and predictability. The model suggests that dependability and predictability is related to trust.

The participants found it difficult to assess dependability in the specific DSS tested in the case. However, in general they were concerned about the use of technology in relation to dependability as systems can fail or be out of power. As systems can fail, some of the participants stressed that users must avoid becoming too dependent of the DSS as discussed above. It seemed that the emergency managers where more concerned about not getting too dependent in relation to dependability. However, concerns about not to become too dependent of the system can be an expression of concerns about trust.

It was difficult for the participants to evaluate predictability as they had little to base expectations of what that the system should recommend. This was a limitation of the study. Therefore, this will not be further discussed.

Differences in trust between the MCDM and ML group. There were no major differences between the two groups related to trust and the model from K. A. Hoff and Bashir (2015).

The author of this thesis expected some differences between the groups as there have been major technological developments making systems based on ML and other forms of AI more autonomous than systems based on more traditional approaches.

Parasuraman and Riley (1997) stated that trust in the system reflects trust in the designers of the system and using ML this makes the designers of the system less involved. Therefore, one could expect trust to differ between the groups as experts are more involved in designing DSS based on MCDM.

There was one participant from the MCDM group that expressed that that it was easier to trust the system knowing that experts had been involved in the development supporting the view of Parasuraman and Riley (1997). At the same time, the other 9 did not emphasize this.

This indicates that trust might have shifted from trust in the designers to trust in the actual system, regardless of how it has been developed.

One explanation might be that as ML and AI have become widespread and available (DeBrusk, 2018) people are more familiar with it. Companies are using it to develop systems to target ads or analyzing the stock market (Jin, 2018). Further, personal assistants that users interact with, such as Amazon's Alexa and Apple's Siri, is based on AIby (Strayer, Cooper, Turrill, Coleman, & Hopman, 2017). Thus, users are familiar with AI and interacting with such systems and are less focused on how such systems are developed.

Another explanation is the context that the DSS are to operate in. As touched upon in the section about differences in reflections the emergency managers focused more on performance, the possibilities, how the systems worked and so on. All related to the functionality and operations of the system. This might indicate that for this context, the trust in the system is primarily focused on the performance, as reflected by the participants. Hence, making whether the system is MCDM or ML secondary. If it works, they have trust in it.

In relation to *understanding* however, there were some differences between the groups. Both groups indicated that to understand how the system functioned helped to trust the system. Still, the ML group seemed more interested in how the system worked compared to the MCDM group.

This was mostly related to how the DSS learned from actual use and how it would update advices based on new extreme weather events. This might indicate that the users of ML system had a different interaction with the system because of how the system worked compared to the users of MCDM.

On the other hand, because the focus was more related to the interaction the users had with the system, and not how the system learned from prior data, this might indicate that the users care about what affects their interaction and not how the system works and therefore not affecting trust, but rather was an expression of curiosity in interacting with a ML system.

Another explanation might be that they had some problems understanding how the system learned from their decisions whilst the use of earlier data was easier to comprehend.

Summing up this section have discussed learned trust from K. A. Hoff and Bashir (2015). The model show applicability within this context, with some differences related to experience, understanding and attitudes, expectations and reputation. Further, differences between the groups have been discussed. The data from the second part of the interviews supports the notion that trust is important in relation to the use of DSS in this context. The next section will elaborate on the use of the results can be used outside of this study.

General discussion

This thesis has so far covered how emergency managers work during extreme weather events, the use and challenges of automation in this context, and the importance of trust and human centered automation in order to achieve a sufficient relationship between the emergency managers and the technology. But how do the findings discussed in the thesis matter? How can we apply them outside of this study? The next section seeks to put the results of this study in a broader context.

As the aim of this study is twofold, there are some practical implications related to further development of DSS in extreme weather events and some theoretical implications related to trust and the model of K. A. Hoff and Bashir (2015).

Practical implications. Many design inputs and perspective are of great importance to achieve a good human centered automaton. This section will cover implications for design based on this study.

Firstly, this study demonstrates that importance availability of raw data. This is also in line with Atoyan, Duquet, and Robert (2006) who argue that access to raw data as this is beneficial to build trust and to let users validate the output from the system, and should therfore be available to the user. To include raw data in the system can also be time-saving by letting the users cross-check them with the results directly in the system. It can also be beneficial to include sources of information that are known to the users in a format they are familiar with. One issue with this, however, might be the need to make formal agreements which can lock the designers into using one system for a certain time. This can make the system less flexible.

Secondly, designers should be aware of how output of the system is communicated. The way the output is communicated can be interpreted different by different users. One possible solution to mitigate the effects of the communication style might be to provide the users with a detailed description of the rationale behind the advices from the system. Knowing why the system recommends a specific action can make it easier for the users to understand the advice. It can also make the job of verifying the advices easier. This was attempted in the prototype in this study, but the description of the status for each input factor was not sufficient. One possibility is to communicate the output in a more human form as a written or spoken message.

However, in using ML there are “black box” problems. This means that only the input and the output of the system are apparent to the user. In the context of extreme weather, the user will only see the input variables such as wind and rain and the advice which is the output.

Current research is looking into how ML can contribute with a full explanation of the rationale behind the output (Grunning, 2018) This is an important next step to utilize the full potential of ML in DSS.

Thirdly, based on this study, there is no need for specific considerations in utilizing systems based on ML in relation to how users interact with such systems. However, organizations must be aware of having proper data to ensure that the system is not trained on faulty data, hence making wrong recommendations(DeBrusk, 2018).

Lastly, proper training and implementation is recommended. The training should clarify possibilities and limitations of the DSS so that the users can understand when they can trust the system and when they cannot. Further, the training should deal with concerns related to manipulation and systems failing. Guidelines for how to react in these scenarios should be made available throughout the organization. In addition, organizations should clarify the status of the system to avoid situations where emergency managers follow the system regardless of their own assessments.

Theoretical implications. The second aim of this study was to examine learned trust from the model of K. A. Hoff and Bashir (2015) in the context of DSS during extreme weather events.

In general, the model fit the data well. Most of the statements fell within the model indicating that the model is relevant in this context. There where however some aspects that the model did not fully capture. These aspects can be important to be aware of and should be addressed in further development of the model.

Firstly, the model adjoins attitude and expectations, and leaves reputation as a single item. In the study, expectations and attitudes seemed to be two separate aspects of pre-existing knowledge. Moreover, expectations were somewhat guided by rumors for the one participant. This might indicate that there are some nuances within attitudes, expectations and reputation that the model does not capture.

Secondly, the participants were more concerned about general technological experience than system specific experience. One possible explanation is that the participants were concerned about technological competency in general. Hence, technological competency might be included as a separate aspect in the model. Another explanation is that it such skills affect self-confidence.

Thirdly, the model is somewhat unclear on the understanding of how the system works. This can be both related to type of system or to understanding, but as discussed K. A. Hoff and Bashir (2015) emphasizes training in relation to understanding of system and does not go into what type of system covers. The understanding of system should be further clarified.

Lastly, the relationship between usefulness and trust should be further addressed as this relationship were not clear within this study.

In sum this section has elaborated on the practical and theoretical implications of this study. The practical implications seek to contribute to a human centered development of DSS to mitigate the challenges of automation. The theoretical implications seek to contribute to the field of trust in automation, and specifically to the model of K. A Hoff and Bashir (2015).

Limitations

The following section will address the limitations of this study and comment on steps that were taken to minimize the impact of these limitations.

Differences between MCDM and ML. The real difference between the two groups was the explanation of how the system worked, and not the actual algorithm behind the system. This means that the users did not interact with a working ML algorithm. However, as the system would look the same regardless of the logic behind it and would give the same advices in the small context in the case, it was decided that it was sufficient to alter the explanations of the system. The two explanations were made in collaboration with experts within MCDM and ML and tested during the pilots prior to the interviews. This was to ensure that the explanations were similar in form and structure. However, it remains that the real difference was in the explanation.

Simulation. A limitation is that the study was done in a simulation rather than in an actual working environment. One problem with simulations is that the actors may decide to rely more on the advice from the system than they would have in a real situation (*Guidelines for Trust in Future ATM systems: Measures*, 2003). Dahl, Alsos, and Svanæs (2010) argues that for simulation to be a part of a design process appropriate fidelity must be applied. However, the primary function of the simulation was to serve as a basis for the interview and not to represent the real world. It was communicated to the participants that the system was a prototype, and that the focus was on how they experienced the use of DSS.

Further, due to how the case and prototype worked, the participants found it difficult to assess the dependability predictability of the DSS which limited the data about these aspects in relation to learned trust.

The participants were asked if they experienced the case as realistic. Their feedback was mostly that it was realistic and that such incidents can happen. Some of the participants expressed that there was a lack of collaboration with others, which they are used to in real incidents. Some indicated that recommended evacuation of the hospital and schools where options that are very seldom carried out, but others did not think so. The recommendations were reviewed by a panel of experts within the field prior to the test to mitigate this. One participant also felt that the answer sheet had few action selections and the case were limited. All in all, the participants expressed that the case was satisfactory to test the DSS.

Interviews. A semi-structured interview has a weakness in that different follow-up questions can be asked in different interviews. This can affect the reliability (Kvale & Brinkmann, 2009) which is in this study further weakened as four of the interviews were done by two different interviewers. To mitigate this, the first six interviews were done by the two interviewers together. This helped develop a similar style in follow-up questions. In addition, by doing the first six interviews together, the interviewers became more conscious about their follow-up questions and developed an interview style that were more consistent throughout the study than what might have happened with only one interviewer.

Sample. The sample consisted of 10 participants. According to Pratt (2009) there are no agreed upon number of how many participants that are appropriate to reach saturation, but others claim that demonstrated “saturation” is often reached between 10-15 (Guest, Bunce, & Johnson, 2006). The saturation was monitored by the author of this thesis and the researcher at SINTEF. It was a continuous dialog between the interviewers about whether the participants contributed with new information. After 10 participants, it was decided that this number was sufficient. One argument could be that there should be more than five participants in each group, but as saturation was reached it was decided that there was no need for additional participants.

Furthermore, the sample was constrained to a Norwegian context. This will have its limitations, but as this is an early study of a prototype that will be implemented amongst other in a Norwegian setting, it was deemed fit. Chien, Sycara, Liu, and Kumru (2016) found that people from Turkey hade less initial trust in automation than people form USA indicating that there can be cultural differences.

Future studies

This study has enlightened some of the challenges and opportunities that a DSS in emergency management during extreme weather events can have. Further, trust has been examined through the model by K. A. Hoff and Bashir (2015). The study has shown some interesting results regarding further development of DSS, trust and the differences between MCDM and ML systems. However, more research is needed as more complex automation is expected.

Firstly, this study examines only an excerpt of the model from K. A. Hoff and Bashir (2015). Because this is a vast model, more research is needed to further examine the relations suggested by the authors. The model is useful to guide development of systems, organize existing and new research and to explain trust. However, this study uncovered some aspects of the model that needs more research.

Secondly, whilst ML and AI become more and more present in our society, it is still quite new. Future research should continue to examine how the interaction between humans and machines is affected by systems becoming more and more autonomous. This will be especially interesting with the development of more advanced systems such as voice-controlled systems, chatbots and systems based on augmented reality. Such systems can further complicate and/or ease the human-machine interaction.

Thirdly, one limitation in this study was that the participants did not see the extreme weather and the consequences and could there for not assesses the performance of the tool. Future studies can utilize complex simulations and gamification methods using either games presented on a computer or in VR. This can create a more dynamic study setting where the users experience the consequences of the forecasted weather and of their decisions.

Fourthly, this study provides initial reflections and a modest insight in trust for a limited time in a simulated setting. Future studies should seek to study the implementation and use of advanced DSS in a natural setting.

Conclusion

This study provides insights into trust and further development of DSS. It has identified five themes that capture the reflections emergency managers have about the use of DSS in the context of extreme weather. The findings indicate that access to raw-data, consciousness on how systems communicate recommendations and proper training that address concerns about manipulation and failing systems, in addition to the possibilities and limitations of the DSS, can be beneficial. Further, this study shows that pre-existing knowledge and system performance from the model of trust by K. A. Hoff and Bashir (2015) is applicable in this context. There were, however, some issues related to aspects of experience with similar systems and understanding. In addition, the placement of attitudes, expectations and reputation differed from the model. Lastly, there were no major differences in reflections or learned trust between ML and MCDM, indicating that how the system is developed is secondary to the users in this context. The study is, however, explorative and based on a small sample and further research is needed to address the findings.

References:

Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005). *Towards improving trust in context-aware systems by displaying system confidence.* Paper presented at the Proceedings of the 7th international conference on Human computer interaction with mobile devices & services.

ANYWHERE. (2017). About Us – Objectives. Retrieved from http://anywhere-h2020.eu/about-us/objectives/

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: an empirical study. *Sociology, 31*(3), 597-606. doi:10.1177/0038038597031003015

Atoyan, H., Duquet, J.-R., & Robert, J.-M. (2006). *Trust in new decision aid systems.* Paper presented at the Proceedings of the 18th Conference on l'Interaction Homme-Machine.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J., Shariff, A., . . . Rahwan, I. (2018). Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation. *arXiv preprint arXiv:1803.07170*.

Aybar-Ruiz, A., Jiménez-Fernández, S., Cornejo-Bueno, L., Casanova-Mateo, C., Sanz-Justo, J., Salvador-González, P., & Salcedo-Sanz, S. (2016). A novel Grouping Genetic Algorithm–Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs. *Solar Energy, 132*, 129-142. doi:10.1016/j.solener.2016.03.015

Bannon, L. (2011). Reimagining HCI: Toward a more human-centered perspective. *interactions, 18*(4), 50-57. doi:10.1145/1978822.1978833

Bawden, D., & Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science, 35*(2), 180-191. doi:10.1177/0165551508095781

Berg, O. T. (2015, 29.04.2015). fylkeskommune. Retrieved from https://snl.no/fylkeskommune

Biffis, E., & Chavez, E. (2017). Satellite data and machine learning for weather risk management and food security. *Risk Analysis, 37*(8), 1508-1521. doi:10.1111/risa.12847

Bohanec, M. (2017). DEXi:A Program for Multi-Attribute Decision Making Retrieved from http://kt.ijs.si/MarkoBohanec/dexi.html

Boyatiz, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development.* . Thousand Oaks, CA: Sage Publications.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. doi:10.1191/1478088706qp063oa

Bründl, M., Etter, H.-J., Steiniger, M., Klingler, C., Rhyner, J., & Ammann, W. (2004). IFKIS-a basis for managing avalanche risk in settlements and on roads in Switzerland. *Natural Hazards and Earth System Sciences, 4*(2), 257-262. doi:10.5194/nhess-4-257-2004

Chien, S.-Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014). *Towards the development of an inter-cultural scale to measure trust in automation.* Paper presented at the International Conference on Cross-Cultural Design.

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). *Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Clarke, C., & Milne, R. (2001). *A national evaluation of the PEACE Investigative Interviewing Course*: Home Office London.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). *Trust in decision aids: A model and its training implications.* Paper presented at the in Proc. Command and Control Research and Technology Symp.

Dahl, Y., Alsos, O. A., & Svanæs, D. (2010). Fidelity considerations for simulation-based usability assessments of mobile ICT for hospitals. *Intl. Journal of Human–Computer Interaction, 26*(5), 445-476. doi:10.1080/10447311003719938

DeBrusk, C. (2018). The Risk of Machine-Learning Bias (and How to Prevent It). Retrieved from https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/?utm_source=facebook&utm_medium=social&utm_campaign=sm-direct

Dekker, S. W., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human–automation co-ordination. *Cognition, Technology & Work, 4*(4), 240-244. doi:10.1007/s101110200022

DSB Rapport. (2015). *Evaluering av forebygging og håndtering av flommen på Vestlandet høsten 2014*. Retrieved from https://www.fylkesmannen.no/Documents/Dokument

FMSF/Samfunnstryggleik/%C3%98vingar/Kommune%C3%B8ving 17.06.15/Evaluering flom Vestlandet 2014.pdf

Dyson, R. G. (2004). Strategic development and SWOT analysis at the University of Warwick. *European journal of operational research, 152*(3), 631-640. doi:10.1016/S0377-2217(03)00062-6

Ekstremværrapport Hendelse: Synne. (2015). *Ekstremværrapport Hendelse: Synne, 4-6.12.2015*. Retrieved from

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64. doi:10.1518/001872095779049543

Endsley, M. R. (2015). Building resilient systems via strong human systems integration. *Defense AT&L*, 7-12.

Endsley, M. R. (2015). Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making, 9*(1), 4-32.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(2), 381-394. doi:10.1518/001872095779064555

Erdogan, G., Refsdal, A., Nygård, B., Randeberg, B. K., & Rosland, O. P. (2017). *Experiences from Developing an Algorithm to Support Risk-Based Decisions for Offshore Installations.* Paper presented at the Proceedings of the 14th International Symposium on Operational Research.

Eskandarpour, R., & Khodaei, A. (2017). Machine learning based power grid outage prediction in response to extreme events. *IEEE Transactions on Power Systems, 32*(4), 3315-3316. doi:10.1109/TPWRS.2016.2631895

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1-4. doi:10.11648/j.ajtas.20160501.11

Field, C. B., Field, C. B., Barros, V., Stocker, T. F., Qin, D. J., Dokken, D. J., . . . P.M, M. (Eds.). (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge, UK and New York, NY, USA,: Cambridge University Press.

FinansNorge. (2018). Naturskadestatistikk. *Naturskadestatistikk.* Retrieved from https://www.finansnorge.no/statistikk/skadeforsikring/Naturskadestatistikk-NASK/

Flores, F., Graves, M., Hartfield, B., & Winograd, T. (1988). Computer systems and the design of organizational interaction. *ACM Trans. Inf. Syst., 6*(2), 153-172. doi:10.1145/45941.45943

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Dordrecht: Springer Netherlands.

Furevik, E. (2012). *Sivilbeskyttelsesloven - En praktisk og jurisik veileder*. Oslo: Universitetsforlaget.

Fylkesmannen i Oppland. (2014). *Del I – Evaluering av forberedelser til vårflom i 2014 Del II – Oppfølging av rapporten etter flommen i 2013* Retrieved from https://www.fylkesmannen.no/Documents/Dokument FMOP/Samfunnssikkerhet/Beredskap/flomrapport 2014.pdf

Fylkesmannens samfunnssikkerhetsinstruks. (2015). Justis- og beredskapsdepartementet Retrieved from https://lovdata.no/dokument/INS/forskrift/2015-06-19-703.

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega, 28*(6), 725-737. doi:10.1016/S0305-0483(00)00021-9

Grunning, D. (2018). Explainable Artificial Intelligence (XAI). Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods, 18*(1), 59-82. doi:10.1177/1525822X05279903

Guidelines for Trust in Future ATM systems: Measures. (2003). Retrieved from

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527. doi:10.1177/0018720811417254

Hardman, D. (2009). *Judgement and Decision Making Psychological Perspectives*. West Sussex, UK: British Psychological Society and Blackwell Publishing.

Harper, G. R., & Utley, D. R. (2001). Organizational culture and successful information technology implementation. *Engineering Management Journal, 13*(2), 11-15. doi:10.1080/10429247.2001.11415111

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research *Advances in psychology* (Vol. 52, pp. 139-183): Elsevier.

Ho, W. (2008). Integrated analytic hierarchy process and its applications–A literature review. *European journal of operational research, 186*(1), 211-228. doi:10.1016/j.ejor.2007.01.004

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407-434. doi:10.1177/0018720814547570

Hoff, T. (2009). Mapping the Organizational Climate for Innovation: Introducing SWOT as a Process Based Tool. In W. Østreng (Ed.), *Confluence* (pp. 75-79): Centre for Advanced Study – Norwegian Academy of Science and Letters.

Hoff, T., Flakke, E., Larsen, A.-K., Lone, J. A., Bjørkli, C. A., & Bjørklund, R. A. (2009). On the validity of M-SWOT for innovation climate development. *Scandinavian Journal of Organizational Psychology, 1*(1).

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53-71. doi:10.1207/S15327566IJCE0401_04

Jin, G. Z. (2018). Artificial Intelligence and Consumer Privacy *Economics of Artificial Intelligence*: University of Chicago Press.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*(6), 515. doi:10.1037/a0016755

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). *Rapid decision making on the fire ground.* Paper presented at the Proceedings of the Human Factors Society Annual Meeting.

Kvale, S., & Brinkmann, S. (2009). *Det kvalitative forskningsintervju* (2 ed.). Oslo: Gyldendal Norsk Forlag AS.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. doi:10.1518/hfes.46.1.50_30392

Lee, J. D., & Seppelt, B. D. (2009). Human Factors in Automation Design. In S. Y. Nof (Ed.), *Springer handbook of automation*. Verlag Berlin Heidelberg: Springer Science & Business Media.

Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & Management, 40*(3), 191-204. doi:10.1016/S0378-7206(01)00143-4

Lone, J. A., Riege, A. H., Bjørklund, R., Hoff, T., & Bjørkli, C. (2017). The relationship between the broader environment and the work system in a university setting: a systems approach. *Studies in Higher Education, 42*(1), 3-22. doi:10.1080/03075079.2015.1034259

Lubchenco, J., & Karl, T. R. (2012). Predicting and managing extreme weather events. *Physics Today, 65*(3), 31. doi:10.1063/PT.3.1475

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301. doi:10.1080/14639220500337708

Madsen, M., & Gregor, S. (2000). *Measuring human-computer trust.* Paper presented at the 11th australasian conference on information systems.

Maguire, M. (2001). Methods to support human-centred design. *Int. J. Hum.-Comput. Stud., 55*(4), 587-634. doi:10.1006/ijhc.2001.0503

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). *Sacrifice one for the good of many?: People apply different moral norms to human and robot agents.* Paper presented at the Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction.

Mardani, A., Jusoh, A., MD Nor, K., Khalifah, Z., Zakwan, N., & Valipour, A. (2015). Multiple criteria decision-making techniques and their applications–a review of the literature from 2000 to 2014. *Economic Research-Ekonomska Istraživanja, 28*(1), 516-571. doi:10.1080/1331677X.2015.1075139

Marsh, S., & Dibben, M. R. (2003). The role of trust in information science and technology. *Annual Review of Information Science and Technology, 37*(1), 465-498. doi:10.1002/aris.1440370111

Mayer, A. K., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2006). *Don't let me down: The role of operator expectations in human-automation interaction.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Mays, N., & Pope, C. (1995). Rigour and qualitative research. *BMJ: British Medical Journal, 311*(6997), 109. doi:10.1136/bmj.311.6997.109

McGurk, B. J., Carr, M. J., & McGurk, D. (1993). *Investigative interviewing courses for police officers: An evaluation*: Citeseer.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS), 2*(2), 12. doi:10.1145/1985347.1985353

Meld. St. 10 2016-2017. *Risiko i et trygt samfunn Samfunnsikkerhet*. Oslo: Justis- og beredskapsdepartementet Retrieved from https://www.regjeringen.no/no/dokumenter/meld.-st.-10-20162017/id2523238/.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don’t know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520-534. doi:10.1177/0018720812465081

Meyer, M. D., Rowan, E., Savonis, M. J., & Choate, A. (2012). Integrating extreme weather risk into transportation asset management. *American Association of State Highway and Transportation Officials*.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5-6), 527-539. doi:10.1016/S0020-7373(87)80013-5

Nass, C., Moon, Y., Fogg, B., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International journal of human-computer studies, 43*(2), 223-239. doi:10.1006/ijhc.1995.1042

National Transportation Safety Board. (1997). *MARINE ACCIDENT REPORT GROUNDING OF THE PANAMANIAN PASSENGER SHIP ROYAL MAJESTY ON ROSE AND CROWN SHOAL NEAR NANTUCKET, MASSACHUSETTS JUNE 10, 1995*. Retrieved from WASHINGTON, DC https://ntsb.gov/investigations/AccidentReports/Reports/MAR9701.pdf

Nau, D. S. (2009). Artificial Intelligence and Automation. In S. Y. Nof (Ed.), *Springer handbook of automation*. Verlag Berlin Heidelberg: Springer Science & Business Media.

Nedre Eiker Kommune. (2013). Evaluering av Nedre Eiker kommunes håndtering av flommen i mai 2013.

Nof, S. Y. (Ed.) (2009). *Springer handbook of automation*. Verlag Berlin Heidelberg: Springer Science & Business Media.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230-253. doi:10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30*(3), 286-297. doi:10.1109/3468.844354

Parkes, A. (2009). Persuasive decision support: Improving reliance on decision support systems. *PACIS 2009 Proceedings*, 11.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Beverly Hills, CA: SAGE Publications, inc.

Perrow, C. (1999). *Normal accidents: Living with high risk technologies*. Princeton, New Jersey and Chichester, West Sussex: Princeton university press.

Polit, D., & Tatano Beck, C. (2010). *Generalization in Quantitative and Qualitative Research: Myths and Strategies* (Vol. 47).

Pratt, M. G. (2009). From the editors: For the lack of a boilerplate: Tips on writing up (and reviewing) qualitative research. *Academy of management journal, 52*(5), 856-862. doi:10.5465/amj.2009.44632557

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*: Malaysia; Pearson Education Limited.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science, 15*(2), 134-160. doi:10.1080/1463922X.2011.611269

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377-400. doi:10.1177/0018720816634228

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., . . . X., Z. (2012). Changes in Climate Extremes and their Impacts on the Natural Physical Environment. In C. B. Field, C. B. Field, V. Barros, T. F. Stocker, D. J. Qin, D. J. Dokken,

K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, S. K. Allen, M. Tignor, & M. P.M (Eds.), *Changes in Climate Extremes and their Impacts on the Natural Physical Environment*. Cambridge, UK and New York, NY, USA,: Cambridge University Press.

Sivilbeskyttelsesloven. (2010). *Lov om kommunal beredskapsplikt, sivile beskyttelsestiltak og Sivilforsvaret*. Retrieved from https://lovdata.no/dokument/NL/lov/2010-06-25-45.

Stephenson, D. B. (2008). Definition, diagnosis, and origin of extreme weather and climate events. In H. F. Diaz & R. J. Murnane (Eds.), *Climate Extremes ans Society* (Vol. 348, pp. 11-23). New York: Cambridge University Press.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 71*(2), 93. doi:10.1037/cep0000104

Tedre, M. (2008). FEATURE What should be automated? *interactions, 15*(5), 47-49. doi:10.1145/1390085.1390096

Terti, G., Ruin, I., Kalas, M., Lorini, V., Sabbatini, T., & Alonso, A. (2017). ANYCaRE: ANYWHERE Crisis and Risk Experiment - A role-playing game for crisis decision-making and communication in weather-related hazards.

Wickens, C., Lee, J., Liu, Y., & Gordon-Becker, S. (2013). *An Introduction to Human Factors Engineering* (2 ed.). Noida: Pearson India Education Services.

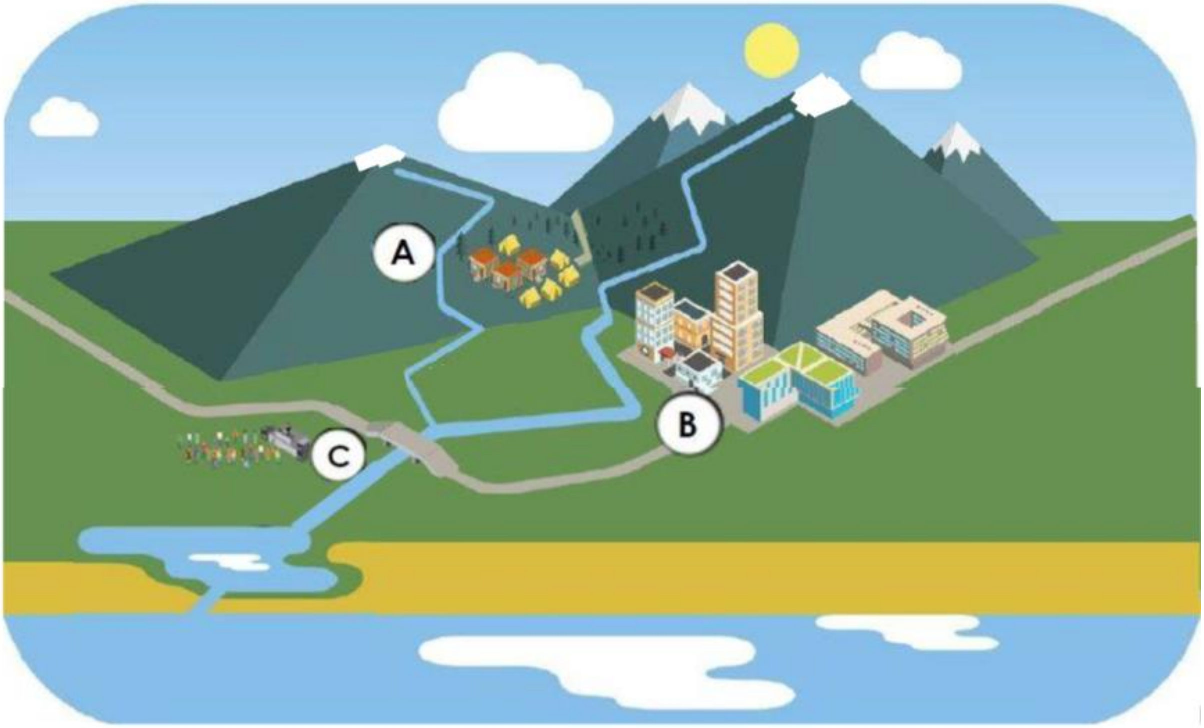
Wiener, E. L. (1988). Cockpit automation *Human factors in aviation* (pp. 433-461): Elsevier.

Workman, M. (2005). Expert decision support system use, disuse, and misuse: a study using the theory of planned behavior. *Computers in Human Behavior, 21*(2), 211-231.

Zachary, W. W. (1988). Decision support systems: Designing to extend the cognitive limits *Handbook of human-computer interaction* (pp. 997-1030): Elsevier.

Zhang, X., Zwiers, F. W., Li, G., Wan, H., & Cannon, A. J. (2017). Complexity in estimating past and future extreme short-duration rainfall. *Nature Geoscience, 10*(4), 255-259. doi:10.1038/ngeo2911

APPENDIX A



Området A: Boliger og campingplass

Området B: Sykehus og skole

Området C: Festivalområde og bro. Festival klokken 18:00

APPENDIX B

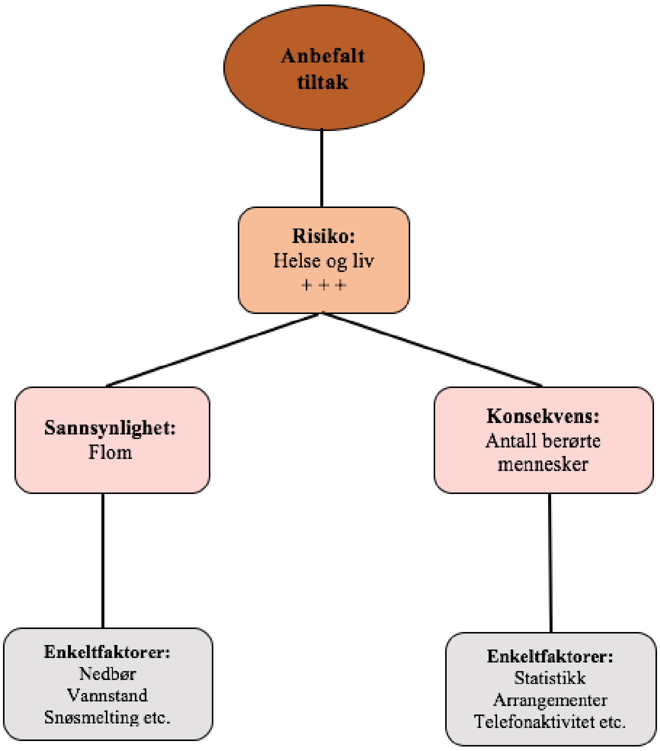


Illustration MCDM

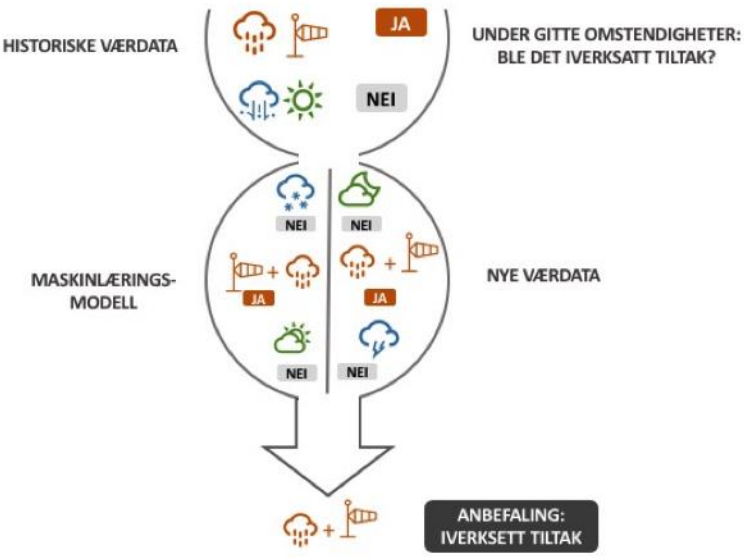
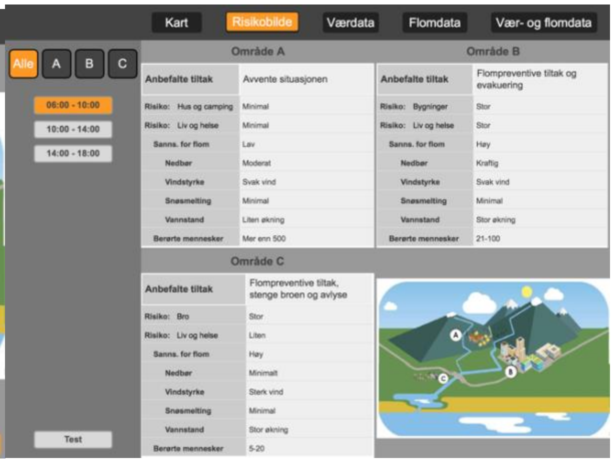


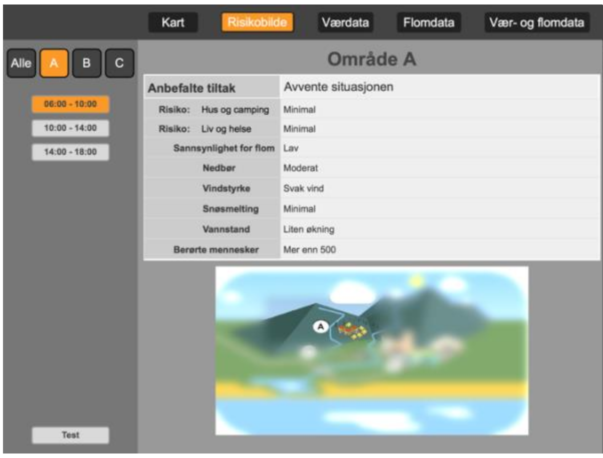
Illustration ML

APPENDIX C

Map



Example of ”all” function in “Risikobilde”

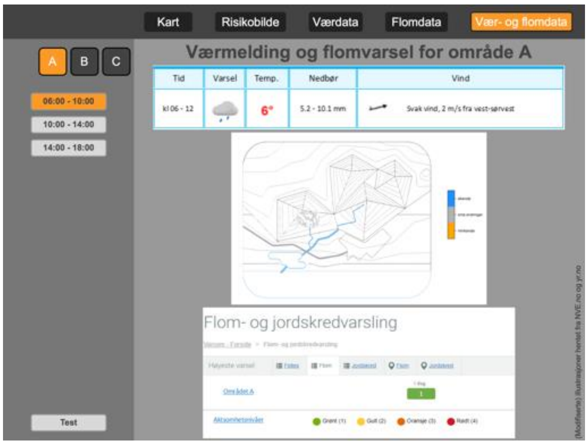


Advice tab MCDM

Advice tab machine learning

Weather report specific area

Flood cast specific area



Combined weather and flood cast

APPENDIX D

ANYWHERE

Invitasjon til deltakelse i forsøk og intervju

SINTEF skal, som en del av et større EU-prosjekt, utvikle et verktøy som skal hjelpe aktører i forbindelse med større natur- og ekstremværhendelser. For å sikre at dette verktøyet blir best mulig ønsker vi å snakke med potensielle sluttbrukere. Vi har kontaktet deg fordi du er en sentral aktør innen beredskapen ved større natur- og ekstremværhendelser, og dermed en potensiell sluttbruker. Selve forsøket samt intervjuet vil ta omlag 90 minutter og gjøres på et sted som passer deg. Vi håper at dette høres interessant ut og at du ønsker å være med.

ANYWHERE er et forskningsprosjekt finansiert av EU Horizon 2020, med en budsjettramme på 18 millioner Euro og med partnere fra 11 ulike europeiske land. I Norge er SINTEF og RAKOS medlemmer av prosjektet. Formålet til ANYWHERE er å gi organisasjoner som har en rolle i å håndtere natur- og ekstremværhendelser bedre verktøy for å forutse hvilke konsekvenser denne type hendelser kan føre til, og dermed gjøre organisasjonene bedre i stand til å håndtere slike hendelser. For mer informasjon om ANYWHERE, se prosjektets nettside www.anywhere-h2020.eu.

I Norge skal prosjektet fra juli 2018 gjennomføre en ettårig pilot-studie med to kommuner – Stavanger og Sauda – hvor verktøyene som utvikles i prosjektet vil bli testet ut i virkelige omgivelser. Prosjektet skal før dette, basert på de behovene potensielle sluttbrukere har, tilpasse løsninger som vil støtte prosessen med å gjøre bedre risikobaserte beslutninger før og under natur- og ekstremværhendelser. For å kartlegge hvilke brukerbehov og –krav verktøyet må ta hensyn til gjennomfører SINTEF og RAKOS forsøk og intervjuer med potensielle sluttbrukere.

Det vil i forbindelse med dette også være sentralt å undersøke hvordan sluttbrukere forholder seg til teknologisk beslutningsstøtte. Mats Ekre Wang er masterstudent ved Universitetet i Oslo samt tilknyttet ANYWHERE-prosjektet, og vil skrive sin oppgave rundt dette.

Om deltakelsen
Forsøket og intervjuet vil ta ca. 30-40 minutter hver, men vi ber deg sette av omlag 90 minutter. Forsøket går ut på å løse en case knyttet til en værrelatert hendelse. Her skal du fatte noen beslutninger og du vil under casen ha tilgang til et risikobasert beslutningsstøttesystem som skal bistå deg i beslutningsprosessen. Etter at casen er avsluttet vil vi gjennomføre et intervju knyttet til din opplevelse med dette systemet.

Behandling av data og personopplysninger
Intervjuet vil bli spilt inn på en lydopptaker for så å bli transkribert (skrevet ut i tekst). Dette gjør vi for å lettere kunne analysere dataene i etterkant. Innsamlet data vil bli behandlet konfidensielt, og kun være tilgjengelig for prosjektpartnere fra SINTEF og RAKOS. Den informasjonen du gir i intervjuet og dine personopplysninger vil bli oppbevart hver for seg.

SINTEF

Personopplysninger vil kun være tilgjengelig for den ansvarlige for studien ansatt ved SINTEF og masterstudenten (Marita Skjuve og Mats Ekre Wang). Under ingen omstendigheter vil andre enn disse få tilgang til dine personopplysninger.

Informasjonen du gir vil bli benyttet i Mats Ekre Wang sin masteroppgave, samt i prosjektrelaterte presentasjoner, rapporter og vitenskapelige publikasjoner. I alle slike publikasjoner vil du som deltaker være anonym, det vil si at det ikke blir inkludert informasjon som gjør det mulig å spore informasjonen tilbake til deg som enkeltperson. Datamaterialet anonymiseres fullstendig senest ved prosjektslutt i 2019, og senest på dette tidspunktet vil navnelisten over deltakere slettes.

Studien er meldt inn til NSD – Norsk Senter for Forskningsdata. Vi understreker at din deltakelse i denne studien er frivilling. Du kan når som helst velge å trekke tilbake ditt samtykke uten å oppgi noen grunn for dette, ved å kontakte Marita Skjuve på telefon eller e-post. I så tilfelle slettes alle data vi har om deg, og den vil ikke bli benyttet i studien.

Med vennlig hilsen,

Marita Skjuve
Forsker ved SINTEF digital
Tlf: 936 00 565
E-post: marita.skjuve@sintef.no

Mats Ekre Wang
Masterstudent ved UiO
Tlf.: 464 45 864
E-post: mats.wang@gmail.com

SINTEF

APPENDIX E

TIDSROM 06:00 - 10:00 **Tiltak**	Sett kryss for tiltak
1. ***Avvente:*** **Ingen handlinger utføres**	
a. Du følger væroppdateringene og monitorer situasjonen for område A	
b. Du følger væroppdateringene og monitorer situasjonen for område B	
c. Du følger væroppdateringene og monitorer situasjonen for område C	
2. ***Flompreventive:*** **Du bestemmer deg for å iverksette en eller flere av følgende tiltak**	
a. Setter opp flompreventive tiltak i område A	
b. Setter opp flompreventive tiltak i område B	
c. Setter opp flompreventive tiltak i område C	
3. ***Evakuering:*** **Du bestemmer deg for å iverksette en eller flere av følgende tiltak**	
a. Evakuerer campingområdet og boliger i område A	
b. Evakuerer skoler og sykehus i område B	
4. ***Annet:*** **Du bestemmer deg for å iverksette en eller flere av følgende tiltak**	
a. Avlyser festivalen i område C	
b. Stenger broen i område C	

APPENDIX F

S: Basert på din erfaring fra casen du nettopp gjennomførte: Kan du fortelle meg om hvilke styrker du ser ved et slikt beslutningsstøttesystem

W: Basert på din erfaring fra casen du nettopp gjennomførte: Kan du fortelle meg om hvilke svakheter du ser ved et slikt beslutningsstøttesystem?

... *nå har vi snakket litt om styrker og svakheter ved systemet. Jeg ønsker nå at vi retter blikket litt mer inn i fremtiden og at du tenker litt utover opplevelsen i casen.*

O: Kan du fortelle meg om hvilke muligheter du ser ved et slikt beslutningsstøttesystem i framtiden?

T: Kan du fortelle meg om hvilke trusler du ser ved et slikt beslutningsstøttesystem i framtiden?

Nå er vi interessert i hvilke tanker du hadde om systemet *før* du fikk brukt det. Dette kan både være systemet du fikk teste, og generelt slike systemer.

- 1 Hva tenker du generelt om å ta i bruk teknologi i en slik kontekst?
- 2 Hvilke forventinger hadde du til systemet før du fikk teste det?
    - Hvordan påvirket informasjonen du fikk om systemet i starten hvordan du forholdt deg til systemet?
- 3 Hadde du hørt om slike systemer før i dag?
    - Hva hadde du hørt om slike systemer?
    - Tenker du at det finnes rykter, eller en felles oppfatning som slike systemer?
- 4 Hvor mye erfaring har du med bruk av systemer for beslutningsstøtte?
    - Hvordan påvirket erfaringen/manglende erfaring opplevelsen av systemet?
    - Hvordan påvirket erfaringen/manglende erfaring beslutningsprosessen?
    - Ser du noen utfordringer knyttet til manglende erfaring med bruk av slike systemer?
- 5 I starten så fikk du informasjon om hvordan systemet fungerte. Opplevde du at du fikk en forståelse for hvordan det var bygd opp?
    - Hvordan påvirket din forståelse i forkant, din opplevelse ved bruk systemet?

Nå har vi snakket litt om hvilke tanker du hadde om systemet i forkant. Du fikk også muligheten til å teste systemet, og nå skal vi snakke litt om hvordan du opplevde prestasjonen til systemet.

- 6 Hvor pålitelig opplevde du systemet?
- 7 Opplevde du at systemet korrekte anbefalinger?
    - Kan du si noe mer om hva som gjorde at du opplevde at du fikk korrekte anbefalinger?

    - Hvorfor opplevde du at systemet ga korrektere anbefalinger?
    - Opplevde du at systemet ikke ga korrekte anbefalinger?
- 8 Opptrådte systemet i tråd med det du forventet?
    - Kan du si noe mer om hva som gjorde at du opplevde at systemet opptrådte i tråd med det du forvente?
    - Hvorfor opplevde du at systemet ga korrektere anbefalinger?
    - Ble du overasket over noen av rådene systemet kom med?
- 9 Opplevde du at systemet var nyttig?
    - Kan du si noe mer om hva som gjorde at du opplevde systemet som nyttig?
    - Hvorfor opplevde du at systemet var nyttig?

Helt til slutt

- 10 I hvilken grad vil du si at du hadde tillit til systemet?
- 11 I hvilken grad vil du si at casen var realistisk?

APPENDIX G

1) Kjønn (sett kryss)		
Mann		
Kvinne		
Ønsker ikke å svare		
2) Alder		
3) Etat/arbeidsgiver		
4) Hva er din stillingstittel?		
5) Hvor lenge har du jobbet i nåværende stilling?		
6) Hvor lang erfaring har du med arbeid knyttet til beredskap?		
7) Hva er din utdanning?		
8) Hvor mange værrelaterte hendelser har du vært med på? Dersom du er i tvil er et estimat i orden.		
9) Har du noen erfaring med bruk av systemer for beslutningsstøttesystemer? (Sett kryss)		
Ja, jeg har brukt beslutningsstøttesystemer før		
Jeg har hørt om det, men aldri prøv det		
Nei, jeg har hverken hørt om eller prøvd det		

Jeg vet ikke		
10) Hvis ja, hvilke systemer har du brukt?		

APPENDIX H

Table 3

Overview over deductive analysis results layered on model from K.A Hoff and Bashir (2015)

Layers, factors and sources of variability	**Groups**	
	DEXI	ML
Dispositional trust[a]	1	2
Age	1	1
Culture	1	0
Personality traits	1	1
Situational trust[b]	4	4
External variability	4	4
Type of system	1	2
System complexity	3	1
Task difficulty	0	1
Workload	0	1
Perceived risks	1	0
Organizational setting	0	1
Internal variability	5	3
Subject matter expertise	4	2
Mood	1	0
Attentional capacity	0	1
Learned trust[c]	5	5
Preexisting knowledge	5	5
Attitudes	4	4
Expectations	5	5
Reputation	4	5
Experience	4	5
Understanding of system	4	5
System performance	5	5
Reliability	2	1
Validity	5	5
Predictability	3	4
Dependability	3	4
Usefulness	5	5

Design	4	4
Ease-of-use	3	4
Communication style	1	0
Transparency/feedback	3	0
Level of control	2	0

Note. Sources that were not mentioned are excluded from the table. These are a) gender. b) perceived benefits, framing of task, self-confidence. c) timing of error, difficulty of error, type of error and appearance. Bold =layers, bold italic = factors, italic = sources of variability. Frequencies indicate number of participants that mentioned each theme