Cybersecurity Mindfulness in the Age of Mindless AIs

Investigating AI Assistants Impact in High-Reliability Organizations

HENRIK TENGE HANSEN
TRULS RØSAND VALØ

SUPERVISOR
PAOLO SPAGNOLETTI

University of Agder, 2023
Faculty of Social Science
Department of Information Systems

PREFACE

This Master Thesis marks the end of our degree in Cybersecurity here at the University of Agder. The last two years have been a rewarding experience both in terms of the social- and educational aspects with new challenges overcome and knowledge gained.

The completion of this thesis could not have been done without the help and continued support of our supervisor. We want to thank Associate Professor Paolo Spagnoletti for giving us guidance, constructive feedback and for always being available. We also want to thank all the professors we have had these past two years; your knowledge has been invaluable. Lastly, we want to give a huge thank you to all the informants who took part in this study, without you there would be no completed thesis.

Kristiansand
02.06.2023

Henrik Tenge Hansen	Truls Røsand Valø
Henrik Tenge Hansen	Truls R Valø

ABSTRACT

The Focus: The focus of this Master Thesis is to investigate how AI tools, such as Large Learning Models (LLMs), impact cybersecurity operations in organizations that are regarded as highly reliable. To understand the impacts of AI tools on such operations, we also need to understand the nature of AI tools, their context of use and the experience of users that rely on them.

Research Approach: This thesis is structured around two different methods of investigation. First a systematic literature review was conducted, where related articles was found in different databases, i.e. Google Scholar, Web of Science and the Basket of Eight publications. After this a Qualitative study was conducted where a multiple case study with interviews and random sampling was utilized. A total of 8 informants were interviewed for this study, each lasting ~30 minutes where the questions were based on the findings from the literature.

Findings: From the literature it became clear that AIs, while better than humans in many things such as analyzing Big Data, intrusion detection and other pattern recognition activities, does bring with it many difficulties to the individual and the organization. AIs and LLMs are prone to making you develop an overreliance on them where you accept their answers because of your own biases, while the information itself might be fundamentally wrong or even deceitful. This phenomenon is called AI Hallucination and is vital to understanding an AIs effect on individuals. The literature highlighted that when using any tool, it was important to realize that the AI tool is simply a machine and might be wrong, question everything and do not accept any information at face value. Quite simply, think things through.

LLMs have a problem with transparency, it is impossible to know its ‘reasoning’ behind the information it provides. This fact is supported by both the literature and the interviews themselves. Overreliance, hallucination, cultivating the wrong kind of trust and lack of transparency all lead to an individual acting mindless who takes the information as true. While they have been deceived by trusting something that essentially is untrustworthy or at the very least should have been looked more into.

Implication: The practical implications for this study is that an organization, especially if it is of high reliability should carefully identify measures to avoid the negative impact of AI Assistants when used in day-to-day work in cybersecurity operations.

TABLE OF CONTENTS

List of Figures

List of Tables

1 INTRODUCTION

Artificial Intelligence (AI) and Large Learning Models (LLMs) are topics that have risen in popularity and relevance the last couple of years, especially with the rise of ChatGPT, Alexa, Google Nest, Siri etc. These tools are incredibly easy to use, you just ask them a simple question and they respond with the information you, hopefully, are looking for. However, that is a key distinction, hopefully, because can you really trust the information you are fed?

In early 2023 a young student here in Norway used ChatGPT in order to help with an assignment regarding Norwegian heroes, so she asked ChatGPT about it and one of the examples that it gave back to her were Anders Behring Breivik, the terrorist who killed 77 people in 2011 (Falk, 2023). Now, of course every Scandinavian would instantly recognize this as preposterously false, but it makes you wonder, if it is so wrong about this simple thing then what else might it be wrong about as well? This is a concept called AI hallucination. Hallucination can be described simply as the AI providing false information by contradicting itself or it makes things up with false references or statements that cannot be verified (Athaluri et al., 2023, p.1; Curtis, 2023, p.275; Alkaissi & McFarlane, 2023, p.2). Hallucination is further explained in Chapter 1.2.

This points to the AI not being particularly trustworthy as it is prone to misleading you. Would using such technology both privately or within a professional organization cause several problems related to individual problem solving and decision making? How can employees within that organization be aware and verify that the AI’s answers or recommendations are correct at any time during the usage of that tool? This leads us to the concept of individual mindfulness.

Mindfulness is the concept of an individual being aware of, and thinking through, what exactly they are doing. An individual who has a high level of mindfulness would then question events in order to create their own interpretations, which in turn would only heighten their mindfulness (Langer & Imber, 1980; Langer & Moldoveanu, 2000, referenced in Dernbecher & Beck, 2017, p.122). Mindlessness however is the polar opposite. Mindlessness is about running on ‘autopilot’ and exerting no critical thinking. Examples of mindless actions and behavior are breathing, blinking and simply walking around. There is no extra brain power invested into breathing in, blinking your eyes or moving one leg in front of the other but you can take conscious control of these actions, thus making them more mindful.

The level of mindfulness regarding cybersecurity is seen through the lens of security awareness, which in the context of this thesis is an important element of mindfulness. The definition of security awareness is that “*information security awareness deals with the use of security awareness programs to create and maintain security-positive behavior as a critical element in an effective information security environment*.” (Kruger & Kearney, 2006, p. 289). Security Awareness is also used to “*refer to a state where users in an organization are aware of - ideally committed to - their security mission (often expressed in end-user security guidelines)*” (Mathieson, 1991, referenced in Siponen, 2000, p. 31). This would mean that in a mindful scenario the employees are firmly aware of the guidelines and the threats that AIs and LLMs pose to their organization, they would not automatically believe that any information they receive from the AI or LLM is correct without verifying it first. On the other hand, a mindless scenario would imply that the employees are just assuming that the AI is correct without thinking through the potential dangers of trusting and relying on it for important tasks.

Of course, AIs are not just simply chatbots that you can interact with. AIs have been used for many years already in many organizations, for instance in the analysis of Big Data, i.e., data of such quantity that it would be impossible for humans to properly analyze it within a satisfactory timeframe. It is also used in other pattern recognitions through Machine Learning (ML), such as during an attack on an organization's systems. However, how will the use of AIs affect an individual's mindfulness?

For this study we want to see what kind of impact AIs have on an individual's mindfulness. In order to best do that we have interviewed key personnel in critical infrastructure and other High-Reliability Organizations (HROs). Why? It is based on an assumption we have made, that in those organizations, because they are important, the level of mindfulness is already quite high, so we get a picture of how AIs affect them.

It is important to have a clear definition of what constitutes critical infrastructure, Figure 1 illustrates what we have based our choice of informants on, and Chapter 1.1 quickly explains every sector.

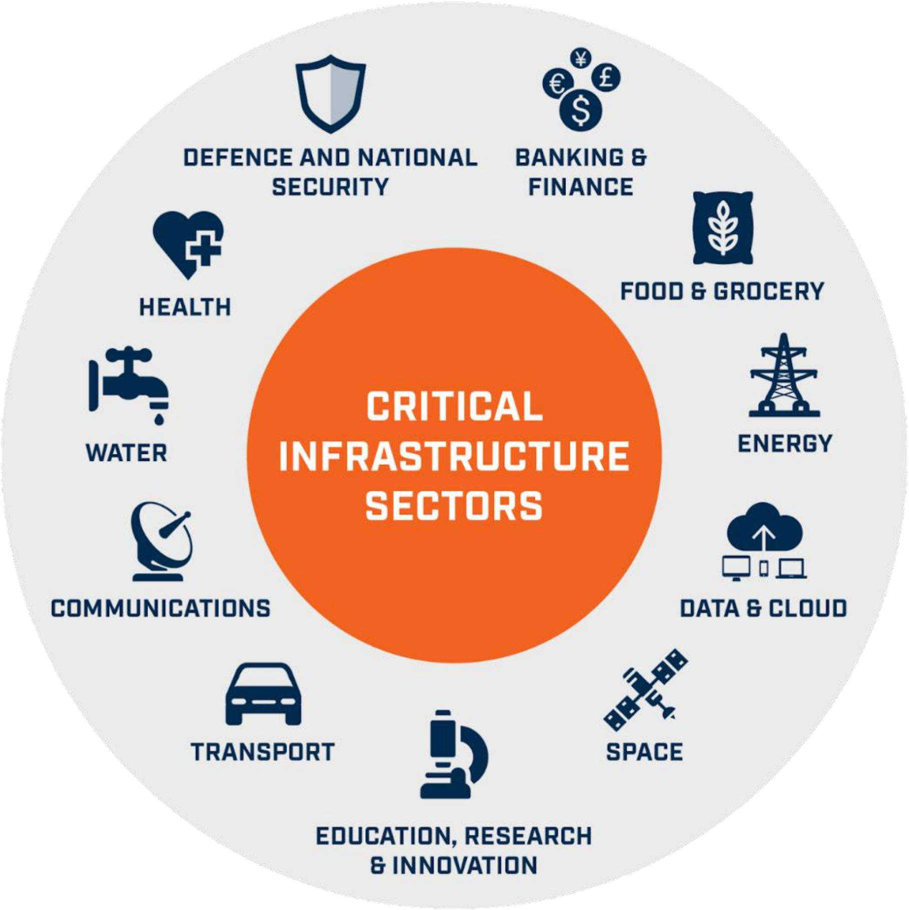

CRITICAL INFRASTRUCTURE SECTORS
DEFENCE AND NATIONAL SECURITY
BANKING & FINANCE
FOOD & GROCERY
ENERGY
DATA & CLOUD
SPACE
EDUCATION, RESEARCH & INNOVATION
TRANSPORT
COMMUNICATIONS
WATER
HEALTH

Figure 1 Critical Infrastructure Sectors[1]

1.1 Critical Infrastructure Sectors

Figure 1 presents the different sectors that are critical for society, however, they are perhaps a bit too abstract so there will be a short explanation of them here.

Banking & Finance. This sector revolves around the continued circulation of money. For instance, if DNB, the largest bank in Norway, were to suffer a catastrophic failure, millions would not be able to access or move their money around. The sector also considers the stock market, ensuring that stock prices are not unnaturally tampered with.

Food & Grocery. This is not just your local grocery store keeping open, it is everything from the growth of vegetables and produce, the creation and use of artificial fertilizer to the packaging of raw materials and its transport to the shelves in your local store.

[1] Figure 1; Critical Infrastructure Sectors” [Figure]. u.y. By Huntsman Security. Retrieved from https://www.huntsmansecurity.com/industries/critical-infrastructure/

Energy. Everything from, hydro-electric dams, generators along waterfalls, nuclear power plants (in other countries) etc. to its transport in power lines and transformers all the way to the output you are charging your phone in. All of this is encompassed in the energy sector.

Data & Cloud. In today's highly digitalized society where virtually everything is stored in some form of cloud it is vital that these are kept up along with the general access to the internet. If access to these systems were lost, then that would spell problems. For example, doctors are not able to access your medical files to see what kind of medicine they should give you.

Space. Closely linked with Communication, which revolves around the satellites and their receivers/transmitters that allow for accurate GPS and communication.

Education, Research & Innovation. The education of the next generation is critical for any society along with research into new fields in order to innovate and find new solutions to old and new problems. For example, cancer research at hospitals.

Transport. This sector is not about your Toyota, it is about the roads, tunnels, bridges, trams, airports and railway systems that connect every part of Norway to each other. The continued maintenance and planning for new transport systems are also a part of it.

Communication. Closely linked with Space, however this sector is more along the lines of phone masts that allow people to call each other, and internet masts that give the population access to the internet.

Water. The collection and purification of water, making it drinkable for the population. This is along with the waste disposal and major pipelines that distribute this to every home here in Norway.

Health. Hospitals and homes for the elderly, along with all the infrastructure in place to take care of people with disabilities that need help. In addition, the distribution of medicine that the public needs. Included are also all the operative systems that connect these with each other.

Defense & National Security. The military infrastructure, fighter jets, bombers, tanks, warning systems, radar systems and the production and development of weapon systems. This sector also encompasses the civilian agencies and departments that plan and protect the population from natural disasters and other events that are crucial for national security.

Now that the different sectors have been given a short description, there needs to also be a definitive description of what constitutes as a LLM and how it differs from previous AI assistants.

1.2 AI & LLM

AIs are not something new. Apple’s built in AI assistant Siri had its initial release in 2011, Amazon Alexa came out in 2014, Google Nest arrived in 2016 and basic chatbots on various websites have been available for a long time. So why exactly have the popularity and interest in AIs blown up in the last year? The answer is both simple and complex, LLMs.

LLMs differ from the ordinary AI assistants you find in your home or on your phone. Siri, while good at telling the time or cracking the odd joke, is not comparable at all to the capabilities present in for example ChatGPT which already in January 2023, 2 months after its initial release, had over 100 million active users in that month alone (Hu, 2023).

Why is Siri not comparable? The first hint lies in the name GPT, it is an abbreviation of Generative Pre-trained Transformer. LLMs like ChatGPT are trained on vast amounts of data which allows it to generate human-like responses and function with high accuracy (Kasneci et al., 2023, p.1). LLMs are simply far more comprehensive tools than the basic assistants in that the responses you as the user receive are more ‘thought out’ and seem better. We asked ChatGPT ‘what is the difference between Siri and a large language model?’ it responded with a 355 word long, detailed, response outlying all the differences and similarities. Of course, getting responses like that would most likely resonate with more users than a basic short answer would.

The increase in popularity of AIs have also led to an arms race between the big tech companies i.e., Microsoft, Google etc. where all of them are attempting to get a majority share in the market by having the leading AI Assistant. At the AI Forward 2023 event in May 2023, Bill Gates explained that the goal was to create an AI that would inevitably do anything for you so that you never again had to use search engines or go onto Amazon in order to buy products (Vanian, 2023). AIs and LLMs are thus here to stay, and with more and more money pumped into creating satisfying models, along with the millions of users helping train them, the quality will only increase in the years to come.

1.3 Motivation

In the beginning of this chapter, hallucination was briefly introduced where ChatGPT identified Breivik as a Norwegian hero. This opens up a whole lot of possibilities for mis- and disinformation to be spread with the use of LLMs.

The concept of hallucination is important in order to get any clear idea on how AIs, especially LLMs affect mindful individuals, but what exactly is it? AI hallucination is defined as "*a phenomenon where AI generates a convincing but completely made-up answer*" (Athaluri et al., 2023, p.1) where it can use fake references in order to be more compelling (Curtis, 2023, p.275; Alkaissi & McFarlane, 2023, p.2). Not just that but hallucination can be split into two different parts, intrinsic- and extrinsic hallucinations. Intrinsic hallucinations "*refers to the LLM generation that contradicts the source/input*" (Bang et al., 2023, p.17). Whereas extrinsic hallucinations "*refers to the LLM generations that cannot be verified from the source/input content (i.e., output that can neither be supported nor contradicted by the source)*" (Bang et al., 2023, p.19).

The problem with this is multifold as "*it hampers a user's trust in the AI system, negatively impacts decision-making, and may give rise to several ethical and legal problems*" (Athaluri et al., 2023, p.4). The hallucination problem is not just present in ChatGPT, it is present in all types of LLMs as they "*hallucinates with human-like fluency and eloquence on things that are not based on truth; and as a general-purpose language model trained from everything on the web, its language coverage is questionable*" (Bang et al., 2023, p.1). Having the AI provide false information would be catastrophic in many sectors, for instance in the Defence and National Security sector AIs will be used for "*intelligence, surveillance, reconnaissance, logistics, cyberspace operations, information operations, command and control, semiautonomous and autonomous vehicles and autonomous weapon systems*" (Hartman & Steup, 2020, p.328). It is therefore obvious that a fault in the AI could result in huge damages and threats. Of course, this may be an extreme scenario, but the point is still true whether AI is used in military, financial or just as an assistant in a job outside of the critical infrastructure. One mistake and your entire system may be brought down, or confidential information may be taken.

ChatGPT as a LLM seems to have especially many problems as it "*shows more weakness in inductive reasoning than in deductive or abductive reasoning. ChatGPT also lacks spatial reasoning while showing better temporal reasoning. ChatGPT also lacks mathematical reasoning*" (Bang et al., 2023, p.2). It is not just that however, ChatGPT also suffers from "*hallucination problems like other LLMs and it generates more extrinsic hallucinations from its parametric memory as it does not have access to an external knowledge base.*" (Bang et al., 2023, p.1). It has also been shown to have an accuracy of 63,41% on 10 different reasoning categories (Bang et al., 2023, p.1).

In December 2022, the CEO of OpenAI, the creators of ChatGPT, Sam Altman wrote on Twitter that "*It's a mistake to be relying on [ChatGPT] for anything important right now*" (Altman, 2022, referenced in Bang et al., 2023, p.1). Trusting LLMs too much right now has the potential of creating false experts, thus it can be harmful if those 'experts' are in key positions such as the medical field (Alkaissi & McFarlane, 2023, p.4).

Now that it has been established that LLMs can essentially 'lie' to the users, this motivates the topic for this thesis. When tools like this get increasingly popular, they will most likely be used by employees in critical infrastructure eventually, if they are not already in use, sanctioned or not. Therefore, it is of importance to investigate how it affects cybersecurity mindfulness in HROs.

1.4 Research Question

We are interested to see what effect the usage of AIs has on the mindfulness of cybersecurity operations; therefore, we will find informants from critical infrastructure to question within the context of cybersecurity mindfulness.

Thus the research question (RQ) guiding this thesis is the following; *How will AI Assistants Affect Cybersecurity Mindfulness in High-Reliability Organizations?*

1.5 Scope of the Thesis

The scope is important for any thesis. We are 2 students who have to complete a study and write a report on it within a certain timeframe and thus have limited resources. If we have too large of a scope then we will not be able to complete it. If we have too small of a scope then the thesis will neither be innovative nor compelling.

Based on this the scope of this thesis is to research how AI's affect an individual's mindfulness in the context of cybersecurity, first a literature review has been conducted in order to find relevant research on the topic or themes associated with it i.e. Mindfulness, mindlessness, AI, ML, trust etc.. After this we conducted interviews with a select group of individuals that matched our criteria, this is further explained in Chapter 3. Finally we analyzed the results and connected the interviews to the literature review findings.

1.6 Structure of the Thesis

This section contains the structure of out thesis.

Chapter 2 Literature Review. In this chapter the list of the literature we have found and the different findings that have been made are put into fitting headings in order to best investigate what the consensus is regarding our topics.

Chapter 3 Research Approach. This chapter establishes what kind of literature review and research approach we have taken to study the phenomenon we have chosen, whether that be qualitative or quantitative. The chapter also plans and explains in detail how exactly we went forth with collecting the information that we seek. In addition, it also explains how the information was analyzed and what the ethical considerations of such a study is.

Chapter 4 Findings. After the information is collected then it needs to be analyzed in order to find what the informants agreed or disagreed on. The chapter is split into different subchapters based on the NVivo codes from the previous chapter.

Chapter 5 Discussion. In this chapter the information is discussed in relation to the findings from the literature review to see if the informants agreed or disagreed with the literature. Or if the informants had information that was not brought up in the articles.

Chapter 6 Conclusion. The last chapter of the study is the conclusion and culmination of the entire thesis, where not only the conclusion is presented but also the practical implications. There is also a critical look at our work this semester where we analyze what could have been done better or what we would have done if we had more time and resources.

2 LITERATURE REVIEW

In order to find out what the literature says regarding our selected topic, there needs to be a literature review conducted in order to get a good overview of what the current consensus is. This Chapter will sum up our literature findings that are collected by conducting our literature methodology described in Chapter 3. The method includes inclusion and exclusion criteria, primary and secondary keywords and on what platforms the literature was found during this review.

This Chapter will both list and categorize the different literature we found relevant during this review, and explore the findings related to it.

The categories that are identified in this Chapter will be added to our variation of the Waardenburg & Huysmans model on AI implementation (Waardenburg & Huysman, 2022, p.7). Our model can be found in Chapter 2.3 and functions as the culmination of the literature review.

2.1 Literature List

Finding the literature involves the use of the criteria mentioned in Chapter 3.2. The articles we include within this review should contain at least one of the following criteria. The first iteration of this process would mainly include the title of the articles within our search results. The articles would then be reviewed by reading the abstract, introduction and conclusion. This process is called the filtration or screening process, where we filter every article we find relevant to narrow our list of articles found (Xiao & Watson, 2019).

Table 1 lists up every article that is being included within this systematic literature analysis. Each article that either relates to mindfulness/mindlessness, AI or cybersecurity, with a complete study that shows interesting findings fully/partially related towards our research scope would be included for this review and are listed under the **main topic**. The research articles could also include other topics listed within the **secondary topics**, which we would qualify as relevant if they also are connected to either mindfulness/mindlessness, AI or cybersecurity. The articles listed contribute to our research, which would establish a better foundation for this thesis.

Table 1 Literature List

Authors	Main Topic			Secondary Topics			
	Mindfulness/ Mindlessness	AI	Cybersecurity	Awareness	Trust	Organizational	ML
Dernbecher, et al. (2017)	X						
Thatcher, et al. (2018)	X						
Zhu, et al. (2015)	X		X				
Esfahani, et al. (2020)	X						
Araujo. (2018)	X	X					
Li. (2018)		X	X				
Morales-Forero, et al. (2022)		X	X			X	
Thuraisingham. (2020)		X	X				
Zhang, et al. (2022)		X	X				
Hartmann, et al. (2020)		X	X				X
Trim, et al. (2022)		X	X			X	
Wazid, et al. (2022)			X				X
Canbek, et al (2016)		X					
Benzaïd, et al. (2020)		X	X				X

Authors	Main Topic			Secondary Topics			
	Mindfulness/ Mindlessness	AI	Cybersecurity	Awareness	Trust	Organizational	ML
Mirsky, et al. (2023)		X	X			X	
Cunneen, et al. (2020)		X					
Gratch, et al (2022)		X					
Jensen, et al. (2017)	X		X				
Burns. (2019)	X		X				
Ansari. (2022)	X	X	X				
Timmers. (2019)		X	X		X		
Passi, et al. (2022)		X					
Sayan, et al (2017)		X	X	X			
Siau et al. (2018)		X	X		X		X
Puthal et al. (2021)		X	X		X		X
Buchanan (2020)		X	X			X	

The next section would discuss the different findings extracted from these articles, each in their own category.

2.2 Literature Findings

After finding all the relevant articles for the thesis it is important to find out what they as a collective say the current understanding of the topics are. From the literature there has been created two different categories from which this thesis will base itself around. These categories are; Mindfulness & Mindlessness and AI Benefits & Disadvantages. They will be discussed in the coming sub-chapters.

2.2.1 Mindfulness & Mindlessness

The concept of mindfulness and mindlessness is one that is foundational to our entire thesis, thus in order to investigate what effect an AI has on mindfulness we first have to identify what defines a mindful or mindless process.

Mindfulness is described as a process where the individual has a certain awareness of their surroundings and their consequences. An individual who has a high level of mindfulness would then question events in order to create their own interpretations, which in turn would only heighten their mindfulness (Langer & Imber, 1980; Langer & Moldoveanu, 2000, referenced in Dernbecher & Beck, 2017, p.122). This is further expanded upon when also adding that they can identify "*novel aspects of context that can improve foresight and functioning"* (Langer, 1989, referenced in Thatcher, Wright, Sun, Zagenczyk & Klein, 2018, p. 832). This can be boiled down to 5 simple aspects: "*1) a preoccupation with failure, (2) a reluctance to simplify interpretations, (3) a sensitivity to operations, (4) a commitment to resilience, and (5) a deference to expertise*" (Burns, 2019, p.14). One of the ways in order to get mindful individuals is to conduct mindfulness training.

Conducting mindfulness training on individuals can also make a business much more secure from a cybersecurity standpoint, especially against phishing attacks because the training makes them stop and think about certain actions before they do them (Jensen, Dinger, Wright & Thatcher, 2017, p.602). If the individuals are more resistant to phishing attacks, then the arguably largest attack surface of the organization, namely the employees, have shrunk and thus making them more cyber secure.

On the other hand, mindlessness is the antithesis of mindfulness. Where mindfulness is cognitive thoughts and reactions that an individual consciously makes, mindlessness is the unconscious or rather 'automatic' actions an individual makes (Zhu, Carpenter & Kolimi, 2015, p. 1067). Examples of unconscious actions can be something as simple as breathing or blinking, you breathe and blink automatically you don't need to invest brainpower for that function. People also tend to mindlessly apply social rules and expectations to computers and if the individuals are in a good mood, they have a high probability of performing a task mindlessly (Zhu, et al., 2015, p. 1068). It also appears that an individual acting mindlessly is far more likely to divulge sensitive personal information based on unsound reasonings given to them (Zhu, et al., 2015, p. 1072).

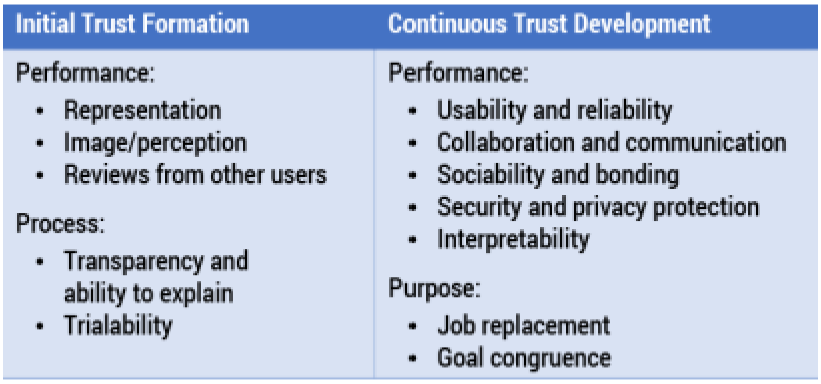
An additional factor is that if an AI seems anthropomorphic. Anthropomorphism is the concept of how human something appears. For instance, a chatbot with rudimentary answers would perhaps not seem that human, whereas an AI assistant like Siri who has a human voice may seem much more human. The more human a machine appears the more it will increase the users trust and lower their uncertainty (Hoeffler, 2003; Castano and Giner-Sorolla, 2006, referenced in Esfahani, Reynolds & Asleigh, 2020, p.1). This is also corroborated upon because if you remove the human cues from the AI, people are far more likely to understand on an instinctive level that it is just a machine and those human cues help build a 'relationship' between the human and the machine (Araujo, 2018, p.188).

This would indicate that the individuals may act more mindless around AIs that appear human, especially if we combine this information with Zhu et al. mentioned earlier (Zhu, et al., 2015, p. 1068).

2.2.2 AI Strengths & Weaknesses

One of the biggest risks regarding AIs is the development of an overreliance on them. The overreliance is defined as "*users accepting incorrect AI recommendations—i.e., making errors of commission. Overreliance generally happens when users are unable to determine whether or how much they should trust the AI*." (Passi & Vorvoreanu, 2022, p. 2). The level of trust between the human and the machine is based on the so-called AI literacy, Expertise and Task Familiarity that the individual has (Passi & Vorvoreanu, 2022, p. 4-5). Similarly there are certain biases that are inherent in the human mind, such as automation- and confirmation bias. Automation bias is self explanatory, it revolves around an individual who would rather favor the recommendations from automated services than non-automated services. If you have a high automation bias you will be unable to regulate your reliance even if the AI makes mistakes, which it inevitably will do (Passi & Vorvoreanu, 2022, p. 6). The other bias, confirmation bias, is simply that you want answers that confirm your assumptions. Individuals with this bias wrongly attribute the AI with logic and reasoning (mindfulness) when in reality it has none of those (mindlessness) (Passi & Vorvoreanu, 2022, p. 7). The overall level of performance is also tanked because of the overreliance compared to when the human and the AI work separately (Passi & Vorvoreanu, 2022, p. 10).

Overreliance is simply when you trust the AI too much, but establishing trust is important. This is because trust is the primary reason for acceptance of any type of information (Siau & Wang, 2018, p.47). However, it is not enough to simply establish trust, it is important to cultivate and maintain it as Figure 2 shows how to do.



Initial Trust Formation	Continuous Trust Development
Performance: • Representation • Image/perception • Reviews from other users Process: • Transparency and ability to explain • Trialability	Performance: • Usability and reliability • Collaboration and communication • Sociability and bonding • Security and privacy protection • Interpretability Purpose: • Job replacement • Goal congruence

Figure 2 Technology features of AI that affect trust building[2]

What becomes the issue then is how to have trust in the AI system but not so much that it becomes an overreliance on it.

However, overreliance is not the only major threat that is associated with AIs, another is the fact that an AI may be attacked or deceived and is susceptible to poisoning attacks. This would lead it to giving you the wrong answers and predictions (Li, 2018, p.1463; Morales-Forero, Bassetto & Coatanea, 2022, p.6; Mirsky et al., 2022, p.1; Thuraisingham, 2020, p.1116; Benzaid & Taleb, 2020, p.143). Data poisoning is a concept that can threaten an AI when you train it (Puthal & Mohanty, 2021, p.33). If an adversary could change the training data the AI could act more adversarial rather than a helpful tool as it should. Data poisoning needs more research, especially in national security organizations because that is usually where the dedicated adversaries are and ready to exploit any flaw or mistake they can (Buchanan, 2020, p.10). Considering this thesis focuses on critical infrastructure, which in most cases can be described as national security, this is something that is important.

This would also be especially prevalent considering that Deep Learning AIs are not transparent in how they reach certain answers or conclusions (Timmers, 2019, p.637). This may of course be circled back to the biases that were brought up in the previous paragraph where it also may provide the users with a sense of naivety

[2] Figure 2 "Technology features of AI that affect trust building" [Figure]. 2018. By Siau & Wang. Retrieved from https://www.researchgate.net/profile/Keng-Siau-2/publication/324006061_Building_Trust_in_Artificial_Intelligence_Machine_Learning_and_Robotics/links/5ab8744baca2722b97cf9d33/Building-Trust-in-Artificial-Intelligence-Machine-Learning-and-Robotics.pdf

and false sense of security while interacting with the AI and thus affect their security awareness (Timmers, 2019, p.637; Ansari, 2022, p.1).

What this then tells us is that an AI is, as all technology, hackable. If the AI then is hacked and gives out the wrong information there is the possibility that the users may take it for granted that it is correct, as we discussed in the previous paragraph.

AI Assistants (AIA), have the possibility of making the life easier for the users by letting the user see what is most relevant to them based on their search pattern. However " *the framing and agreement to tailoring obfuscates other possible data use contexts that may lack user benefits and may present risks by permitting commercial analytics and insights regarding user behavior*." (Cunneen, Mullins & Murphy, 2020, p.625). This obfuscation would then obviously lead to the user not getting the best information they would need to fulfill their tasks. In addition to this some researchers argue that the use and personalization, i.e the AI is affected by your search patterns, makes people tend to act more deceptive than if they had not used an AIA (Gratch & Fast, 2022, p.1-2). It appears that this does not explicitly lower the level of mindfulness, it however makes the individuals much more prone to knowingly bend the rules because it is not them directly that is bending them, but rather the AIA (Gratch & Fast, 2022, p.2). This fact can of course negatively impact the employees' security behaviour and make the organizations more at risk.

Of course, there are not just risks connected with the use of AIs, there are also great possibilities that can be achieved. An AI within a system has the possibility of performing User Access Management, Network Situation Awareness, Monitoring of Dangerous Behaviour and Identification of Abnormal Traffic (Zhang, 2022, p.1031-1037; Wazid, Das, Chamola & Park, 2022, p.314; Sayan, Hariri & Ball, 2017, p.313-314). This relieves a huge workload of an individual where the AI is much more suited for this kind of work and the individuals can focus their time on responding to events in real time (Trim & Lee, 2022, p.10). AIs have also been shown that they may be a good defense for 5G networks and beyond (Benzaid & Taleb, 2020, p.147)

AIs are not simply advantageous on the overall business side as you can see above, having AI assistants, or so called Intelligent Personal Assistants (IPA) has been shown to increase the users knowledge in the specific field in which it assists and could possibly make even the least expert individual safe on the web (Canbek & Mutlu, 2016, p.596; Sayan, Hariri & Ball, 2017, p.314).

2.3 Summary

To summarize our findings it can be said that you want to have a high level of mindfulness. Mindfulness is simply the ability to have critical and rational thinking while also daring to ask questions. Whether it is just an individual citizen or your employee, the benefits are quite clear in that regard compared to the risks that are appearing in this cyber age regarding mindlessness.

AIs on the other hand are, in a simplified way, just lines of unfeeling code. An AI does not have the ability to be mindful as its answers come from previous search patterns. The problem then becomes if a mindful individual can become compromised by an overreliance on AI assistants in their daily routine? While the research is clear that an AIA certainly has both positives and negatives regarding what it can do, the fact that there seems to be a possibility of it being poisoned or targeted to behave more adversarial is a crucial negative. The fact that it might contradict itself thus giving an unstable basis of knowledge at best is concerning, add in the fact that it can completely make up false references, as seen in Chapter 1.3, to back up its claims then it seems completely factual, whereas in reality it is not. It is this that this thesis will look deeper into.

Figure 3 is based on the Waardenburg & Huysmans model on AI implementation (Waardenburg & Huysman, 2022, p.7). This model is a representation of our Literature review where 4 topics have been identified when Users work together with AI Assistants. The figure illustrates the interaction between these two entities which forms the area of interaction within cybersecurity operations.

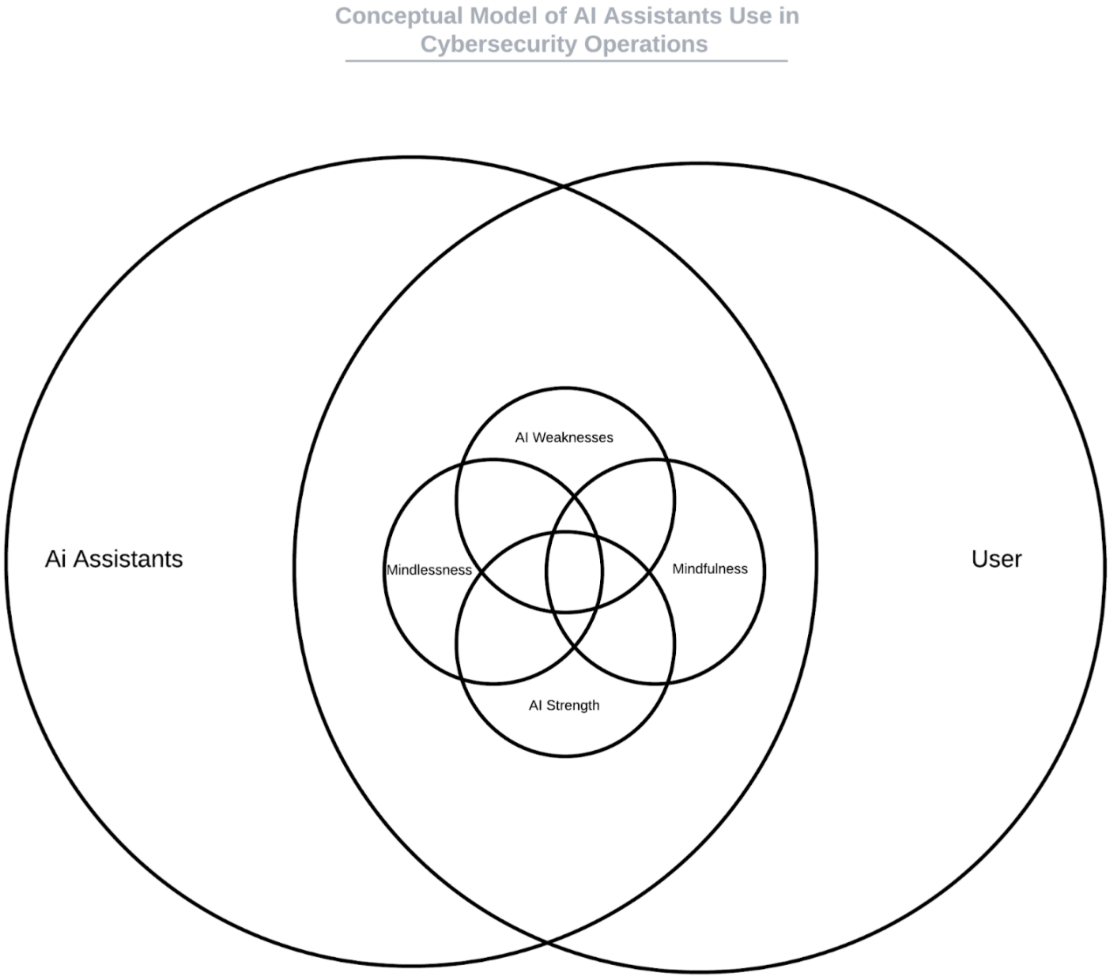

Conceptual Model of AI Assistants Use in Cybersecurity Operations
Ai Assistants
User
AI Weaknesses
Mindlessness
Mindfulness
AI Strength

Figure 3 Conceptual Model of AI Assistants Use in Cybersecurity Operations

2.3.1 *The Gap*

The Gap in the research is that our RQ in itself has not been investigated thoroughly before, mainly because LLMs is new and the research has not come out yet. In Figure 3, and in the previous chapters, it has become apparent that AIs have both strengths and weaknesses. Depending on how you see it a strength can also be a weakness in a different context. For instance, automation; you streamline the processes and make the organization more efficient, however you remove the human, thus there is perhaps less oversight into the decisions the AI makes and the employees utilizing the AI become over reliant on it. There is a fine margin between what is a strength and a weakness. In conjunction with this, the mindfulness and mindlessness aspects also overlap, as seen in Figure 3. An AI strength or weakness can lead to both mindlessness and mindfulness based on how the tool is utilized. In Chapter 1.3 the concept of AI hallucination was brought up, this concept can easily lead to mindlessness by tricking the user. However, if the user is aware of the phenomenon then they can counter it by reference checking the information and thus acting more mindful.

The literature however, has not gone in depth to understand what kind of an impact an AI has on individuals already considered mindful. While some articles have brought up concepts like overreliance and shown that it might lead to mindless actions from the individual, it has not been set in the context of an HRO. An HRO has a high level of mindfulness and they think things through, to avoid being deceived as best as they can. Because of this there is a gap in the knowledge, and is why it would be prudent to investigate this phenomenon in HROs, to see just what kind of effect AIs and LLMs have on an individuals mindfulness.

This has led us to formulating this RQ for the thesis: *How will AI Assistants Affect Cybersecurity Mindfulness in High-Reliability Organizations?*

3 RESEARCH APPROACH

In this chapter, the different methodologies utilized in this thesis will be explained. This thesis is built up by first conducting a literature review, then a qualitative study in order to thoroughly investigate the RQ.

The literature review is foundational for the entire thesis, you have to start with that before you conduct any type of qualitative or quantitative study. Why? Because it is through the literature review that it is possible to see what is already understood and what the gap in the literature is. This is important as the gap guides the questions in the study that is conducted after the review.

3.1 Literature Methodology

This section covers the process of conducting a systematic literature analysis. Yu Xiao and Maria Watson's report on “*Guidance on Conducting a Systematic Literature Review*” (Xiao & Watson, 2019) describes this process in great detail. They visualize these guidelines in their report, step by step. We have recreated their model in our literature review, Figure 4 is based on the same steps and fulfills the same purpose.

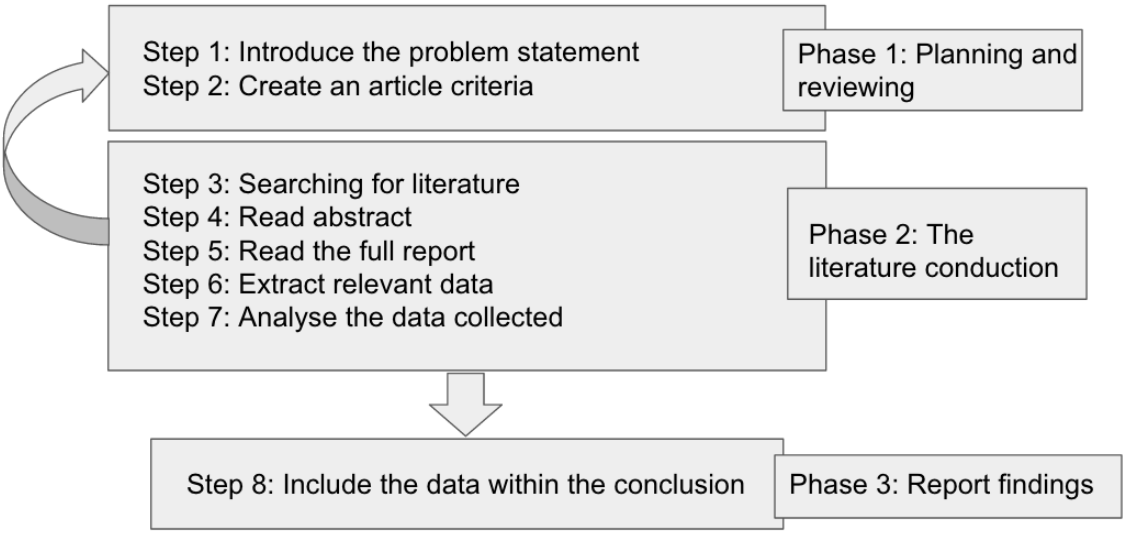

Step 1: Introduce the problem statement
Step 2: Create an article criteria
Phase 1: Planning and reviewing
Step 3: Searching for literature
Step 4: Read abstract
Step 5: Read the full report
Step 6: Extract relevant data
Step 7: Analyse the data collected
Phase 2: The literature conduction
Step 8: Include the data within the conclusion
Phase 3: Report findings

Figure 4 Process of Systematic Literature Review

Figure 4 is divided in 3 phases, where each phase includes one or more steps that are related to this methodology. The figure represents an iterative process which is represented by the arrow on the left side. This process is done to continuously clarify the research question after completing an iteration, each iteration increases the knowledge of the current problem statement, which results in better and more precise literature conduction for each iteration that is done in this analysis.

3.2 Criteria

The criteria is constructed as a blueprint for filtering articles that are deemed relevant/irrelevant. It is an important protocol of this review, as Xiao & Watson writes that "*It is necessary for enhancing the quality of review because it reduces the possibility of researcher bias in data selection and analysis*" (Kitchenham and Charters 2007, referenced in Xiao & Watson, 2019, p. 103). The purpose of each criteria is to collect valuable data from each article that concerns AI, mindfulness/mindlessness or cybersecurity. The articles should be written in an understandable language and be relatable towards the scope of this thesis. We have constructed an inclusion criteria table, which lists the necessary information an article should include, and an exclusion criteria table, which exclude any articles that fall into the different categories listed.

Table 2 Inclusion Criteria

Criteria	Justification
The study must focus on Mindfulness/Mindlessness and AI risks or AI.	Our study investigates the effect mindless AIs have on an individual's mindfulness. It is therefore important to collect articles that revolve around this issue. We need good descriptors of what defines mindfulness and mindlessness. In addition we need information on the possible risks that can occur when using AIs in order to see if they could negatively affect an individual.
The study should be related to cybersecurity.	Our thesis is based on cybersecurity related content. Articles that cover

	aspects that are not directly concentrated towards cybersecurity topics, but rather could be related towards our scope, would be included within this analysis.
The study must be written either in Norwegian or in English.	We would not be able to understand them if they were in another language.

Table 3 Exclusion Criteria

Criteria	Justification
The study excludes topics that can not be related to cybersecurity.	We are writing a thesis that includes cybersecurity related content. Any research articles that are completely irrelevant towards this topic would be excluded from this review.
The study is not complete.	Articles with partially or incomplete studies would be disregarded. As studies with a complete conclusion would be considered a more valuable contribution to our studies.
The study is not related to our thesis scope.	Studies that cover different aspects of the same topics we find irrelevant.

3.3 Finding Literature & Relevant Topics

The process of conducting a literature review is done by following the different processes in Figure 4 in an orderly manner. There are several strategies to search for literature, the use of electronic databases, forwards and backwards searching (Xiao & Watson, 2019, P. 103). In this review, we will firstly need to narrow down which databases we conduct our searches in. For this review we have landed on Web of Science and Google Scholar as the primary sources. When using Google Scholar it is easy to end up with millions of results if you do not specify your searches, therefore we tried to be as specific in our searches as possible, and only looked at the first 3 pages of results. This was because we can not go through all articles, and based it on that if it was not in the first 3 pages then it was not really relevant or any good. We also chose the Basket of Eight as one of the most reliable database sources, as these articles often reflect high value to the research community.

The next step in this review is to search within these databases with certain types of keywords that are related to the topic of this report. This is done to “*dissect the research question into concept domains*” (Kitchenham and Charters, 2007, referenced in Xiao & Watson, 2019, p. 103), which would narrow the search results within relevant categories for this review. In order to find these keywords the research question has to be looked into and broken down into distinctive elements.

3.3.1 Theoretical Background

Our research question is specifically targeting high reliability organizations based on their professionality and responsibility towards the society. The research question also asks how the impact of mindfulness can be affected by using mindless AI assistants, which revolves around aspects of cybersecurity, security awareness and trust towards these tools. This creates a foundation of topics that we would further implement throughout our study. The research question is directed towards **cybersecurity** related topics, where **mindfulness** can be affected by **mindless** tools like **AI** assistants. Modern AI tools we see today are both new and untested which can cause skepticism. We believe that **trust** is an important factor when it comes to using it with care, and would implement that as a topic for this analysis. Another important aspect of AI is **machine learning**. Many organizations already use ML in their daily security operations, which is also why we would like to include this topic for our literature analysis as it represents similar technology, but at a much lower level.

Mindfulness and mindlessness within the context of cybersecurity is one of the main elements that is involved within the RQ of this thesis. These elements combined within the context of using AI assistants creates the foundation of our research. These elements are described in detail in Chapter 1 and are one of our main focuses that are included within our literature analysis, which in this review would also be considered as our main keywords. These keywords alone would result in vast numbers of articles that might not be related towards our research topic. We also included other keywords that can be partially connected towards the topic of this thesis. We have implemented security awareness as one of these secondary keywords as mindfulness and mindlessness is closely connected to the individual's awareness as stated in Chapter 1. We are also including several keywords that are connected towards AI implementation and usage within HRO´s. This involves elements like trust towards AI and organizational handling of AI. We are also including machine learning as this type of technology resembles the one used in AI. This thesis mainly concerns topics that are newly introduced to the research community that explores a new type of technology, which is currently in the process of being used in our everyday lives. The consequences that follow is the lack of information about literature online which relates to our topic, and is the main reason why keywords that can be partially related to the topic are included within our literature review strategy. The keywords combined would also result in narrower and more specific search results, where articles that are more relevant can easier be found.

3.4 Research Design

The gap explained in Chapter 2.3.1, focuses on the lack of information about our research topic. This section would propose a suitable research strategy to our research question, where each phase includes different design decisions that can be implemented within this strategy. The strategy would also be affected by elements like timeframe and topic of choice which would be argued throughout the following chapters.

There are several different strategies to follow while conducting a research study, which is required for constructing a suitable answer towards the specific research question. Our study involves analyzing the effects of mindless AIs on individual mindfulness in a cybersecurity context. To understand these changes, we will design a qualitative research strategy.

Table 4 below describes different research designs that we find most relevant, however we will only choose one of these as the limited timeframe and resources would not allow the usage of a combined research design. The two research approaches come with a short description to get an overview of which is more suitable regarding our field of research.

Table 4 Research Design

Qualitative Research Design	Short Description
Case Study	Case study would offer rich and descriptive data for the instance of the phenomenon that is being studied (Johannesson & Perjons, 2014, p. 44).
Multiple Case Study	It is best suited when there are several cases that are connected towards the phenomenon, where connecting similarities are essential when

	conducting this design (Hunziker & Blankenagel, 2021, p. 172).

The qualitative approach could involve several research designs, but is highly dependent regarding elements such as topic, cost, time and resources. In our case, the timeframe and available resources to conduct our research is very limited. Which would affect our decision when it comes to the choosing of a qualitative approach. It is also important to choose the right approach concerning the topic and relevance towards our research.

3.4.1 Longitudinal or Cross-Sectional

When we are creating this study we also have to take into consideration what kind of design we will follow, either a cross-sectional or a longitudinal. A cross-sectional design is a series of studies taken at one specific time or a short period, it gives a snapshot of the phenomenon we are studying (Johannessen, Tufte & Christoffersen, 2016, p.70). A longitudinal design is in contrast a study which is done at more than one occasion, you can say that a longitudinal study is multiple cross-sectional studies in one (Johannessen, Tufte & Christoffersen, 2016, p.71). In this thesis we opt for a cross-sectional research design, because of several reasons. First off is that a longitudinal study would have been better for our research question IF we had done interviews at the beginning, in the middle and the end of the implementation of AIs. This is of course something which is impossible or very time consuming regarding the time limit we have to complete this thesis. Based on all these factors we will do a cross-sectional study. It would also be combined with other design strategies mentioned in Table 4 Research Designs. Which will be further discussed in Chapter 3.5.

3.4.2 Sampling

To conduct any data generating methods, we need to decide which sampling techniques are best suited for our approaches. In Table 5 the different techniques are mentioned with a short description of when to use them.

Table 5 Sampling

Sampling Technique	Short Description
Representative Sampling	This sample represents the total population that is being studied which shares the same relevant characteristics (Johannesson & Perjons, 2014, p. 43).
Exploratory Sampling	This sample has an exploratory purpose which focuses on new explorable areas (Johannesson & Perjons, 2014, p. 43).
Random Sampling	This sample is randomized and would be represented as the targeted population; the randomization process is conducted where participants have an equal chance of being chosen (Johannesson & Perjons, 2014, p. 43).
Purposive Sampling	The sample would provide valuable information and an asset to the research, where the techniques are generated through exploratory purposes of the samples (Johannesson & Perjons, 2014, p. 43).
Snowball Sampling	The sample can suggest other participants to join the sample group throughout the research process (Johannesson & Perjons, 2014, p. 44).

For this study we find Random sampling to be the best suited for us. This is because of the limited time frame we have available and we want to find HROs that want/can participate in our study. Therefore it seems far more prudent to find a list of HROs and send out requests at random to them.

3.5 Qualitative Research Approach

A qualitative approach is better explained where “*Instead of providing a broad view of a phenomenon that can be generalized to the population, qualitative research seeks to explain a current situation and only describes that situation for that group.”* (Lowhorn, 2007, p. 3). The approach relies heavily on data collection directed towards a phenomenon connected to a group of individuals. Lowhorn also mentions qualitative methods focusing on the behavior aspect of these individuals and other characteristics as it is more inductive rather than deductive, where the approach has the ability to reproduce results (Lowhorn, 2007, p. 3).

There are several qualitative methods that have their own strengths and weaknesses depending on the circumstances. Each method is based on human interaction, which creates potential for a deeper understanding of a phenomenon, but also increases the probability of biased data. Some relevant methods used with this approach are interviews, observations and focus groups.

The research question regarding how individuals' mindfulness is affected by AI Assistants within the context of cybersecurity is a new topic. This creates a challenge based on the availability of the information needed within this type of research. For that reason, we have concluded that a type of case study would be the most suitable approach for this thesis.

3.5.1 Case Study

A case study is explained as to focus "*on one instance of a phenomenon to be investigated, and it offers a rich, in-depth description and insight of that instance.*"(Johannesson & Perjons, 2014, p. 44). What differs this strategy is the level of details painted by researching a phenomenon with the use of different methods as observation or interviews. Johannesson & Perjons sums up all the characteristics of a case study, where focus on one instance and depth within a natural setting, creating relationships and processes are studied by using multiple sources and methods (Johannesson & Perjons, 2014, p. 44). A case study can be conducted in several ways, exploratory, descriptive and explanatory. Exploratory focuses on the unknown, where information about the phenomenon is scarce. Descriptive aims to produce a large amount of information regarding the phenomenon, while explanatory focuses not only on the ’what?’, but also the ’why?’ certain events happen (Johannesson & Perjons, 2014, p. 44). However, in this thesis we aim have several informants from different HROs, thus creating multiple cases.

3.5.2 Multiple Case Study

In instances where there are several cases, a multiple case study research design would be the preferred approach. While a single case study focuses on one instance of the phenomenon, a multiple case study compares similarities and differences within the same environment, among several cases (Hunziker & Blankenagel, 2021, p. 172). The advantages by conducting multiple case studies is the amount of collected data that is both comparable and trustworthy rather than a single case study (Hunziker & Blankenagel, 2021, p. 183).

One of the biggest weaknesses within this type of design is the generalization of the results, this counts for both single and multiple case study. For example if a case study is conducted towards a small scale business, the findings would be only relevant towards businesses with the same scale and scenario. Which makes larger businesses irrelevant when it comes to that certain case scenario (Johannesson & Perjons, 2014, p. 45). There are also disadvantages with multiple case studies, where such design is both time consuming and costly to conduct (Hunziker & Blankenagel, 2021, p. 183-184).

The findings from the literature analysis presents the negative and positive consequences by AI integration within any organization and business. It also focuses on the different aspects such as mindfulness, mindlessness, and cybersecurity within this context. To further investigate this phenomenon, it requires several case studies that cover this unfamiliar topic, where we can investigate and compare these cases to form a better understanding of this phenomenon. For that reason, we have concluded that a multiple case study combined with a cross-sectional research design is the most preferable approach within this thesis.

There are multiple methods that can be conducted towards the multiple case study design. To conduct an observation can be extremely hard as we are studying cybersecurity related content, which would require permission to observe employees that are connected towards those environments. To conduct a focus group session would also be difficult because of the availability issues that may arise when inviting several participants that are also connected towards these environments. The thesis is expected to start and finish within a 5-month period, which is important to note when it comes to method of choice, that it can generate enough relevant data within this time period. The most efficient method is to conduct several interviews with targeted participants, which can give an in-depth description of the phenomenon we are studying. An interview is easily the most practical decision as it also is time efficient, cheap and can cover several elements that are connected towards our research question.

The multiple case study will have an exploratory focus and will include interviews as the method of data collection. The sampling would be random, as it targets several random HROs. The interview would be conducted one time per HRO, as our research also includes a cross-sectional design approach. The interviews will be transcribed, where the data is analyzed with the use of NVivo, the process is described in both Chapter 3.5.4 and Chapter 3.6. The validity of the data is based on several factors which we are aware of when conducting this design approach and described in Chapter 3.5.5.

3.5.3 Interview

An interview is described as being "*a communication session between a researcher and a respondent, in which the researcher controls the agenda by asking questions of the respondent*." (Johannesson & Perjons, 2014, p. 57). In our case our respondent is related to the cybersecurity department within a HRO. An interview can be constructed in many different ways to "*allow for a more or less structured interaction between the researcher and the respondent*" (Johannesson & Perjons, 2014, p. 57). The information we retrieve is hugely dependent on how we construct our interview. An interview can either be constructed as unstructured, semi-structured or structured.

Unstructured interviews are when "*the researcher is as unobtrusive as possible and lets the respondent talk freely about a topic without being restricted to specific questions*" (Johannesson & Perjons, 2014, p. 57). This technique is more effectively used when the topic of the research is not well known. A semi structured interview is used when "*investigating complex issues, as the respondents can express their ideas and feelings in a more unrestricted way*" (Johannesson & Perjons, 2014, p. 57). A structured interview is a more restrictive technique that "*follows a predefined protocol and is similar to a questionnaire*" (Johannesson & Perjons, 2014, p. 57).

The unstructured and semi structured interview is best suited for investigating complex issues, where the informants are less restricted in comparison to a structured interview (Johannesson & Perjons, 2014, p. 57). In our context a semi structured interview would be best as we do have categories and questions we want to ask, but we do not want it to feel restrictive for the informants. We want them to open up and tell us as much as possible and we want to be able to ask follow-up questions or ask for clarifications if we need to. It is also important to note that our RQ concerns a phenomenon that lacks both research and resources such as research papers available online, which would impact on how we will design our interview.

An interview can be conducted within almost any research design approach. It is both cheap and time efficient to construct and conduct, which can be done digitally or physically depending on the situation. The influx of AIs in everyday life is, and will continue to be, a major digital transformation which can lead to many unforeseen consequences. It is this that we want to study in detail, more specifically how the mindfulness of the employees in HROs are affected by, perhaps, using AIs in many of their daily tasks.

There are different challenges we should be aware of while conducting an interview, Anyan describes that "*Both the interviewer's scientific competence and the interviewee's behavior are examples of power manifestations in the qualitative interview research*." (Anyan, 2013, p. 6). This can create backlash during the interview and analysis. It is also important to construct exploratory questions to avoid short answers such as ‘yes’ and ‘no’ when the interview is based on a qualitative semi-structured interview. It is important to interact with the interviewee during the conversation. This can lead to unforeseen topics to be discussed, which grants us a broader understanding of the phenomenon. One of the disadvantages by conducting an interview is that it is time consuming, which relates to the transcription and the analysis of the interviews. It is also dependent on our own characteristics and personal attributes, which can affect the interview's outcome (Johannesson & Perjons, 2014, p. 58).

By using random sampling technique, we are not targeting any specific HRO, instead we are randomizing the sample group by sending emails to a handful of candidates. The first eight or so agreeing to participate in our research would be included within our interview process.

As mentioned, our research question explores an unfamiliar phenomenon which lacks documentation or research. The different topics related to this phenomenon on the other hand is better documented in Chapter 2, which makes the semi-structured interview our best approach for this type of research. The semi-structured interview would contain open questions regarding the topic that are being discussed. Follow-up questions will also provide better details in some sections if we find something unclear or interesting. The interview would also contain several main topics established earlier, such as AI, hallucination, mindfulness/mindlessness within the context of cybersecurity. We would also implement several questions that involve topics like cybersecurity awareness, trust and machine learning within organizations to better understand how these elements can be connected towards the main topics. The interview would be conducted digitally, which allows us to communicate with the informants regardless of geographical location and would be held 1 time each. The interviewee would also be informed that he/she would be pseudonymized to protect their identity and that they are being recorded for transcription purposes only. The consent form for the interview subjects, which describes how the interview will be conducted and how the information will be stored and processed is in Appendix A while the Interview Questions will be in Appendix B.

3.5.4 Research Analysis

The interviews will be held on either Zoom, Teams or Meet, mainly because they are simple to use and it is easy to get a recording of the interview itself where you get an audio file. That audiofile needs to be transcribed. For this we used Microsoft Word which has an integrated function that lets us upload said audio file directly and then it gets transcribed. This gives us a rough version of the transcription, because the AI is not that good with Norwegian yet so we have to manually go over the text while listening to the audio file. This is of course time consuming, but would have been even more so without Word.

When all the interviews are transcribed, we need to analyze exactly what each interview is about. The way we do this is we use Coding. Coding is a way to put labels on what the informant says, this can be done in programs like NVivo. After this is done, we can begin the real analysis and put the findings into context and see what our informants agree or disagree on. For this analysis we will use a thematic content analysis which is described as "*Weeding out biases and establishing your overarching impressions of the data. Rather than approaching your data with a predetermined framework, identify common themes as you search the materials organically. Your goal is to find common patterns across the data set*." (Rev, 2022). While our informants may talk about things that seem unrelated to our RQ, they may unintentionally touch upon something related to mindful individuals, mindless AI, or other relevant information within the context of cybersecurity, which in turn will help our analysis. This gives us a better way to contextualize and compare the information our informants give us.

NVivo is our preferred approach regarding coding transcripts, as the program is free and easy to use. The coding process with NVivo is described in detail in Chapter 3.6.

3.5.5 Validity of Findings

After analyzing the data retrieved from the transcripted interviews, we need to ensure the validity of the data. As mentioned in Chapter 3.5.2, case studies has one major disadvantage regarding the validity of the data, which is generalization. HROs in Norway vary in sizes, available resources, population, equipment and skills. This is an important element to consider when validating the data retrieved. One of the most important methods to ensure a great validity is to make this study replicable.

By collecting other relevant literature, as done in Chapter 2, we can find the gap in the research and make our analysis more comparable, which also increases the validity of our research. Another important step is to ensure that our research approach is motivated by being self-critical towards the construction of the interviews. This would reduce the amount of bias that is reflected within the interview questions.

These steps would ensure and decrease the probability of bias within the data that is being collected. The construction of the interviews will be affected by these elements to ensure greater quality of data. By making sure that other researchers could replicate our study with the same findings, would ensure that our data is valid. Doing this in case studies on the other hand, would be difficult, as the problem with generalization could also affect the result by interviewing other organizations that are not included within this study.

One important aspect for us to consider is our inherent biases when analyzing these interviews. First off, the interviews were held and transcribed in Norwegian. In Chapter 4 we have of course taken snippets from these interviews and translated them into English. The most obvious problems with this are of course that we take their quotes out of context or that we mistranslate or misunderstand the meaning of them for it to fit better into what we want to find. We are human of course and therefore it is natural for us to perhaps want to find something that corresponds with our preconceived beliefs. In Chapter 2.2.2 this phenomenon was described as a confirmation bias and is a problem. However, this must not be allowed to steer the analysis and overall conclusion of the thesis away from what the informants say, even though what they say may be the polar opposite of what we expect to find. Another issue is generalization, which was explained earlier. This thesis aims to include a minimum of 8 interviews. There are of course plenty more organizations that are in the critical infrastructure or HROs, so our findings here cannot be used to state what definitively happens with mindful individuals that use AIs in HROs. However, it can be a good starting point for future research and a guiding pin to understand AIs effect on an individual. We can minimize the issue as mentioned earlier by conducting a literature review beforehand and being self-critical towards constructing the interview questions.

3.6 NVivo

When you are analyzing and coding the results from the interviews into NVivo, or other comparable tools, it is important to adhere to a methodology. This study will adhere to the method described in the report *Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology* (Gioia, Corley & Hamilton, 2012). Figure 5, 6, 7, 8 and 9 in Chapter 4 shows how this method works in practice.

The model is split into 3 parts, the 1st Order Concepts, 2nd Order Themes, and finally the Aggregate Dimensions. The 1st Order Concepts "*which tries to adhere faithfully to informant terms, we make little attempt to distill categories, so the number of categories tends to explode on the front end of a study*" (Gioia et al. 2012, p.20). After coding all the transcribed interviews in this study we ended up with 78 different categories and concepts. This is of course a large number and can be overwhelming and easy to get lost in, but "*You gotta get lost before you can get found*'' (Gioia, 2004, referenced in Gioia et al. 2012, p.20).

When this is completed, we can try to find similarities between the different codes and sort them accordingly into bigger categories as shown in Figures in the 2nd Order. Here it is now "*firmly in the theoretical realm, asking whether the emerging themes suggest concepts that might help us describe and explain the phenomena we are observing.*" (Gioia et al., 2012, p.20). After doing this we ended up with 12 2nd Order Themes.

Lastly there is the possibility to distill these themes further by combining several of them to create what is known as Aggregate Dimensions. As Figure 5, 6, 7, 8 and 9 shows, the themes are combined into far fewer Dimensions based on which fit together the best. This thesis ended up with 5 Dimensions, AI Strength, AI Risks, Mindfulness, Mindlessness, and Trust & Deceptiveness.

3.7 Ethical Considerations

The qualitative approach includes some ethical considerations. The approach should respect the potential ethical issues that may arise from creating and conducting any research. We are basing our ethical considerations with the use of National Research Ethics Committees or in short NREC's guidelines which specifies several ethical principles that should be considered in this thesis. It is important to notice that these guidelines should reflect a general view of the ethical element, and should not be replaced with the subject-specific guidelines (Torp, 2019).

Our report needs to follow the four principles mentioned by NREC, which is:

- To respect any participants that are involved within the study (Torp, 2019).
- Our research should only produce good consequences as the outcome of the study (Torp, 2019).
- Our research should be implemented and designed in a fair manner (Torp, 2019).
- We should contain good integrity, where we behave openly, honestly and responsibly for the public and other colleagues (Torp, 2019).

We also follow their guidelines that are listed on their webpage, which involves several different ethical aspects within this research. The thesis should focus on uncovering the truth, which should reflect academic freedom, where our approach is motivated, chosen and driven by us. We should also ensure high academic quality of the thesis, which possesses all the necessary requirements for creating and conducting this study. Participants should also consent to be a part of it, where confidentiality is used to hide their identity. Our study needs to include impartiality and integrity to add openness and trustworthiness to our research. The report shall credit/cite other authors if their content is cited during this report, or theory that is partially or fully based on any decisions made under this research. We also need to show respect towards our participants, other researchers and the institution we belong to. We and our research shall also show social responsibility by respecting the community within the different organizations that are involved within this study. We are also to follow any national laws and regulations that are connected to the field of study (Torp, 2019).

3.8 Data Sources

For this thesis we have conducted 8 interviews with 8 different Norwegian organizations that fall under the categories of critical infrastructure and HRO. Table 6 shows the list of informants with their pseudonyms (I-1, I,2 etc.) and what occupation they had in their respective organizations. In Table 6 there is also which sector the informant's organization falls under, this is based on the naming of these sectors shown in Figure 1 in Chapter 1.

Table 6 Informants

Informant	Role	Sector
I-1	Head of IT-Department	Defence & National Security
I-2A-D	CISO, Advisors, GRC	Transport
I-3	Senior Engineer	Energy
I-4	Head of Information Security	Health
I-5	Security Engineer	Defence & National Security
I-6	IT Emergency Manager	Transport
I-7	CISO	Energy
I-8	Identity and Access Governance Architect	Food & Grocery

As you can see, I-2 is named I-2A-D, this is because there were 4 informants that took part in this specific interview. For the vast majority of the interview we talked to the CISO, however the other informants came with some clarifications and their own experiences and thoughts as well to back up statements.

In order to get in contact with these informants we sent out emails to every organization that fell under our criteria, and since we have random sampling, as explained in Chapter 3.4.2 we did not prioritize some organizations over others, we simply needed some to participate.

In the emails that we sent out, we presented ourselves quickly by stating what and where we studied, and asked if they had the opportunity to participate in this study. We asked for employees with competence in Cybersecurity, AI, Mindfulness/Mindlessness, Security Awareness and Security Culture. Then we quickly explained what we wanted to research and that we were open to have the interview on any platform they were comfortable with. Lastly we added our contact information and wrote that if they had any questions whatsoever that they could reach us on email. We also added an attachment that you can see in Appendix A that is the consent form where we state the purpose of our study and what we will do with the data, who processes it, when it will be deleted, their general rights according to GDPR etc..

The interviews were held on Microsoft Teams, and we used an external recording software OBS to record them. Before the recording starts in the interview, we will give some information to the informants so that they are aware of their rights.

First off, we start by welcoming them to the interview and thanking them for taking time out of their busy day to talk to us for the ~30 min this interview will take. Then we introduce ourselves shortly by stating our names, age, where we are from and what and where we study. We do this in order to create a more personal feeling between the informant and ourselves so that they are hopefully more comfortable with the setting. We then give a short introduction to the topic we want to talk about and make it abundantly clear that if there are any questions, they do not feel comfortable answering they of course do not need to answer them. Consent is important for us, and we do not want to strongarm them into something they do not want to answer.

Lastly, we tell them about the recording process and ask if they consent to it being recorded and say that their names will not be stated in the report, just a pseudonymized version e.g. ‘I-1, I-2, etc.’.

After this we begin by warming them up with letting them introduce themselves so that we can understand their experience and creating a more personal connection for the duration. The main parts of the interview start by asking what their thoughts on AI are, especially in the context of mindfulness and security culture. We then have some questions regarding awareness connected to AI and Machine Learning in general.

When we are finished, we once again thank them for finding time in their busy day to be able to take part in this study and bid them farewell.

It is important to note that since we had open questions, the informant may answer multiple questions unwittingly in one answer, if that is the case, we will not ask the follow-up question but rather a clarification question if there is something unclear in their answer. For that reason, we have made the follow-up questions italic within our interview questions, as not every question might be asked during the interview if the context does not match the current topic.

The transcription is made with the use of Microsoft Word and the analysis is done with the help of NVivo. In the coding process we made 5 overarching codes, Mindfulness, Mindlessness, AI Positives, AI Negatives and finally Trust & Deceptiveness. In the analysis mindfulness and mindlessness have been combined into one subchapter Mindfulness vs Mindlessness, the same applies to AI Positives and AI Negatives which are combined to AI Positives & Negatives. We chose these because they reflect the most prevalent aspects of our research question.

4 FINDINGS

In this thesis we have conducted 8 interviews in the Norwegian critical infrastructure sector, the interview subjects were mostly CISOs and department heads of IT, but also security engineers and architects so the informants were all well informed about topics that our research question regarded.

A quick reminder, our research question is: ***How will AI Assistants Affect Cybersecurity Mindfulness in High-Reliability Organizations?*** In this chapter we will analyze what the informants had to say and see what they agreed and possibly disagreed on. For this purpose, we will conduct a thematic content analysis based on the Codes we created in NVivo which were explained in the previous Chapter. The informants themselves will only be referenced as their pseudonyms, i.e. I-1, I-2 etc.

The codes we created in NVivo are shown in Figure 5, 6, 7, 8 and 9. Where they first are made into concepts, then grouped into themes and finally aggregated as explained in Chapter 3.6. The aggregated Code is representative for how this chapter will be structured.

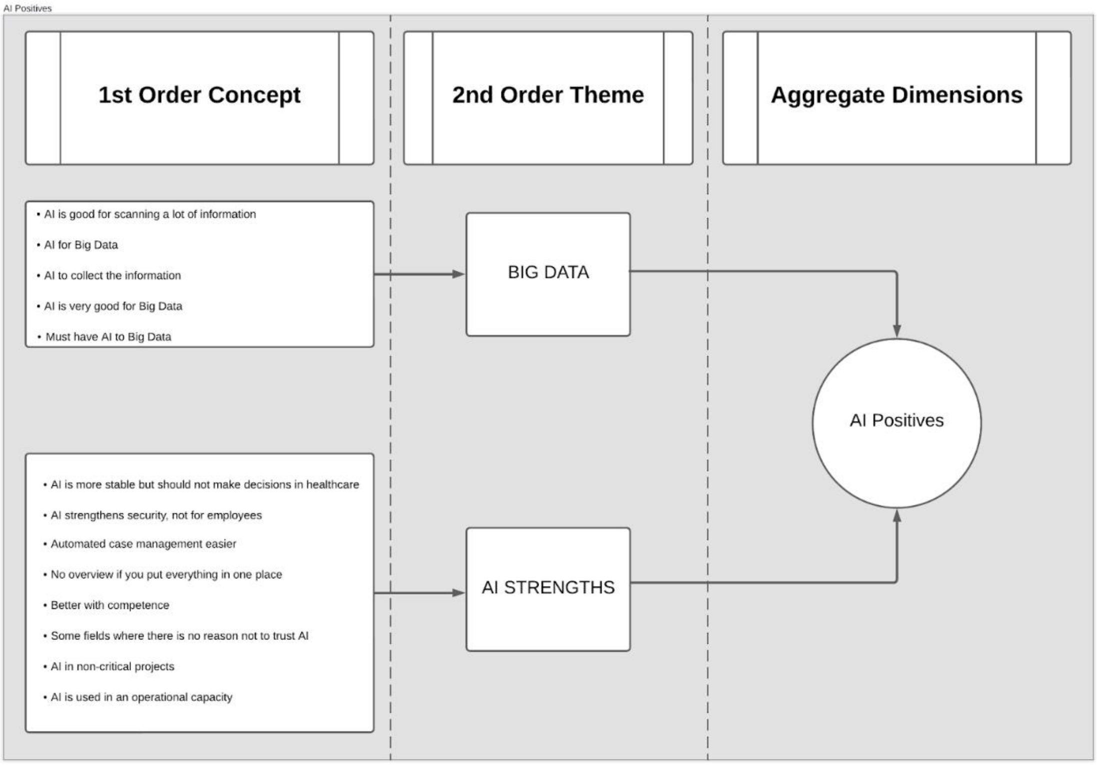

AI Positives
1st Order Concept
2nd Order Theme
Aggregate Dimensions
• AI is good for scanning a lot of information
• AI for Big Data
• AI to collect the information
• AI is very good for Big Data
• Must have AI to Big Data
BIG DATA
AI Positives
• AI is more stable but should not make decisions in healthcare
• AI strengthens security, not for employees
• Automated case management easier
• No overview if you put everything in one place
• Better with competence
• Some fields where there is no reason not to trust AI
• AI in non-critical projects
• AI is used in an operational capacity
AI STRENGTHS

Figure 5 AI Positives NVivo Model

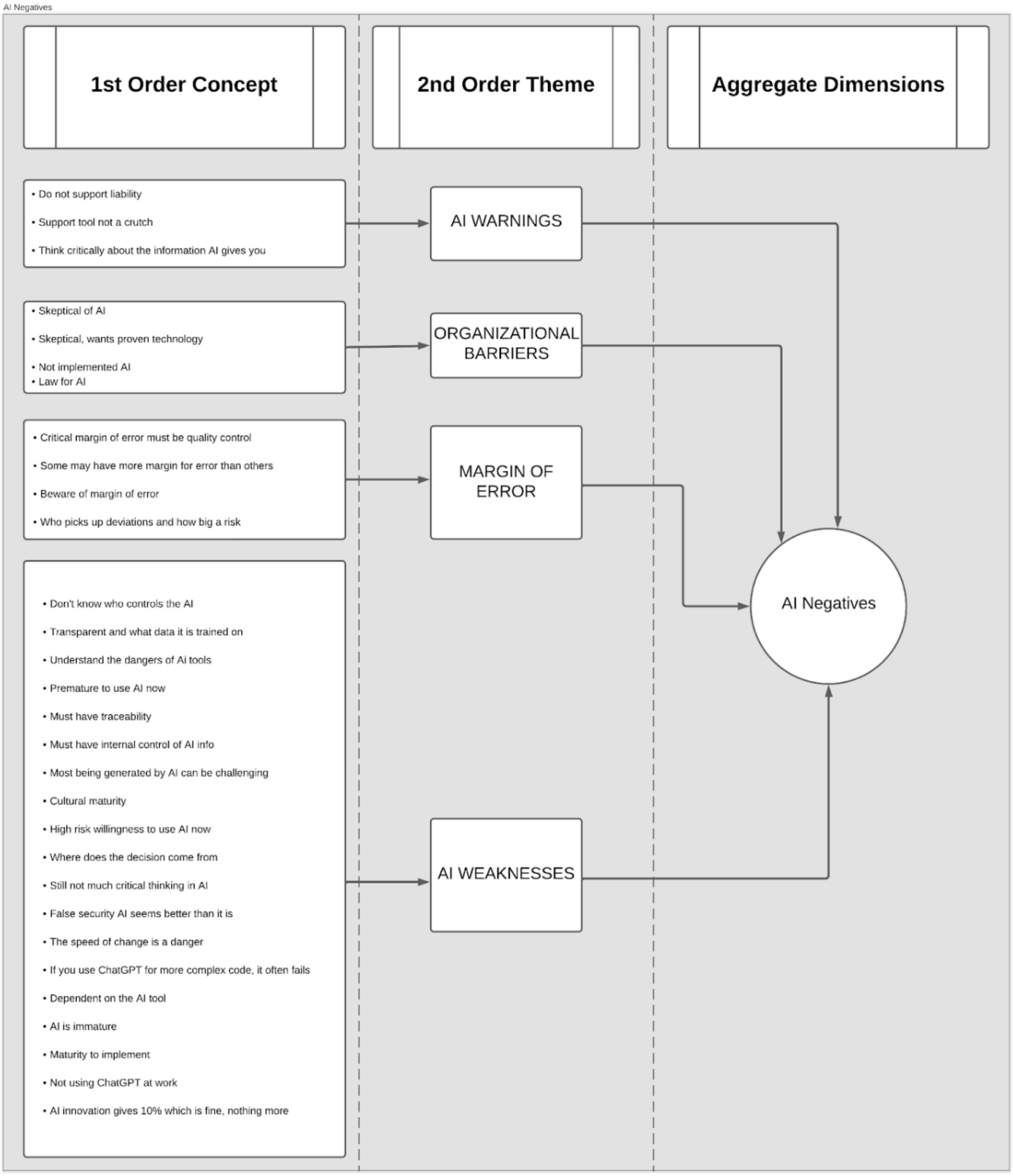

AI Negatives
1st Order Concept
2nd Order Theme
Aggregate Dimensions
• Do not support liability
• Support tool not a crutch
• Think critically about the information AI gives you
AI WARNINGS
• Skeptical of AI
• Skeptical, wants proven technology
• Not implemented AI
• Law for AI
ORGANIZATIONAL BARRIERS
• Critical margin of error must be quality control
• Some may have more margin for error than others
• Beware of margin of error
• Who picks up deviations and how big a risk
MARGIN OF ERROR
• Don't know who controls the AI
• Transparent and what data it is trained on
• Understand the dangers of Ai tools
• Premature to use AI now
• Must have traceability
• Must have internal control of AI info
• Most being generated by AI can be challenging
• Cultural maturity
• High risk willingness to use AI now
• Where does the decision come from
• Still not much critical thinking in AI
• False security AI seems better than it is
• The speed of change is a danger
• If you use ChatGPT for more complex code, it often fails
• Dependent on the AI tool
• AI is immature
• Maturity to implement
• Not using ChatGPT at work
• AI innovation gives 10% which is fine, nothing more
AI WEAKNESSES
AI Negatives

Figure 6 AI Negatives NVivo Model

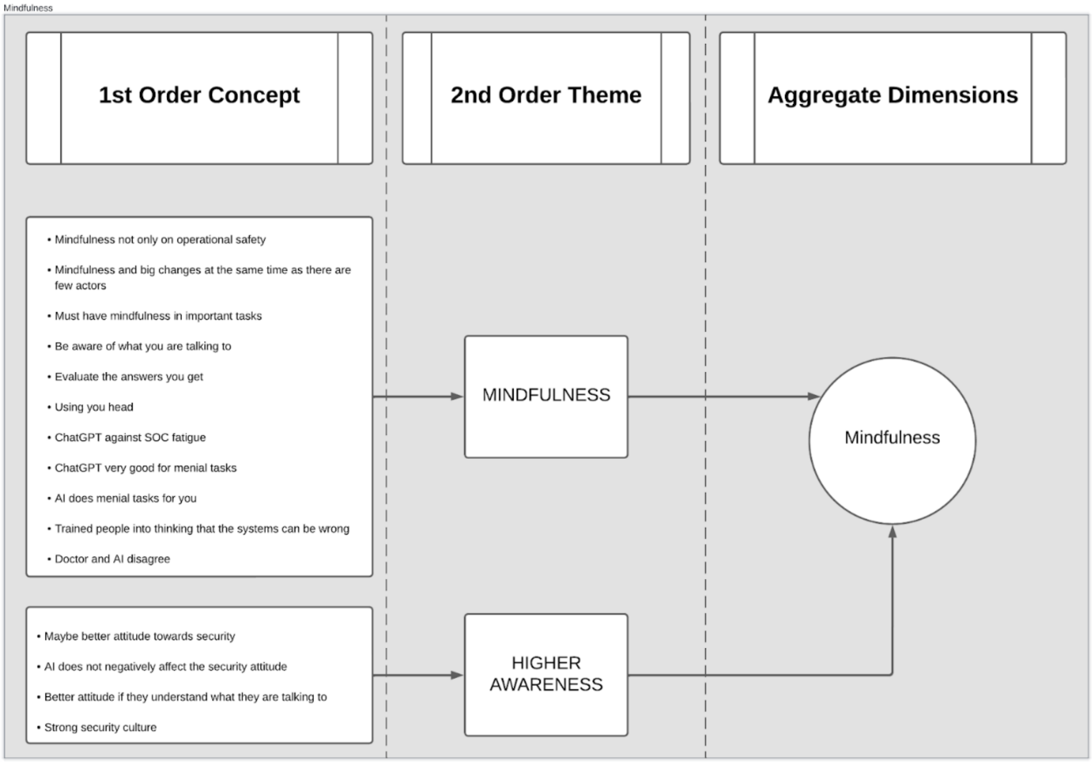

Mindfulness
1st Order Concept
2nd Order Theme
Aggregate Dimensions
• Mindfulness not only on operational safety
• Mindfulness and big changes at the same time as there are few actors
• Must have mindfulness in important tasks
• Be aware of what you are talking to
• Evaluate the answers you get
• Using you head
• ChatGPT against SOC fatigue
• ChatGPT very good for menial tasks
• AI does menial tasks for you
• Trained people into thinking that the systems can be wrong
• Doctor and AI disagree
MINDFULNESS
Mindfulness
• Maybe better attitude towards security
• AI does not negatively affect the security attitude
• Better attitude if they understand what they are talking to
• Strong security culture
HIGHER AWARENESS

Figure 7 Mindfulness NVivo Model

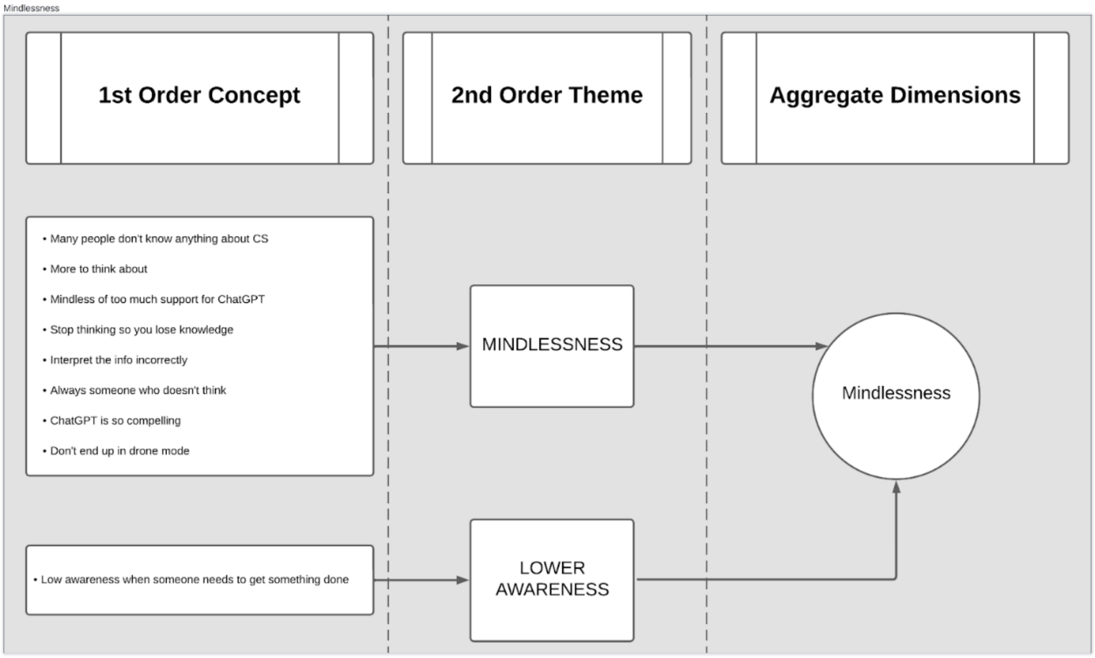

Mindlessness
1st Order Concept
2nd Order Theme
Aggregate Dimensions
• Many people don't know anything about CS
• More to think about
• Mindless of too much support for ChatGPT
• Stop thinking so you lose knowledge
• Interpret the info incorrectly
• Always someone who doesn't think
• ChatGPT is so compelling
• Don't end up in drone mode
MINDLESSNESS
Mindlessness
• Low awareness when someone needs to get something done
LOWER AWARENESS

Figure 8 Mindlessness NVivo Model

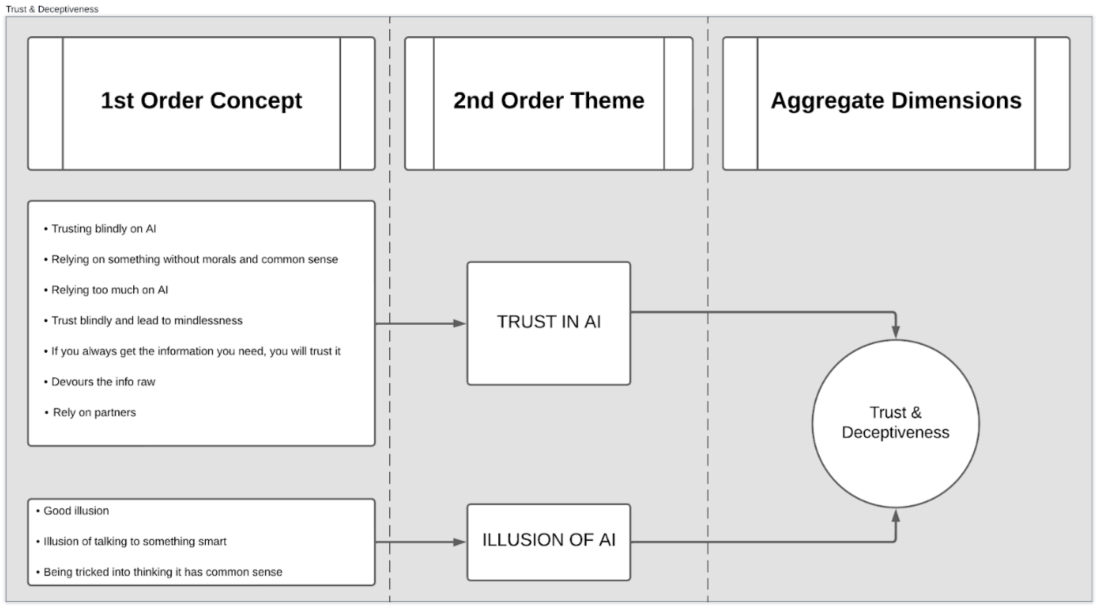

Trust & Deceptiveness
1st Order Concept
2nd Order Theme
Aggregate Dimensions
• Trusting blindly on AI
• Relying on something without morals and common sense
• Relying too much on AI
• Trust blindly and lead to mindlessness
• If you always get the information you need, you will trust it
• Devours the info raw
• Rely on partners
TRUST IN AI
Trust & Deceptiveness
• Good illusion
• Illusion of talking to something smart
• Being tricked into thinking it has common sense
ILLUSION OF AI

Figure 9 Trust & Deceptiveness NVivo Model

This is how the data from our study is represented within the Gioia methodology. From this it is far easier to have a structured way of presenting our findings. These five figures illustrate the coding process of NVivo that is explained in Chapter 3.6 and represents the foundation of our findings.

4.1 AI Positives & Negatives

The concept of AIs in today's modern world is a topic whose public interest seems to rise exponentially, especially with the introduction of LLMs and other AI tools which can drastically help its users generate answers to whatever question they may have.

One of the biggest concerns seems to be in regards to how the AI was trained and on which data. I-5 expressed his concern by stating that:

> "*An AI is not perfect… You can train an AI to recognize pictures of a dog for example and it is correct 99% of the time, but as I mentioned earlier there will be a deviation and who will pick that up? And how big of a risk is it?*".

Of course most of the businesses in the world would be happy with just 1% deviation, however I-5 makes it clear that an example business 'Olas Leatherjackets "*could implement machine learning with a 9% deviation without any major consequences, but we cannot accept such a deviation, it's not possible. A deviation of 1% for us is critical.*". However I-4 is of the opinion that an AI "*is* *repeatedly more accurate than a human is*" especially in regards to scanning through big data sets such as x-rays to look for breast cancer and that in these scenarios "*AI can play a decisive role*". This was corroborated by I-6 who used the same example regarding medical scans and said that "*there is no reason not to trust it*" in regards to AI. I-4 was quick to offer a dilemma however "*What do you do if the AI suggests that the patient is sicker than what the doctor's opinion states. And what becomes the doctor's responsibility if their recommendation is different to the AIs*?". This is really the crux of the problem, who should you listen to? The doctor who is educated and has working experience or an AI which is simply lines of codes trained to recognize patterns, both may be wrong but it is an interesting dilemma.

I-6 gives us some insight into how you could think about it by asking "*How transparent is it? ... everything is automated. You don't get any insight. How did it come to this conclusion? What data is it trained with?*". This hints at a level of skepticism regarding how much responsibility you assign the AI. Whereas when considering Big Data I-5 states that "*A good model can eat through it and give you the pattern you would not have seen*", I-7 also says that "*Automatic detection and understanding large amounts of data is essential because there is so much of it.*" and while this is obviously a good thing I-5 is also insistent that "*There must be quality assurance control*" because there will always be a deviation. I-5 also shared his thoughts regarding the danger of source code poisoning, where he stated the difficulty of executing this type of attack because the "*amount of data on which it is trained is completely absurd*" and quickly followed up with "*An AI model requires you to have quite a lot of data to twist it, but at the same time it is already the case that it does that*". I-6 was under the impression that one of the biggest problems was the "*rate of change ... and complexity*" while I-4 complained that there still was not much "*critical thinking in these AI products*".

The organizations however seem to be quite skeptical and reluctant regarding implementing more than simple pattern recognition of Big Data AIs, not simply because they feared it but because many of these organizations are state-owned. I-1 claimed that "*It's a long way to go, but this is a legal issue. The technology is here.*" while also expressing that there is a need for AI but that first you need to have "*laws and regulations*" for the use of it. I-4 stated that in their organization they had a positive pilot of an AI but that they "*have not implemented anything in response to this*". In addition, I-2 regarding LLMs said that they think "*those are tools that we don't use at work, quite yet.*".

However, many informants agree that using AI can strengthen the cybersecurity aspect of their organizations, at least in some areas. I-2 thought that "*Within IT development, AI can be a strength for security, but for the normal case manager then I doubt that AIs will help any.*" while I-7 simply stated that "*I think it has a positive effect*".

I-5 explained how an AI can positively affect an employee by explaining the concept of SOC-fatigue, which simply is that a first-line, or second-line, employee must go through essentially menial tasks and many false positives, and they get fatigued by it. Regarding SOC-fatigue the informant offers a solution to it "*When it comes to AI there is ChatGPT*". The biggest positives that LLMs offers is according to I-5 "*You are going to answer a generic email you have answered a hundred times before, then it is very quick to give you a block of text that you can proofread and send off, saves you 15-20 minutes, right?".* While I-5 is not enthusiastic about LLMs making entire code sets that will be deployed, he is positive regarding making "*boilerplates, in order to not spend 15 minutes coding something, I have coded 100 times before*". From this we can see that AIs can help by alleviating the employees from unnecessarily menial tasks and thus keeping their 'head in the game' so to speak.

4.2 Mindfulness & Mindlessness

As we saw in the previous part, there is a lot of skepticism and issues regarding AIs, but also positives as well. The informants reflected the mindful aspects regarding the usage of AI, which delves into different elements as cybersecurity awareness, training, validity of information regarding AI output and mindlessness. As I-7 stated, "*again, it's a matter of understanding, what does it mean to use a new tool of this type? What risks? What threats?"* regarding AI's capabilities towards implantations and usage.

One of the risks covered within the interview process was explained by several informants regarding the validity of AI output and its sources to that certain output. I-2 explained that "*there have been quite a few AIs popping up that can be used to misinform, for example, Russian propaganda and that sort of thing*", the informant further explains the problem with the biggest AI's available today as they might have a long queue-time. This can cause users to explore alternate AI tools, which can for example spread different information, or organize the output in a certain way to misguide or influence the user in a certain way. I-3 explains one the problems with AI's accuracy within their output response is the validity of the sources attained within the output itself, which he further explained that you should "*make sure that you have a description of the assessments, where do they come from, what are the sources? What did I use for this particular assessment?*". The validity of the information retrieved by using AI can certainly be a risk as it would require additional knowledge to verify the output, but as I-7 further explains the issue when the AI structure the answer to look believable, "*then the question is whether you should trust it blindly?*" and further explains with an example that "*implementing ChatGPT in outlook [and receive a] button that says 'generate answers for me' then we start to get into mindlessness*", but rather use it more as a support tool.

The importance of adapting to this new technology is stated by I-6 where "*90% of the content would be AI generated on the internet*" within 2026-2027, which might be a plausible scenario as the following growth of AI's popularity has risen within this year. It can also be a dangerous scenario as the AI might construct answers that are purely based on present AI generated data, which again can complicate the validity of the AI outputs. This can also create a problem of misinformation, where every user of an AI should be critical of certain outputs that might be false. To combat this fast growth and avoid such scenario, I-1 said that "*if you are aware and you kind of take 2 steps back to make sure that you actually use your head, then you can use this for something positive.*", where he also added that "*Then perhaps it will be a new everyday life for the employees, where they have to think for themselves and take care of the information they have that is their company, for example on a phone.*" related to AI implementation. Some informants were also rather skeptical of the AI in its current state, where I-3 shared his concerns about starting using AI too early, where the informant stated that it is simply too "*premature, a number of other frameworks need to be in place in order to be able to use it.*". Where he also added that it might have a negative effect towards security awareness based on this reason. People within the security context can in some cases misuse this type of technology as they might lack any awareness training towards this field. I-4 explained that their mindfulness towards information security has indeed seen an increase in the last couple of years within the health sector. This is not just within the "*operational security*" but also how employees "*maintain confidentiality in user processing of clinical data.*" within the health sector, but he later quotes "*that there is always a village idiot*" which refers to employees with lower awareness that might make mistakes. I-4 also explains the risks of using this tool without "*critical sense towards the decisions or the recommendations that the tool gives you*", where I-8 also added within the same scenario, that users might get an illusion effect by using it without critical thinking. This type of critical thinking and being aware of the illusion that is represented by the AI, could also play an important role when compared to the same scenario mentioned by I-6.

Several informants explained the need for training regarding AI implementation, where I-3 states that when you "*order an answer from an AI, then you must be able to assess the answer you get, so it's a new competence that you have to have. And it is demanding.*". This is also explained by I-4 where critical thinking is important to have when using such tools. A user should always verify the output of an AI, and always be aware that the system sometimes can be wrong. I-4 also explains that "*we use both tools, for training, and then we use campaigns and* *training talks with people*.” to increase the awareness towards cybersecurity within the organization. He also believes that this training should be “*individually adapted*” within the context of using AI and implementing “*internal control mechanisms built into these processes*” where verification towards the system output is understood and interpreted correctly. I-7 also explains that AI could be a helpful tool to better “*arrange more effective training*” towards user's awareness, which also reflects how useful it is to implement and use AI in a helpful way.

4.3 Trust & Deceptiveness

The skepticism surrounding AI often affects the overall trust towards its abilities. The concerning factor of relying too much on this type of new technology can be both scary and fascinating at the same time. AI, which has recently gotten an increase of popularity and accessibility makes the technology easy to use by everyone, which has never been done before. This can create some concerns towards the accuracy and credibility of the AI’s outputs, how it stores/uses the data which is shared with it and the algorithm behind it.

These concerns mainly occur as this type of technology is both new, lacks regulations and research. This view is shared with I-2, who stated that:

> *“I too am generally skeptical of AI as it is today. When you work with cybersecurity in a company that operates critical infrastructure, you are a bit ‘stiff” about not wanting to be first in the queue to open new solutions. We want to use proven technology.”.*

This skepticism may be due to the immaturity of AI output, which can provide inaccurate solutions. As I-5 stated that you should avoid AI to provide solutions around complex tasks, which in return could give “*often wrong or not fully optimized*” answers, depending on the end user's knowledge surrounding the initial problem.

Several informants stated their concerns surrounding how convincing the AI in many cases could be. I-4 further elaborated this by explaining the danger to “*blindly trusting the AI*”, where the informant explained that AI “*provides decision-making support, it does not take over decision-making responsibility*.”. The informant referring to AI as a support tool, this statement is also shared by I-5, who further explains the accuracy of AI output, which can “*hit 99 out of 100 times*”, but further shares his concerns about the 1% inaccuracy. Sometimes the AI can give wrong output, why is that? I-7 states that users “*don't really have the opportunity to go in and check and really understand why that output comes*.”. The output is not verified by the AI itself in most cases, where I-3 also explains that AI should have output that include "*traceability, until then, no matter how good it is, it is worthless*.".

Several informants expressed their opinion on the overall trustworthiness towards the AI's output, but this concern on the other hand can also be related to other tools like Google, which I-7 states that

> "*I think that in a way you want to trust it, the same way that you trust a lot of other things today when you google everything you do, then there are many people who just swallow the information raw, and I think so in a way it will be with AI.".*

He further states that being convinced of the AI's output being continuously correct can lead to higher "*trust, and it can, in a way, take a huge hit*" in the future. Another aspect shared with I-3 points out his concerns regarding the "*decision-making processes*" of the AI, which is currently new and untested, which is yet to be "*constructed in a good way*".

Lastly, I-8 described an AI as "*super autistic*" which can be both "*incredibly efficient and has an intelligence that surpasses everything*". I-8 corroborated the same as the other informants, but also added the comparison which metaphorically humanizes the AI. The informant also stated that an AI has a "*smart spectrum of intelligence [but lacks] common sense*" and added that an AI can be pictured as "*a sociopath".* He also stated that an AI has "*no limits to behavior*", which he further corroborated by saying "*You can trust the AI as much as you want, as you would from such an individual, total absence of common sense and no morals*". The same informant adds that the user gets "*the illusion that you are talking to a sentient being*", and further states that "*there is no conscious creature there, but the illusion is good*". The informant also explains his concerns about this, where people might for example "*be tricked into giving up information*". I-8 also sums up a mitigation towards this problem where he states that "*In other words, there are new risks that must be taken into account, and if they do, they develop an understanding of security in order to handle them properly.*".

5 DISCUSSION

In order to discuss a possible solution to the problem statement, it is important to mention the research question which is as follows: ***How will AI Assistants Affect Cybersecurity Mindfulness in High-Reliability Organizations?***. This is to ensure that the discussion is highly relatable towards the scope of the thesis and how our analysis could contribute to a conclusive answer towards the RQ.

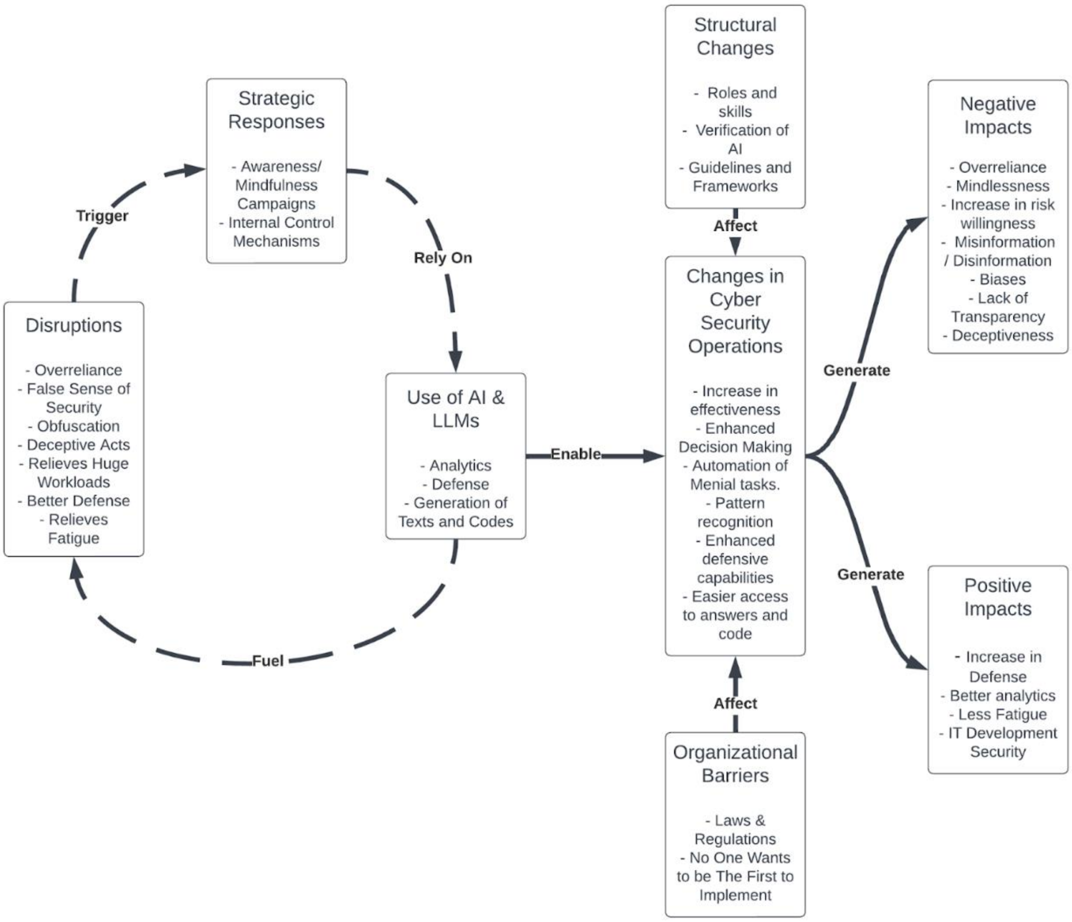

Strategic Responses
- Awareness/ Mindfulness Campaigns
- Internal Control Mechanisms
Trigger
Rely On
Disruptions
- Overreliance
- False Sense of Security
- Obfuscation
- Deceptive Acts
- Relieves Huge Workloads
- Better Defense
- Relieves Fatigue
Use of AI & LLMs
- Analytics
- Defense
- Generation of Texts and Codes
Fuel
Enable
Structural Changes
- Roles and skills
- Verification of AI
- Guidelines and Frameworks
Affect
Changes in Cyber Security Operations
- Increase in effectiveness
- Enhanced Decision Making
- Automation of Menial tasks.
- Pattern recognition
- Enhanced defensive capabilities
- Easier access to answers and code
Affect
Organizational Barriers
- Laws & Regulations
- No One Wants to be The First to Implement
Generate
Negative Impacts
- Overreliance
- Mindlessness
- Increase in risk willingness
- Misinformation / Disinformation
- Biases
- Lack of Transparency
- Deceptiveness
Generate
Positive Impacts
- Increase in Defense
- Better analytics
- Less Fatigue
- IT Development Security

Figure 10 Implementation of AI & LLMs

Figure 10 is based on Vial's model of the digital transformation process (Vial, 2019, p.5). This model is the culmination of the findings in both the literature review and the interviews that have been conducted. The use of AIs and LLMs fuels disruptions in the field, these disruptions trigger the strategic responses that an organization has (or should have), these however circles back and relies on the use of AIs. In addition, the use of AI and LLMs also enable certain changes in cybersecurity operations, these changes are affected by both the organizational barriers and what structural changes need to be added to support these changes. For example, an organization can not implement AIs if it is illegal or breaches policies and regulations. All of this leads to the generation of both positive- and negative impacts. The different factors shown in Figure 10 include several subsections that are related towards that specific factor. Each subsection would be discussed further down in this chapter, which is related towards the literature and the interview findings.

From this model it can be distilled what both the positive and negative impacts of the implementation of AIs and LLMs are. Like any piece of technology, it has both its positives and its negatives associated with it. This is both reflected in the literature where the AI could be 'tricked' by being targeted by adversaries to make the wrong predictions by what is called poisoning attacks (Li, 2018, p.1463; Morales-Forero, Bassetto & Coatanea, 2022, p.6; Mirsky et al., 2022, p.1; Thuraisingham, 2020, p.1116; Benzaid & Taleb, 2020, p.143). However, according to our study it would be difficult to perform source code poison attacks because the amount of data the AIs are trained on are absurd but that there were already examples of it happening. So, while it may be difficult to perform such attacks it seems to be feasible.

There was a consensus among both the literature and the informants regarding AI transparency, i.e., where it gets its information from, and obfuscation which is simply the LLM hiding information (Cunneen, Mullins & Murphy, 2020, p.625). This is a huge issue for the general level of mindfulness in any organization, not just HROs. It was explained that there were a lot of chatbot AIs that have popped up that can spread Russian misinformation for example, and if the employees in critical infrastructure were to base their decisions on that misinformation it might lead to severe consequences.

This is also a great example of how an AI can make you mindless, by essentially making you over reliant on it. In Chapter 2.2.2 overreliance was explained simply as you accepting erroneous information from the AI because you are unable to determine how much you should trust it (Passi & Vorvoreanu, 2022, p. 2). The risks with this were that you essentially rely on something mindless to make decisions, which in turn lowers your own mindfulness. The informants agreed with the literature where they talked about the dangers of blindly trusting AI but made sure that AIs are not the decision makers but provide support for the decision makers. They also highlighted that AIs are equivalent to a 'sociopathic super autist' completely without emotions and not understanding of social queues, so trusting it too much could lead to problems. It also highlighted another aspect that showed in the literature, they stated that LLMs gave the illusion of talking to a sentient human.

In the literature in Chapter 2.2.1 the concept of anthropomorphism was discussed. Anthropomorphism is how human something seems, the more human it seems the more you are likely to trust it and lower your uncertainty (Hoeffler, 2003; Castano and Giner-Sorolla, 2006, referenced in Esfahani, Reynolds & Asleigh, 2020, p.1). Thus, the more human an AI is the more prone we as humans are to fall into overreliance on it, and as discussed earlier, overreliance leads to mindlessness.

However, LLMs can also be used to make the employees more mindful. It was brought up that it could limit soc-fatigue as explained in Chapter 4.1 and could automate the more mundane and menial tasks that an employee could be burnt out by doing, thus keeping their head more in the game.

Transparency and obfuscation are closely linked with another concept this thesis brought up in the Introduction namely AI Hallucination. In Chapter 1.3 this topic was explained as essentially the AI 'lying' or contradicting itself, this is called extrinsic- and intrinsic hallucination (Bang et al., 2023, p.17-19). We tested this ourselves and asked ChatGPT to write a small article regarding mindfulness and provide references. It gave us ~30 of them and none of them existed at all.

The informants expressed concern that some employees may just swallow the information the AIs give raw and explained that you always had to be on your guard regarding AIs, no AI is perfect and if it is correct 99/100 times then you can possibly miss or just still assume it is correct the one time it is wrong. Most of these problems boil down to the fact that there is no transparency in the AIs, how did it come to its conclusions? That is the trap many seem to fall into. It was stated that if it does not have traceability, then the AI is useless, no matter how good it seems.

One additional problem with this is that it was explained that 90% of content on the internet would be generated by AIs by 2027. Without any transparency this would be highly problematic, especially since the AIs hallucinate and then suddenly 90% of all content on the internet is questionable at best in its veracity. This could be solved with a more mindful approach where you take 2 steps back and use your head thus creating your own interpretations, which match the literature (Langer & Imber, 1980; Langer & Moldoveanu, 2000, referenced in Dernbecher & Beck, 2017, p.122).

What needs to be added into the process is some strategic responses as seen in Figure 10, where it was put an emphasis on mindfulness training along with internal control mechanisms in order to verify the output of the AI and LLM but also keeping the employees more aware of the dangers of trusting the AI.

In Chapter 2.2.2 Figure 2 was brought up on how you continuously cultivate the trust of an AI. Two of the factors brought up are Reliability and Communication (Siau & Wang, 2018, p.51). After having analyzed other literature and the interview transcripts we can see that with the lack of transparency, hallucination and the margin of error in all AIs are detrimental to the factors highlighted in the model, thus perhaps not cultivating the right kind of trust. Figure 2 also brings up the factor of Bonding (Siau & Wang, 2018, p.51). Bonding can be seen in tandem with trust in general and with anthropomorphism which as explained earlier, an AI with human-like features creates a stronger bond and better trust with the human that operates it. In that sense anthropomorphism cultivates the right kind of trust, but if we see this in relation to the 2 previous factors brought up, then it would seem that AI assistants and LLMs then are detrimental to the overall trust in the organization.

As explained earlier, AIs do have many positives as well. Both the literature and the interviews agreed that organizations really do need AIs, especially when it comes to analyzing Big Data, intrusion detection etc. (Zhang, 2022, p.1031-1037; Wazid, Das, Chamola & Park, 2022, p.314; Sayan, Hariri & Ball, 2017, p.313-314). It was corroborated that a good AI model could eat through all the data and find patterns a normal human would never find. In addition, they also saw its positives in scanning through medical scans to find cancer or other illnesses, which according to the informants it did far more accurately than a human. The informants also brought up the dilemma of what happens if the doctor and AI disagree. From the literature we can gauge that the AI might be correct, but it needs to be looked at critically as it could easily be wrong as well. It is important that the personnel using it does not simply say ‘it has always been correct, so it is correct now as well’. That would lead to overreliance and mindlessness within the organization.

The CEO of OpenAI said that it would be a mistake to trust ChatGPT for any big decisions (Altman, 2022, referenced in Bang et al., 2023, p.1). This is a sentiment that is reflected in the interviews as most stated that it could be used as a support tool, but it could not make decisions yet.

Because we interviewed different organizations given our random sampling, they inevitably fall under different sectors, as shown in Table 6. What can we then try to see if there were any different viewpoints? There are 5 different sectors represented in our study: Defense & National Security, Energy, Transport, Food & Grocery and Health.

The health sector said that the AI were repeatedly more accurate than a human and seemed to be slightly more in favor of utilizing more AIs, however the Defense sector were more reserved and could not take the risk of implementing more even if the AI only had a margin of error of 1%, because that would be irresponsible of them. The transport sector also thought as the health sector did that there is no real reason not to trust AIs. The energy- and defense sectors were both in agreement that the usage of AIs were critical in analyzing Big Data.

The transport sector was under the impression that LLMs were something that they did not use at work, at all. Whereas the defense sector thought that LLMs had their use now already, such as alleviating menial and mundane tasks.

Every sector agrees that LLMs and AIs in general are here to stay, and they are positive that one day it will be more commonly implemented. However, for it to get to that stage, there needs to be further improvements made to it, especially regarding the transparency of how it reaches its decisions.

6 CONCLUSION

After having analyzed both the literature and the interviews conducted in this thesis we can conclude that the usage of AI Assistants such as LLMs in HROs is problematic. It lacks any sort of transparency and in a sector such as critical infrastructure there is a need for transparency because their work is obviously important. It is also the fact that LLMs is prone to hallucination and has, as discussed, a large chance to cause overreliance on it, thus relying on false information and the employees acting mindlessly. Thus the conclusion is that while AI Assistants may provide some benefits in regards to menial and mundane tasks and pattern recognition, it is detrimental, at least by using tools like LLMs, on the level of mindfulness of the individuals in HROs. The literature and the interviews all point to the fact that it is far too immature and prone to erroneous statements to be able to be used in critical infrastructure.

The practical implications for this study is that an organization, especially if it is of high reliability should carefully identify measures to avoid the negative impact of AI Assistants when used in day-to-day work in cybersecurity operations.

6.1 Limitations & Future Work

This section will explain the limitations regarding the thesis and some future research directions to further develop a better understanding of the relationship between impact of the individual mindfulness and using AI. The theme of this thesis is mainly based on AI, which has previously been a restricted technology, only being used by professionals. At the time of writing this thesis, this technology has been newly made accessible among the general population, enabling people to use it regardless of their skills and knowledge within the field of AI. This accessibility can provide benefits and consequences that are both known and unknown. For that reason, there aren't a lot of research articles that have been published within this field of research that could provide valuable information towards our research problem. This limitation however can also be seen as an opportunity as the contribution towards the field of research can be seen as valuable and important to the research community. The big difference between the current AI that is available to the general public and the AI in the past is the advanced technology behind it. The current AI technology which in many cases can implement different tasks as for example: writing a poem, summarizing large quantities of text, making complex presentations or creating code that is based on the users requirements. These functions can in many cases be seen as a large leap within the IT technology which is rapidly becoming more advanced. As seen from that point of view, we can better understand the importance of researching this current topic.

Another limitation towards this thesis is the conduction of the interviews. As we lack experience to plan, conduct and analyze interviews which have had some impact on this thesis. This limitation has been mitigated by exercising this method last semester and using additional information to further improve the planning phase of the interviews. It is also important to mention that the interviews that were conducted last semester were directed towards a much smaller assignment, which the master thesis is more comprehensive than comparatively. The master thesis also includes a more comprehensive research report rather than the research design itself, which was only included in last semester's assignment.

This leads us to the final limitation, which is the report itself. This is the first time we both have written and completed a thesis, where elements like scope, literature analysis, research design, conducting research method, analysis and findings are included with a comprehensive mindset. We have written reports which include parts and every element mentioned above, but on a smaller scale. The semester has also included meetings with a specific supervisor, which has given us continuous feedback on every section of the report. This iterative process has helped to mitigate this limitation.

The amount of research done in this thesis establishes a foundation of some aspects towards AI and mindfulness within a cybersecurity context. As this area of research is both new and unknown, it can further establish new directions in future research. LLMs ha recently become available to the general public, which is considered new and untested. We will recommend new research within the problem areas we have investigated, when AI will become more mature within the business sector.

REFERENCES

Alkaissi, H. & McFarlane, S.I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus 15(2): e35179. DOI 10.7759/cureus.35179

Ansari, M. F. (2022). A Quantitative Study of Risk Scores and the Effectiveness of AI-Based Cybersecurity Awareness Training Programs. International Journal of Smart Sensor and Adhoc Network.. 1-8. 10.47893/IJSSAN.2022.1212.

Anyan, F. (2013). The Influence of Power Shifts in Data Collection and Analysis Stages : A Focus on Qualitative Research Interview. The Qualitative Report, 18(18), 1-9. https://doi.org/10.46743/2160-3715/2013.1525

Araujo, T. (2018) Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions, *Computers in Human Behavior*, Volume 85, 2018, Pages 183-189, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2018.03.051.

Athaluri, S.A., Manthena, S.V., Kesapragada, V.S.R.K.M., Yarlagadda, V., Dave, T., Duddumpud, R.T.S., (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. Cureus 15(4): e37432. DOI 10.7759/cureus.37432

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... Fung, P., (2023). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. Pages 1-52, https://doi.org/10.48550/arXiv.2302.04023

Benzaïd, C. & Taleb, T. (2020) "AI for Beyond 5G Networks: A Cyber-Security Defense or Offense Enabler?," in IEEE Network, vol. 34, no. 6, pp. 140-147, November/December 2020, doi: 10.1109/MNET.011.2000088.

Buchanan, B. (2020). *A National Security Research Agenda for Cybersecurity and Artificial Intelligence*. Center for Security and Emerging Technology. Retrieved from: https://cset.georgetown.edu/wp-content/uploads/CSET-A-National-Security-Research-Agenda-for-Cybersecurity-and-Artificial-Intelligence.pdf

Burns, A.J. (2019). Security Organizing: A Framework for Organizational Information Security Mindfulness. SIGMIS Database 50, 4 (November 2019), 14–27. https://doi.org/10.1145/3371041.3371044

Canbek, N.G., & Mutlu, M. E. (2016). On the track of Artificial Intelligence: Learning with Intelligent Personal Assistants. Journal of Human Sciences, 13(1), 592–601. Retrieved from https://www.j-human-sciences.com/ojs/index.php/IJHS/article/view/3549

Cunneen, M., Mullins, M. & Murphy, F. (2020). Artificial intelligence assistants and risk: framing a connectivity risk narrative. AI & Soc 35, 625–634. https://doi.org/10.1007/s00146-019-00916-9

Curtis, N., (2023). To ChatGPT or not to ChatGPT? The Impact of Artificial Intelligence on Academic Publishing. *The Pediatric Infectious Disease Journal* Volume 42, Page 275. https://journals.lww.com/pidj/_layouts/15/oaks.journals/downloadpdf.aspx?an=00006454-202304000-00001&casa_token=yFhaRKQaxcMAAAAA:2PF7P8qksXNAjZtoSGKHOK3Dc_cMsjTTj3r76t501XKXNWiuKT0VehC7dhNzT1OPoq3oXLETtoesbAT_po1JiMMA

Dernbecher, S., & Beck, R. (2017) The concept of mindfulness in information systems research: a multi-dimensional analysis, *European Journal of Information Systems*, 26:2, 121-142, DOI: 10.1057/s41303-016-0032-z

Esfahani, M.S., Reynolds, N., & Ashleigh, M. (2020). Mindful and Mindless Anthropomorphism: How to Facilitate Consumer Comprehension Towards New Products, *International Journal of Innovation and Technology Management* Vol. 17, No. 03, 2050016 (2020). https://doi.org/10.1142/S0219877020500169

Falk, J. (2023, February 2nd). ChatGPT foreslo Anders Behring Breivik som «norsk helt». *VG*. Retrieved from: https://www.vg.no/nyheter/i/WRkK5K/chatgpt-foreslo-anders-behring-breivik-som-helt

Gioia, D,A., Corley, K,G., Hamilton, A,L., (2012). Seeking Qualitative Rigor in Inductive Research : Notes on the Gioia Methodology. *Organizational Research Methods,* 2013. Pages 15-31. DOI: 10.1177/1094428112452151

Gratch. J., Nathanael J. Fast. J. N. (2022). The power to harm: AI assistants pave the way to unethical behavior. Current Opinion in Psychology. Volume 47. 101382. ISSN 2352-250X. https://doi.org/10.1016/j.copsyc.2022.101382.

Hartmann K., & Steup, C. Hacking the AI - the Next Generation of Hijacked Systems. 2020 12th International Conference on Cyber Conflict (CyCon), Estonia, 2020, pp. 327-349, doi: 10.23919/CyCon49761.2020.9131724.

Hu, K. (2023, February 2nd). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. Retrieved from: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Huntsman Security (Unknown). *Critical Infrastructure Sectors* [Figure]. Retrieved from: https://www.huntsmansecurity.com/industries/critical-infrastructure/

Hunziker, S., Blankenagel, M. (2021). Multiple Case Research Design. In: Research Design in Business and Management. Springer Gabler, Wiesbaden. https://doi.org/10.1007/978-3-658-34357-6_9

Jensen, M.J., Dinger, M., Wright R.T., & Thatcher, J.B. (2017) Training to Mitigate Phishing Attacks Using Mindfulness Techniques, Journal of Management Information Systems, 34:2, 597-626, DOI: 10.1080/07421222.2017.1334499

Johannessen, A. Tufte, P.A. Christoffersen, L. (5.edition: 2016). Introduksjon til samfunnsvitenskapelig metode. Abstrakt forlag: Oslo.

Johannesson, P., & Perjons, E. (2014). Research strategies and methods. In An introduction to design science (pp. 39-73). Springer, Cham. https://doi.org/10.1007/978-3-319-10632-8_3

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Kasneci, G..(2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, Volume 103, 102274, ISSN 1041-6080, https://doi.org/10.1016/j.lindif.2023.102274.

Kelley, K., Clark, B., Brown, V., Sitzia, J. Good practice in the conduct and reporting of survey research, International Journal for Quality in Health Care, Volume 15, Issue 3, May 2003, Pages 261–266, https://doi.org/10.1093/intqhc/mzg031

Kruger, H.A., Kearney, W.D.. A prototype for assessing information security awareness, Computers & Security, Volume 25, Issue 4, 2006, Pages 289-296, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2006.02.008

Li, J.H. (2018). Cyber security meets artificial intelligence: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1462–1474. https://doi.org/10.1631/FITEE.1800573

Lowhorn, G. L. (2007, May). Qualitative and quantitative research: How to choose the best design. In Academic Business World International Conference. Nashville, Tennessee. Available at SSRN: https://ssrn.com/abstract=2235986

Mirsky. Y., Demontis. A., Kotak. J., Shankar. R., Gelei. D., Yang. L., Zhang. X., Pintor. M., Lee. W., Elovici. Y., Biggio. B. (2023). The Threat of Offensive AI to Organizations. Computers & Security. Volume 124. 103006. ISSN 0167-4048. https://doi.org/10.1016/j.cose.2022.103006.

Morales-Forero, A., Bassetto, S. & Coatanea, E. (2022). Toward safe AI. *AI & Society*. https://doi.org/10.1007/s00146-022-01591-z

Passi. S., Vorvoreanu. M. (2022). Overreliance on AI: Literature review. Retrieved From: https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf

Puthal, D. & Mohanty, S.P. "Cybersecurity Issues in AI," in IEEE Consumer Electronics Magazine, vol. 10, no. 4, pp. 33-35, 1 July 2021, doi: 10.1109/MCE.2021.3066828.

Rev. (2022, March 30th). How to Analyze Interview Transcripts in Qualitative Research. Retrieved from: https://www.rev.com/blog/transcription-blog/analyze-interview-transcripts-in-qualitative-research

Sayan, C.M., Hariri, S., & Ball, G.. "Cyber Security Assistant: Design Overview," 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), Tucson, AZ, USA, 2017, pp. 313-317, doi: 10.1109/FAS-W.2017.165.

Schlienger, T., & S. Teufel, S.. Analyzing information security culture: increased trust by an appropriate information security culture. *14th International Workshop on Database and Expert Systems Applications*, 2003. Proceedings., 2003, pp. 405-409, doi: 10.1109/DEXA.2003.1232055.

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. Cutter business technology journal, 31(2), 47-53.

Siau, K., & Wang, W. (2018). *Technology features of AI that affect trust building [Figure]*. Retrieved from: https://www.researchgate.net/profile/Keng-Siau-2/publication/324006061_Building_Trust_in_Artificial_Intelligence_Machine_Learning_and_Robotics/links/5ab8744baca2722b97cf9d33/Building-Trust-in-Artificial-Intelligence-Machine-Learning-and-Robotics.pdf

Siponen, M.T. (2000), "A conceptual foundation for organizational information security awareness", Information Management & Computer Security, Vol. 8 No. 1, pp. 31-41. https://doi.org/10.1108/09685220010371394

Thatcher, J,B., Wright, R,T., Sun, H., Zagenczyk, T,J., Klein, R. (2018). MINDFULNESS IN INFORMATION TECHNOLOGY USE: DEFINITIONS, DISTINCTIONS, AND A NEW MEASURE, *MIS Quarterly* Vol. 42 No. 3, pp. 831-847/September 2018, DOI: 10.25300/MISQ/2018/11881

Thuraisingham, B. (2020). The Role of Artificial Intelligence and Cyber Security for Social Media. *IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. New Orleans. LA. USA. pp. 1-3. doi: 10.1109/IPDPSW50202.2020.00184

Timmers, P. (2019). Ethics of AI and Cybersecurity When Sovereignty is at Stake. Minds & Machines 29, 635–645. https://doi.org/10.1007/s11023-019-09508-4

Torp, S, I. (2019, July 8). General guidelines. Retrieved from: https://www.forskningsetikk.no/en/guidelines/general-guidelines/

Trim, P. R. J., & Lee, Y.-I. (2022). Combining Sociocultural Intelligence with Artificial Intelligence to Increase Organizational Cyber Security Provision through Enhanced Resilience. Big Data and Cognitive Computing, 6(4), 110. MDPI AG. Retrieved from http://dx.doi.org/10.3390/bdcc6040110

Vanian, J. (2023, May 22nd). Bill Gates says A.I. could kill Google Search and Amazon as we know them. *CNBC*. Retrieved from: https://www.cnbc.com/2023/05/22/bill-gates-predicts-the-big-winner-in-ai-smart-assistants.html

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *Journal of Strategic Information Systems.* Pages 1-27. https://doi.org/10.1016/j.jsis.2019.01.003

Waardenburg, L., & Huysman, M.. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization*. Pages 1-11. https://doi.org/10.1016/j.infoandorg.2022.100432

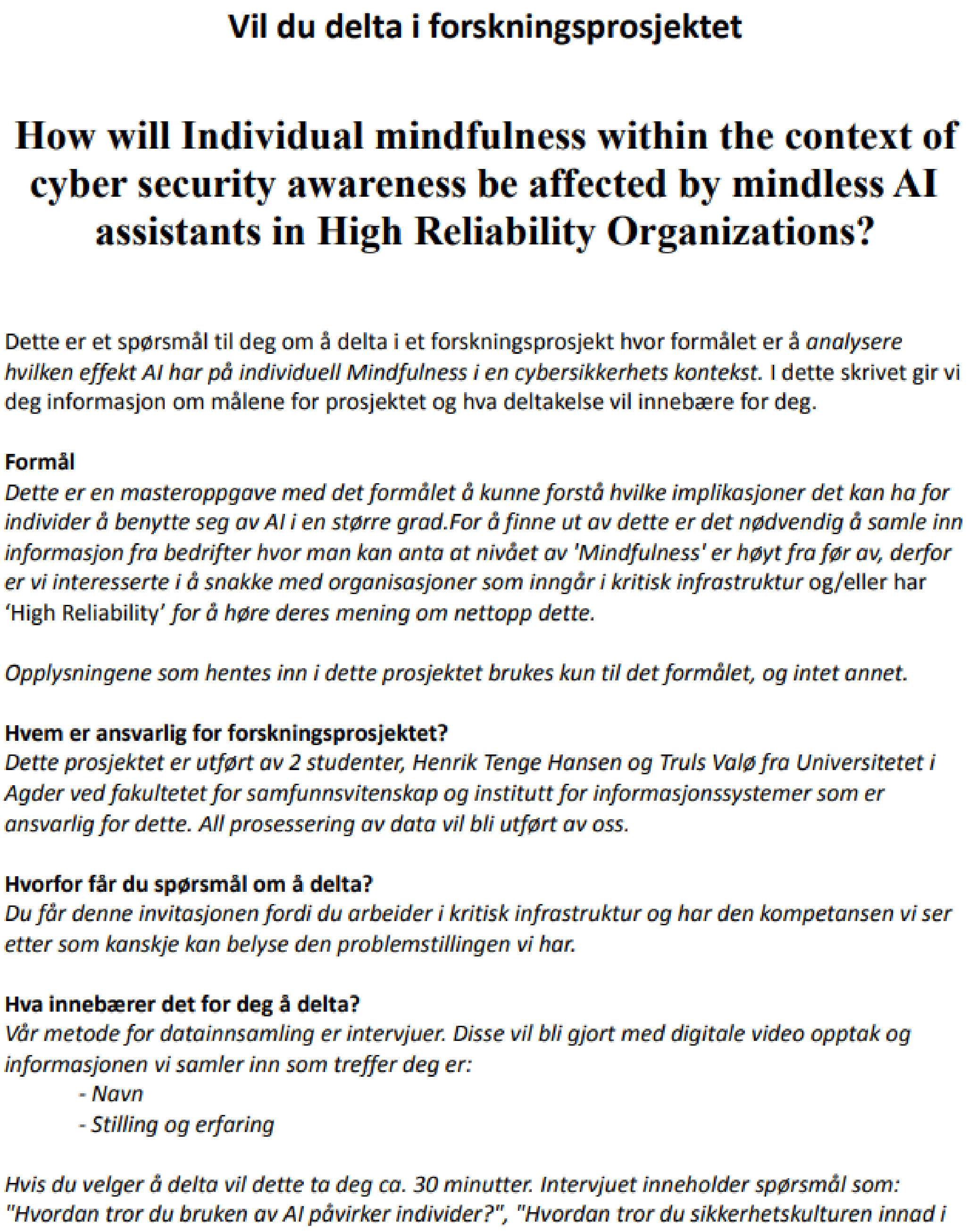
Wazid. M., Das. A. K., Chamola. V., Park. Y. (2022). Uniting cyber security and machine learning: Advantages, challenges and future research. Volume 8, Issue 3. Pages 313-321. ISSN 2405-9595. https://doi.org/10.1016/j.icte.2022.04.007.

Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. Journal of Planning Education and Research, 39(1), 93–112. https://doi.org/10.1177/0739456X17723971

Zhang, Z., Ning, H., Shi, F. et al. (2022). Artificial intelligence in cyber security: research advances, challenges, and opportunities. *Artificial Intelligence Review* 55, 1029–1053. https://doi.org/10.1007/s10462-021-09976-0

Zhu, F., Carpenter, S., Kolimi, S., Mindlessness attacks, *Procedia Manufacturing*, Volume 3, 2015, Pages 1066-1073, ISSN 2351-9789, https://doi.org/10.1016/j.promfg.2015.07.174.

APPENDIX A CONSENT FORM



Vil du delta i forskningsprosjektet

How will Individual mindfulness within the context of cyber security awareness be affected by mindless AI assistants in High Reliability Organizations?

Dette er et spørsmål til deg om å delta i et forskningsprosjekt hvor formålet er å *analysere hvilken effekt AI har på individuell Mindfulness i en cybersikkerhets kontekst.* I dette skrivet gir vi deg informasjon om målene for prosjektet og hva deltakelse vil innebære for deg.

Formål
Dette er en masteroppgave med det formålet å kunne forstå hvilke implikasjoner det kan ha for individer å benytte seg av AI i en større grad.For å finne ut av dette er det nødvendig å samle inn informasjon fra bedrifter hvor man kan anta at nivået av 'Mindfulness' er høyt fra før av, derfor er vi interesserte i å snakke med organisasjoner som inngår i kritisk infrastruktur og/eller har 'High Reliability' *for å høre deres mening om nettopp dette.*

Opplysningene som hentes inn i dette prosjektet brukes kun til det formålet, og intet annet.

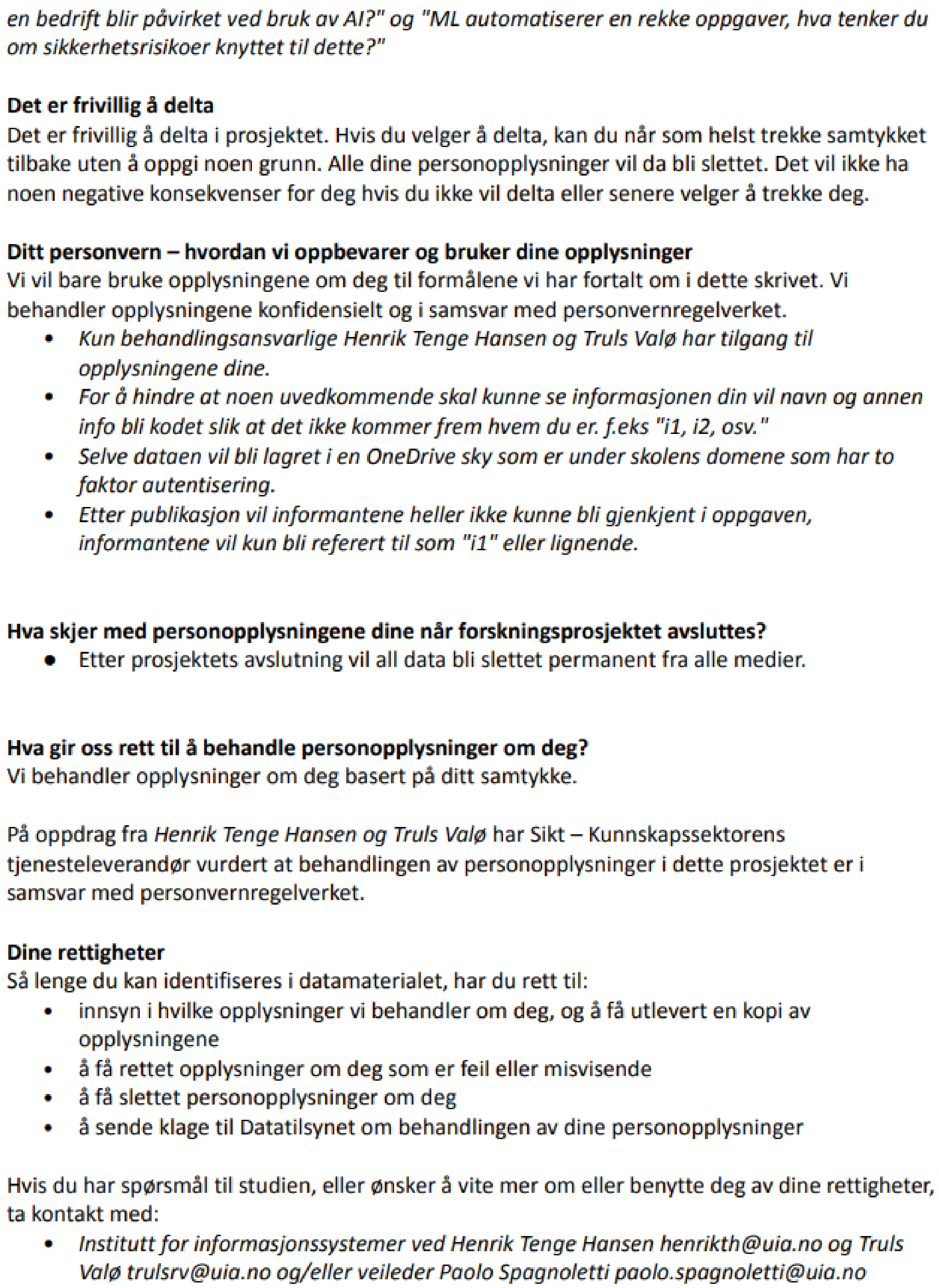
Hvem er ansvarlig for forskningsprosjektet?
Dette prosjektet er utført av 2 studenter, Henrik Tenge Hansen og Truls Valø fra Universitetet i Agder ved fakultetet for samfunnsvitenskap og institutt for informasjonssystemer som er ansvarlig for dette. All prosessering av data vil bli utført av oss.

Hvorfor får du spørsmål om å delta?
Du får denne invitasjonen fordi du arbeider i kritisk infrastruktur og har den kompetansen vi ser etter som kanskje kan belyse den problemstillingen vi har.

Hva innebærer det for deg å delta?
Vår metode for datainnsamling er intervjuer. Disse vil bli gjort med digitale video opptak og informasjonen vi samler inn som treffer deg er:

- *Navn*
- *Stilling og erfaring*

Hvis du velger å delta vil dette ta deg ca. 30 minutter. Intervjuet inneholder spørsmål som: "Hvordan tror du bruken av AI påvirker individer?", "Hvordan tror du sikkerhetskulturen innad i



en bedrift blir påvirket ved bruk av AI?" og "ML automatiserer en rekke oppgaver, hva tenker du om sikkerhetsrisikoer knyttet til dette?"

Det er frivillig å delta

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

Ditt personvern – hvordan vi oppbevarer og bruker dine opplysninger

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

- *Kun behandlingsansvarlige Henrik Tenge Hansen og Truls Valø har tilgang til opplysningene dine.*
- *For å hindre at noen uvedkommende skal kunne se informasjonen din vil navn og annen info bli kodet slik at det ikke kommer frem hvem du er. f.eks "i1, i2, osv."*
- *Selve dataen vil bli lagret i en OneDrive sky som er under skolens domene som har to faktor autentisering.*
- *Etter publikasjon vil informantene heller ikke kunne bli gjenkjent i oppgaven, informantene vil kun bli referert til som "i1" eller lignende.*

Hva skjer med personopplysningene dine når forskningsprosjektet avsluttes?

- Etter prosjektets avslutning vil all data bli slettet permanent fra alle medier.

Hva gir oss rett til å behandle personopplysninger om deg?

Vi behandler opplysninger om deg basert på ditt samtykke.

På oppdrag fra *Henrik Tenge Hansen og Truls Valø* har Sikt – Kunnskapssektorens tjenesteleverandør vurdert at behandlingen av personopplysninger i dette prosjektet er i samsvar med personvernregelverket.

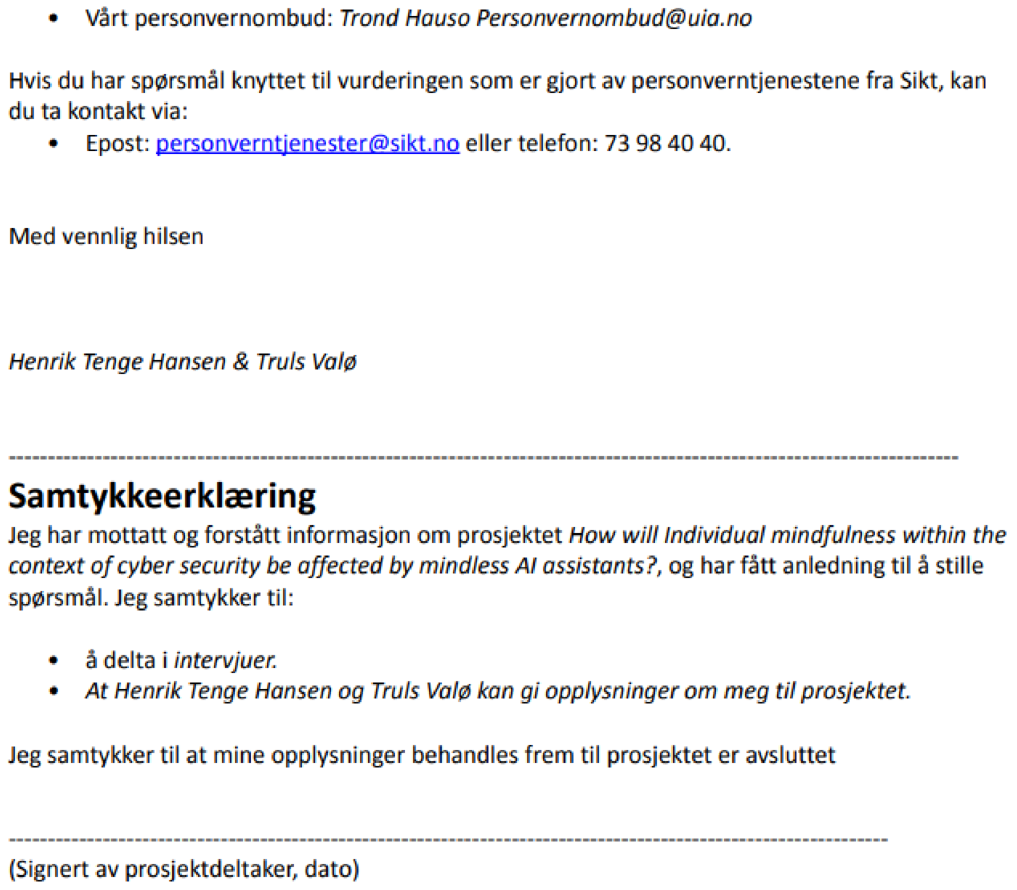
Dine rettigheter

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

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

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

- *Institutt for informasjonssystemer ved Henrik Tenge Hansen henrikth@uia.no og Truls Valø trulsrv@uia.no og/eller veileder Paolo Spagnoletti paolo.spagnoletti@uia.no*



- Vårt personvernombud: *Trond Hauso Personvernombud@uia.no*

Hvis du har spørsmål knyttet til vurderingen som er gjort av personverntjenestene fra Sikt, kan du ta kontakt via:

- Epost: personverntjenester@sikt.no eller telefon: 73 98 40 40.

Med vennlig hilsen

Henrik Tenge Hansen & Truls Valø

Samtykkeerklæring

Jeg har mottatt og forstått informasjon om prosjektet *How will Individual mindfulness within the context of cyber security be affected by mindless AI assistants?*, og har fått anledning til å stille spørsmål. Jeg samtykker til:

- å delta i *intervjuer.*
- *At Henrik Tenge Hansen og Truls Valø kan gi opplysninger om meg til prosjektet.*

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

----

(Signert av prosjektdeltaker, dato)

APPENDIX B INTERVIEW QUESTIONS

Introduksjonsspørsmål:

- Kan du fortelle litt om deg selv og din arbeidserfaring?

Hovedspørsmål

Mindfulness

- Hvordan har utviklingen til ansattes bevissthet vært de siste årene?
  - *Hva tror du det skyldes?*
  - *Er det noe dere gjør som er forskjellig fra tidligere?*
  - *Har denne utviklingen hatt en påvirkning på måten dere opererer?*

AI

- Benytter dere dere av AI/ML i sikkerhetsoperasjoner? Og i så fall, hvor lenge har dere gjort det?
  - *Hva tenker du om farene knyttet til AI?*
  - *Er det en reell fare for å stole på for mye på AI? Isåfall på hvilken måte?*
  - *Hvilken effekt tror du bruken av AI har på den ansatte?*
- Har AI en påvirkningskraft på måten man opererer sikkerhet? Nåtiden og fremtiden

Awareness/Bevissthet

- Hvordan har de ansatte forholdt seg til angrep de siste årene? Er det noe som har overrasket deg, både positivt og negativt?
  - *Har det som har overrasket deg positivt gitt noen endringer i måten dere håndterer nye hendelser?*
  - *Har du sett noen bevissthetsendringer innenfor sikkerheten ettersom trusselen er stadig økende?*
  - *Hvordan har læringsevnen innenfor informasjonssikkerhet til de ansatte utviklet seg gjennom de siste årene?*
- Har sikkerhetsbevissthet og læringsevnen noen påvirkningskraft på måten man opererer et AI-verktøy?

Avslutning

- Blir ansattes holdning til sikkerhet bedre/dårligere av å benytte AI og hvorfor?