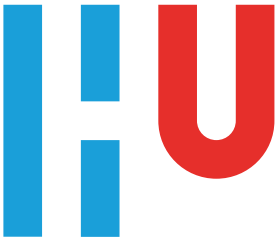
HU University of Applied Sciences

Master of Informatics



HOGESCHOOL UTRECHT

MASTER'S THESIS

Acceptance of AI-driven Autonomous Interventions in Cyber Security Defense

"I'm sorry, Dave. I'm afraid I can't explain that."

Author:
Lennard van der Linden

Supervisors:
Peter Westerveld, Christian-Alexander Bunge

Date:
January 6, 2026

HU University of Applied Sciences
P.O. box 182
3500 AD UTRECHT
The Netherlands

I Abstract

AI, especially generative AI, will enable threat actors to scale up their cyber-attacks. Threat actors will exploit new vulnerabilities much quicker in an automated fashion and speed up the cyber security kill chain. To manage the growing volume of attack and the attack speed, cyber security defense teams will need to automate their defenses and have their AI-powered cyber security tooling respond autonomously.

This study used the Delphi method to research under which conditions financial institutions in the Netherlands accept AI tooling to act autonomously, without a human approving the decisions of AI.

Potential factors associated with acceptance were extracted from the literature. These factors were verified by interviewing a CISO of a Dutch financial institution. Via a mixed-method three round Delphi study the factors that are associated with acceptance of autonomous AI tooling were determined. Eight (chief) information security officers from Dutch financial institutions took part in the Delphi study. The goal of a Delphi study is to reach consensus on statements, usually statements about the future.

This study shows that autonomous AI-driven cyber security tooling will be needed in the next three years to be able to withstand the growing scale of cyber-attacks. For cases with a risk of operational impact, a human *in* the loop is required, and accepted. AI tooling performing the investigation of major security incidents is accepted as well.

This study finds that when AI tooling acts autonomously, the actions of AI need to be understandable to humans. That is, the actions are explainable, and the workings of AI tooling need to be transparent. Another finding is that autonomous actions that could lead to operational impact are not accepted.

This study confirmed that the Mission Critical Technology Acceptance Model (MCTAM) by Weger et al. (2023) fits the cyber security context. As an extension to this conceptual model, the findings reveal that *reliability* influences *trust*.

This study had a small sample size: it was restricted to a small, homogenous group of experts from the financial sector in the Netherlands. The views of a small group may not be representative, which impairs generalization. However, this study does capture a quantified snapshot of the sentiment of the Dutch financial sector about autonomous AI tooling. Future research should investigate acceptance in different sectors or countries, or design guidelines for implementing AI powered tooling. Additionally, how opinions will shift in the next two to three years is an interesting topic for research.

Autonomous AI is not fully trusted yet, because the reliability and accuracy are not perfect. Conversely, the expectation is that autonomous AI will be needed in cyber security defense. Therefore, financial organizations' capability and trust need to grow: Start by implementing autonomous AI in low-risk and low-impact scenarios (Nyre-Yu et al., 2019). As participant 3 (P3) stated, have a "validation and testing process" in place. Build up the required skill and expertise and involve users early on (Cabitza et al., 2021; Hauptman et al., 2023; Solaimani & Swaak, 2023). As P5 suggested, require personnel to get certified. Then, every organization will need to start allowing autonomous AI in higher-impact scenarios, driven by a risk-based approach. Because after all, the threat actors will keep on coming and they will eventually figure out how to use the full capabilities of AI to attack any of us. This advice is based on an interpretation of the findings of this study and the current cyber security and AI landscape in practice, and therefore further research is required to make any scientific claims.

II Table of Contents

III List of Tables and Figures

III.1 Tables

III.2 Figures

IV List of abbreviations

Abbreviation	Explanation
AGI	Artificial General Intelligence
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
ASI	Artificial Super Intelligence
CISO	Chief Information Security Officer
CRA	Cyber Resilience Act
DORA	Digital Operational Resilience Act
GDPR	General Data Protection Regulation
GenAI	Generative AI
GPAI	General Purpose Artificial Intelligence
HitL	Human in the Loop
HotL	Human on the Loop
ICDDSS	Intelligent Clinical Diagnostic Decision Support System
IQR	Interquartile Range
IR	Incident Response
ISO	Information Security Officer
LLM	Large Language Model
MCTAM	Mission Critical Technology Acceptance Model
MMCTAM	Modified Mission Critical Technology Acceptance Model
MSSP	Managed Security Services Provider
NBFI	Non-Bank Financial Institution
NIS2	Network and Information Security Directive
OECD	Organisation for Economic Co-operation and Development
P	Participant
Q1	Lower Quartile (see IQR)
Q3	Upper Quartile (see IQR)
RQ	Research Question
SOAR	Security Orchestration, Automation and Response
SOC	Security Operations Center
SQ	Sub Question
TAM	Technology Acceptance Model

1 Introduction

1.1 Background

With the introduction of generative AI (GenAI) applications like ChatGPT in November 2022, cyber criminals and other types of threat actors in the cyber security space can use this technology to augment their cyber-attacks (OpenAI, 2022). AI-powered cyber-attacks were anticipated as early as 2018 (Brundage et al., 2018). The National Cyber Security Centre of the UK (2024) predicted the volume and impact of attacks to rise in 2025. How generative AI produces output is mostly a black box, although research is being done on making their results interpretable (Ameisen et al., 2025).

GenAI enables creating customized, one-off attacks easily, such as creating deepfakes for social engineering, impersonating websites, creating tailored phishing campaigns, assisting in writing exploits or malware, and rewriting malware to evade detection (Baltariu & Puscasu, 2024; Cato Networks, 2025; Check Point Research, 2025; Cuprik, 2025; ENISA, 2024; Heiding et al., 2024; Insikt Group, 2024; Madjar et al., 2024; Schafer, 2024; Tologonov & Fokker, 2024; Wise, 2023). Threat actors targeting the payment ecosystem continue to use AI technologies (Visa Payment Fraud Disruption, 2023). The following examples illustrate the potential usage of GenAI in cyber-attacks. Threat actors have augmented a phishing kit with a deep fake voice calling feature (Ushakov, 2024). An autonomous voice-enabled AI agent proof of concept has been demonstrated (Fang et al., 2024). In 2024, an AI-powered attack involved a suspected deepfake attack that resulted in a 25 million dollar loss for a British multinational (Magramo, 2024).

According to Trend Micro research in 2024, there was one criminal LLM: WormGPT. Other criminal AI tools were either scams or jailbreaks for regular LLMs (Ciancaglini & Sancho, 2024). Jail breaks bypass security controls enabling malicious usage of regular LLMs (Fire et al., 2025; Lin et al., 2024). After Trend Micro's research was published, WormGPT has been shut down and variants of WormGPT have emerged (Simonovich, 2025). Malicious LLM services that use regular LLMs as their backend included EscapeGPT, LoopGPT, BlackhatGPT, BadGPT, XXXGPT, Evil-GPT, GhostGPT and WolfGPT (Abnormal Security, 2025; Ciancaglini & Sancho, 2024; Lin et al., 2024).

Until recently, GenAI usage by threat actors was limited according to cyber security firms (Hofmans, 2024). However, a recent report from Anthropic shows that threat actors used Anthropic's family of LLMs, Claude. Threat actors used Claude to scale their operations, from developing and distributing ransomware, to masquerading as remote workers, to supporting thousands of romance scams (Moix et al., 2025). Legitimate AI tools have been abused in attacks as well, compromising the (AI) software supply chain (Gooding et al., 2025). Automated exploitation of vulnerabilities at scale saw a big boost via the Hexstrike-AI framework. In August 2025 Citrix Netscaler vulnerabilities were being scanned and exploited within 12 hours instead of weeks (Weigman, 2025).

To conclude, threat actors use GenAI as one more tool in their toolbox, but have not used it for novel techniques (Google Threat Intelligence Group, 2025; Microsoft Threat Intelligence, 2024). However, if automation as in the example of Hexstrike, becomes the norm, the speed and scale of attacks can increase dramatically within months.

AI tools introduce additional security risks. For example, an AI tool proposed to change the limits of its operation. Sandboxing such a tool could mitigate an AI tool from overstepping its boundaries (Sakana AI, 2024). As another example, AI tooling, like Microsoft Copilot, can be vulnerable to prompt injection leading to data exfiltration (Rehberger, 2024). Malicious AI tooling is vulnerable to this as well (Mayoral-Vilches & Rynning, 2025). Another example is that software used to run AI can introduce critical vulnerabilities into a company's IT landscape (Tamari et al., 2024). Design patterns can be implemented to protect against prompt injections (Beurer-Kellner et al., 2025).

AI, however, also can assist in preventing, detecting, and responding to cyber-attacks. Incident response root cause analysis for cloud environments can be enhanced using AI (Chen et al., 2024; D. Hsu et al., 2024). Generative AI can be used to discover vulnerabilities in software (Chang et al., 2024; Google Project Zero, 2024). The expectation in the cyber security industry and beyond is that AI is a necessity in defending against AI-powered attacks (Aubley et al., 2021; Bosch et al., 2023; Columbus, 2023; De Lucia et al., 2019; Lorenz et al., 2023; MIT Technology Review, 2021; Osterman Research, 2024; Ravichandran, 2023; Seetharaman et al., 2023; Xu et al., 2024). The capability of AI to perform more complex tasks is growing, doubling roughly every 7 month (Kwa et al., 2025). Kwa et al. included cyber security tasks in their study. Both defenders and attackers benefit from growing capabilities. There is no consensus between cyber security professionals who benefits more from AI-powered tools: defenders, attackers or neither (Baron et al., 2024; Ciancaglini & Sancho, 2024).

AI is used to boost productivity within and outside of the financial sector, which has led to job cuts at Klarna and UPS (BBC, 2024; Klarna, 2024). However, Klarna is an outlier in the financial sector (Criddle & Quinio, 2024). In a reverse of events, Klarna started hiring people again in 2025 (Ivanova, 2025). Replacing employees in customer service by an AI-powered voice bot failed after a month for an Australian bank (Belanger, 2025).

Autonomy is about who or what is in control. It ranges from a human being in full control to a machine being in full control. Autonomy is related to automation, where tasks are executed according to rules with less or no intervention (Sarker et al., 2024). In the case of autonomy a machine does not follow pre-defined actions but can set its own path (Mayer, 2023; Veitch et al., 2024). There is a distinction between human-in-the-loop (HitL) where humans take key decisions, and human-on-the-loop (HotL) where machines act autonomously while being supervised by humans (Fischer et al., 2021). Schneier (2017) argued that fully automating cyber security incident response is not possible; humans need to be *in* the loop. Paling (2025) stated that AI should augment the humans in the security operations center, but should not be fully autonomous. Having a human in the loop or not is a contentious topic.

GenAI is a recently introduced technology. Usage of GenAI in the financial sector has been limited. Several models describe technology acceptance, which will be used to investigate acceptance of GenAI. Weger et al. (2023) extended the well-known Technology Acceptance Model by Davis (1989) to include mission critical factors. They found that trust in mission critical environments is influenced by the factors reliability, safety/security, human in the loop and perceived usefulness, and transparency of how the system works.

Chapter 2 describes the concepts cyber security, AI, autonomy and technology acceptance in further detail.

1.2 Scope of This Study

This study focuses on financial institutions in the Netherlands. Organizations within the financial sector are relatively mature in cyber security. Financial institutions tend to have Security Operations Centers (SOCs) where dedicated personnel are preventing, detecting, and responding to cyber threats. The financial sector is frequently targeted by cyber-attacks (Check Point, 2023; ENISA, 2023) and is (expecting to be) targeted by more advanced threat actors (e.g., nation state actors). AI assisted tooling is an accepted solution in the Dutch financial sector (De Nederlandsche Bank, 2019) and usage of AI is expected to increase (De Nederlandsche Bank & Dutch Authority for the Financial Markets, 2024). Increased usage of AI can be a solution for the growing shortage of qualified cyber security personnel in the Netherlands and abroad (Coker, 2024; Cyber Security Raad, 2023; Ministerie van Economische Zaken en Klimaat, 2024).

1.3 Problem Statement

As AI technology develops, cyber security teams of organizations are forced to keep up with threat actors by augmenting their cyber security defenses with AI technology. Developing AI cyber security technology is not the core business of most organizations. Those organizations will use AI-powered products and services offered by cyber security firms. These (mostly) black box technologies assist in detecting and responding to perceived threats, boost productivity of Security Operation Centers, and provide tailored threat intel information and reporting (Bisen et al., 2023; Cassetto, 2024; Cyware, 2023; Darktrace, 2024; Microsoft, 2023; Potti, 2024; Tevet, 2024; The Hacker News, 2024). Responses to cyber threats can, depending on the action taken, lead to a temporary loss of availability of data and/or services. Not containing these cyber threats could lead to the loss of availability, but also to defraudment, a successful ransomware attack, or data leakage. Organizations face the dilemma of either trying to contain these threats manually (risking a larger impact on their operations) or accepting automated or autonomous responses to cyber threats to potentially cause impact on operations.

1.4 Research Question

This study *evaluated* the following: What factors are associated with acceptance of AI-driven autonomous interventions in cyber security defense by organizations within the financial sector in the Netherlands?

To illustrate, autonomous actions could range from deleting emails or isolating a laptop of a single employee, to taking an entire organization offline due to a (suspected) ransomware attack. Consider an autonomous agent that sets its own path and can perform similar actions as a human SOC analyst or cyber incident responder. Autonomous interventions can consist of a complex, multiple step plan being executed autonomously (Kott, 2023).

The research function of this study is to evaluate. The research function is limited to evaluating and not theory building, theory testing or designing: AI-powered security tooling that not only detects but also responds autonomously is state-of-the-art and the number of implementations is limited, and academic research is lagging behind the practical use of AI-powered security tooling in a business context. This study is about future developments and it aims to find the factors associated with acceptance of AI interventions, but the study design does not allow for any scientific claims about causation between those factors and acceptance. Note that the term 'factors' is used, as this study resembles a critical success factors study (Moeuf et al., 2020; Solaimani & Swaak, 2023).

1.5 Sub Questions

The sub questions (SQs) that support answering the research question (RQ) are listed in Table 1. Each sub question covers a different part of the research question: SQ1 encompasses technology acceptance of AI; SQ2 encompasses technology acceptance of autonomous response; SQ3 encompasses sectoral requirements for AI in cyber security; and finally, SQ4 combines the findings of the first three sub questions to provide insight in the RQ. With these sub questions, every aspect of the RQ is covered, see Figure 1.

Table 1

Sub questions

Number	Sub question	Research function	Method to answer the SQ
1	What factors are associated with acceptance of AI advising humans on which actions to take?	Define	Literature review + Interview
2	What factors are associated with acceptance of the outsourcing of cyber security response (to external Security Operation Centers)?	Define	Literature review + Interview
3	What requirements exist for AI usage within the Dutch financial sector?	Define	Literature review [a]
4	How do the identified factors relate to acceptance of autonomous interventions within the Dutch financial sector?	Evaluate	Delphi study

[a] The literature review for SQ3 consists of regulations, professional and scientific literature, as opposed to the literature review of SQ1 and SQ2 which consist entirely of scientific literature.

SQ1 resembles the research question. The key difference between the RQ and SQ1 is that in case of SQ1 a human is in the loop: the human has the final decision on performing an action or not. In case of the RQ, the human cedes direct control but monitors the decisions and can act on them after the fact.

SQ2 provides insight into factors related to organizations outsourcing their cyber security function, which results in organizations relinquishing control over cyber security response actions. Relinquishing control can have a negative impact on the information systems of the organization. Autonomous AI cyber security tooling also results in organizations relinquishing control over their response actions. We transpose the results of the former (i.e., outsourcing) onto the latter (i.e., autonomous AI). Therefor factors found via SQ2 can be used for the main RQ, where - instead of outsourcing and handing control over to a third party - control is handed over to AI tooling.

SQ1 and SQ2 are not scoped to the financial sector. The literature review for SQ1 and SQ2 was exploratory. The goal was to find the more common (and more pressing) factors relating to loss of control that have been researched, which apply to the financial sector. The findings from the literature review have been validated by interviewing one chief information security officer (CISO).

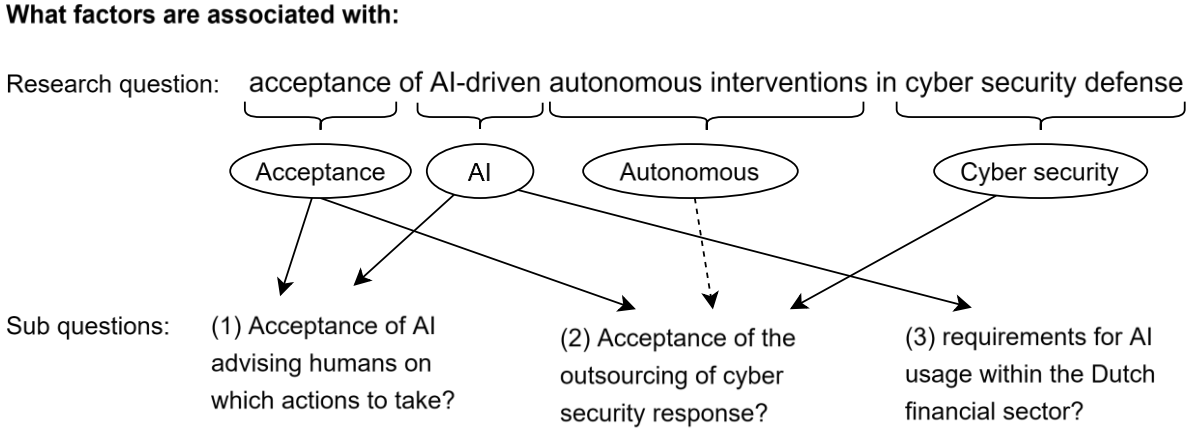
SQ3 aims to identify factors that are specific to the context of the financial sector, and not to cyber security defense. The financial sector is heavily regulated, which could lead to sector-specific factors to take into consideration when answering the main RQ. Several laws and regulations applicable to the Dutch financial sector have been or will be implemented between 2024 and 2027. Therefore, limited academic literature was available and the academic literature search was augmented by reviewing the laws and regulations themselves.

Figure 1 shows how the first three sub questions cover the concepts contained in the research question.

Finally, SQ4 applied the comprehensive set of factors identified by SQ1, SQ2 and SQ3 to the acceptance of autonomous AI tooling in a cyber security context within the financial sector in the Netherlands.

Figure 1

Research question coverage by Sub Questions 1 to 3



Note. The scoping of the research question to the Dutch financial sector has been left out for brevity.

The line from *Autonomous* to SQ2 is dashed to illustrate that *Autonomous* is being investigated indirectly.

2 Theoretical Background

2.1 Initial literature Review

An initial literature review was performed to establish the state of the art in academic research. The concepts and conceptual model, the research question and sub questions were constructed based on the initial literature review. See 3.1 for the research method including search and exclusion criteria.

2.2 Concepts

The following concepts are part of the research question. Each concept has been described based on the literature review. AI and autonomy are combined in one section as there is a significant overlap resulting from the phrasing of the search terms used in the literature review. The cyber security section includes the application of AI.

2.2.1 Autonomy and Artificial Intelligence

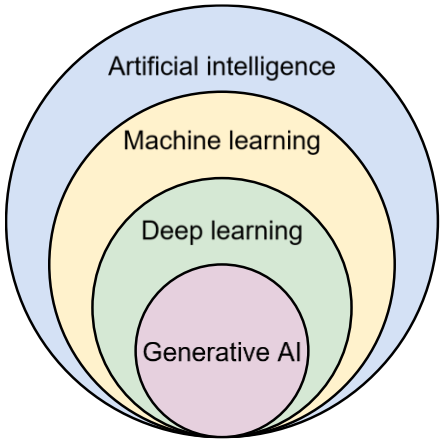
The field of artificial intelligence is broad and diverse. There is no single definition for AI (Veitch et al., 2024; Wang, 2019). Wang provided the following working definition for intelligence, which applies but is not limited to machines and computers: “Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources” (2019, p. 17). Although there is no consensus on the definition for AI, common aspects across definitions are perception of the environment, information processing, decision making, and achievement of specific goals. A distinction can be made between artificial narrow intelligence (ANI) and artificial general intelligence (AGI), where ANI is specialized in specific domains and currently in use, and AGI refers to systems that can perform at the same level as humans (Samoili et al., 2021). Generative AI (GenAI) is largely based on large language models (LLMs). The way GenAI produces output is mostly a black box, although making their results interpretable is an active area of research (Ameisen et al., 2025).

Artificial intelligence encompasses both machine learning and deep learning, see Figure 2. GenAI can be considered a subfield of deep learning, although it can be argued that GenAI partly overlaps machine learning as well (Sarker, 2024; Zhuhadar & Lytras, 2023). GenAI can create text, images, audio and video. Language models (LLMs) make up a large part of GenAI.

LLMs are known to hallucinate, which means they can generate information that is not factually correct. Retrieval augmentation generation, which entails the AI retrieving additional information from the web, does not fully reduce hallucinations (Zhao et al., 2024).

Figure 2

Artificial intelligence, machine and deep learning, generative AI

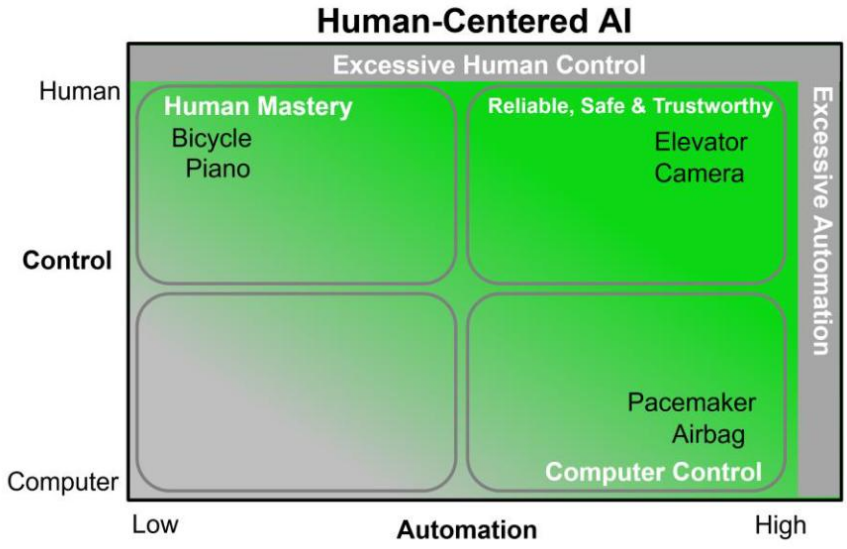


Note. Adapted from “The Application of AutoML Techniques in Diabetes Diagnosis: Current Approaches, Performance, and Future Directions” by L. P. Zhuhadar and M. D Lytras, 2023, *Sustainability, 15*(18), 13484 (https://www.mdpi.com/2071-1050/15/18/13484). CC BY 4.0.

Autonomy is about who or what is in control. It is a sliding scale going from a human being in full control to a machine being in full control, see Figure 3. Autonomy is related to automation, where tasks are executed according to rules with less or no intervention (Sarker et al., 2024). In the case of autonomy a machine does not follow pre-defined actions but can set its own path (Mayer, 2023; Veitch et al., 2024). An autonomous cyber defense agent should be capable to create a complex, multiple step plan and execute it (Kott, 2023). Shneiderman, however, challenged this single dimension concept and argued that humans and computers have different capabilities. By designing automated systems to be reliable, safe and trustworthy, humans can be provided finer control and be in control to reach their goals (Shneiderman, 2020b).

Figure 3

Human-Centered AI plotting automation against human control

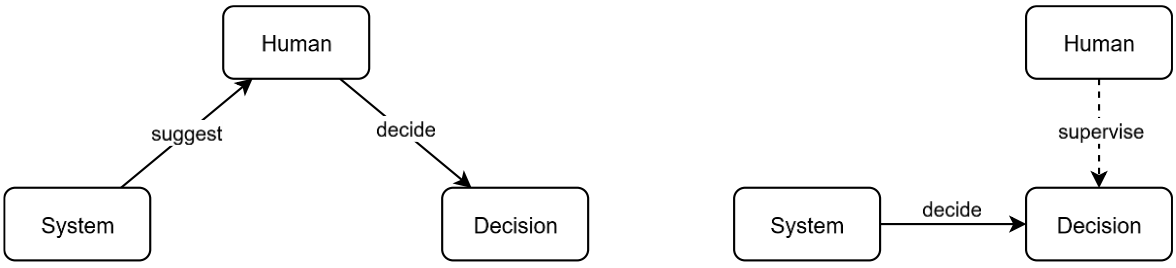


Note. From “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy” by B. Shneiderman, 2020, *International Journal of Human–Computer Interaction*, *36*(6), p. 495–504 (https://doi.org/10.1080/10447318.2020.1741118).

A distinction can be made between human-in-the-loop (HitL), human-on-the-loop (HotL) and human machine teaming based on the following literature. In HitL humans take key decisions, see Figure 4. For example, in the study of Hseih (2023), the physician took the final decisions. In military context there is always a human in the loop (Weger et al., 2023). In HotL systems or machines act autonomously while being monitored by humans (Fischer et al., 2021). According to Tolmeijer et al. (2022) humans have a supervisory role in HotL: They oversee the decisions made by AI and can correct mistakes or veto decisions made by AI. Human machine teaming entails humans and machines collaborate on tasks (Cleland-Huang et al., 2024). This study focuses on HotL: AI tooling making autonomous interventions. The specific way HotL is implemented (e.g., AI agents), is not an explicit part of this study. The first-round questionnaire of the Delphi study included a question about the possibility of vetoing.

Figure 4

Human in versus on the loop



High levels of autonomy impose a risk of deskilling, which is poignant because human operators need to be able to intervene when an autonomous system fails (Lee et al., 2025; Veitch et al., 2024). A higher level of autonomy results in a shared responsibility between AI and humans. A human does not have the ability to control a highly intelligent and automated system; AI and humans will have a shared control (Douer & Meyer, 2020). In another study about shared control; humans did not understand the decision making of an autonomous system when they needed to perform an emergency take-over (Chu et al., 2023).

Weger et al. (2023) found that the level of autonomy does not influence acceptance of autonomous systems. However, according to Hauptman et al. (2023) a higher level of autonomy is accepted when processes are clearly defined, when the team is more experienced and when there is less expected impact from the autonomous actions. This is in line with the guidelines for human-AI interaction that Amershi et al. (2019) proposed.

Liehner et al. (2022) found that automation is task-dependent: there is less reliance on automation in a higher risk context (e.g., risk of personal injuries) or when the automation is less reliable. Note that Liehner et al. used risk in a slightly different context than how cyber security risk has been defined in the Cyber Security section. In a study on cyber security incident response, automation is seen as a solution for low-level tasks only (Nyre-Yu et al., 2019). Low-level tasks could for example entail predefined playbooks that run automatically in certain conditions (Shaked et al., 2023).

AI agents can collaborate with humans as teammates (Hauptman et al., 2023). However, the concept of AI agents has been contested (Cabitza et al., 2021). Cabitza et al. argued that intelligence and agency should be moved to the human collective. In human-AI collaboration, AI is compared to human peers (Cai et al., 2019). Humans rely more on recommendations of AI than on those of humans experts (Tolmeijer et al., 2022). There is a recent trend that states that the design of future autonomous systems need to be human-centered (Weger et al., 2023). Shneiderman (2020b) proposed a human-centered AI framework that distinguishes between the level of automation and the level of human control, see Figure 3. He argued that with a good design highly automated systems can facilitate human control.

Finally, Barredo Arrieta et al. (2020) described Responsible AI as a concept that requires the implementation of several AI principles when putting AI models in practice. An AI model is built according to its design goals and is trained on data. It is the end product responsible for output of an AI functions of systems, for example their decisions (Amershi et al., 2019; Cai et al., 2019; Park et al., 2023). The Responsible AI principles mentioned are fairness, privacy, accountability, ethics, transparency, and safety & security. Transparency has been identified before as influencing technology acceptance. Technology acceptance will be described in the next section.

2.2.2 Technology Acceptance (or Adoption) of Automation

The Technology Acceptance Model by Davis (1989) has been extended by Weger et al. (2023) to include mission critical factors. Trust plays a crucial role in acceptance of technology. There is less reliance on automation in a higher risk context, and therefore less trust in automation. The context of the situation influences the trust in automation: in the medical domain automation was relied upon less than in a physical warehouse (Liehner et al., 2022). Trust in mission critical environments is influenced by factors reliability, safety/security, human in the loop and perceived usefulness, and transparency of how the system works (Weger et al., 2023). Reliability refers to the accuracy and conversely the failure rate of a system. In other words, how well the system performs (Vantrepotte et al., 2023; Weger et al., 2023). Transparency entails the visibility of the inner workings of a system, and by which steps it reaches a conclusion or makes decisions. Tangentially, Rudin (2019) stated that for high-stakes decisions there is a need for interpretable models. Interpretability implies transparency. Transparency is crucial to understand reasoning behind security decisions (Shreeve et al., 2023). AI systems should be understandable and trustworthy (Amershi et al., 2019). Extending transparency to GenAI: The transparency of LLMs, the most prominent technology behind generative AI, is lacking. Explainability builds on transparency: are the inner workings and outcomes comprehensible to the stakeholders of the system (Kuiper et al., 2022)?

Regarding AI assisting humans and human-AI teaming, acceptability of assistance systems (e.g., flight assistance systems) is positively influenced by system reliability and system explicability. By providing the assistance system's confidence in its suggested choice, users of the system felt a higher sense of agency (Vantrepotte et al., 2023). However, the adoption of an autonomous parking feature was low because humans believe that they are better and faster than the autonomous system (de Visser et al., 2023). Training helps in acceptance of AI teammates (Hauptman et al., 2023). Early involvement in the design process also helps in acceptance (Cabitza et al., 2021).

Explicability incorporates both intelligibility and accountability (Floridi et al., 2018). Intelligibility is captured by the concept of explainability; thus, explicability is a broader concept than explainability. Breaking tasks up in smaller steps, a process dubbed chaining, improves transparency and explicability (T. Wu et al., 2022).

There can be individual differences in acceptance. Younger people tend to be more accepting of technology (Hauptman et al., 2023; Weger et al., 2022). Women tend to be less accepting (Weger et al., 2022). Personal involvement and the traits optimism and innovativeness positively influence the adoption of AI assisted diagnosis (Hsieh, 2023).

In a medical context, AI can be mistrusted (Cabitza et al., 2021; Hsieh, 2023). The trust might be a trait of the user of the system (Cabitza et al., 2021). The transparency of the AI system could enhance trust (Cai et al., 2019). Understanding the capabilities and utility of an AI assisted diagnosis tool influences technology acceptance (Cai et al., 2019). However, Rudin (2019) noted that there is a trend in medicine to accept black box models whose inner workings are not transparent.

Finally, Shneiderman (2020a) deducted guidelines to increase user satisfaction of interaction with AI systems. These guidelines include factors trust and reliability. These guidelines will be used as a source for the first-round questionnaire.

2.2.3 Cyber Security

Cyber security is defined as the process of managing risks to protect the confidentiality, integrity and availability of an organization's IT systems and data/information (Berman et al., 2019; Flechais & Chalhoub, 2023; NIST, n.d.; Schatz et al., 2017).

A distinction should be made between cyber security of AI and AI in cyber security. The former is about attacks on or vulnerabilities in AI systems and how to protect against them (Sangwan et al., 2023). The latter is the subject of this study and entails applying artificial intelligence within the field of cyber security.

Formosa et al. (2021) adapted the work of Floridi et al. (2018) to the cyber security context. They confirm that among others the ethical principles of autonomy (i.e., human autonomy) and explicability are important ethical issues in cyber security. As stated above, explicability implies transparency.

In cyber security, context awareness is required (Nyre-Yu et al., 2019). If AI agents are considered as teammates as stated by Hauptman et al. (2023), the AI agents should have the same context awareness in the human-AI team. Having a shared understanding within a team helps in reaching optimal cyber security decisions (Shreeve et al., 2023).

Cyber security risk is a function of threat, vulnerability, and impact. Threat is the likelihood of an attack; vulnerability is the likelihood an attack succeeds; impact consists of the material (e.g., recovery cost) and non-material (e.g., reputation) consequences of a successful attack. Threat and vulnerability can be combined into a single variable: likelihood (Geer et al., 2020; Shreeve et al., 2023).

Cyber security managers take decisions alternating between risk-first and opportunity-first. Most cyber security frameworks prescribe a risk-based approach, but not every cyber security professional is able to grasp the entire risk landscape. Another effective approach is to consider the opportunities of the available cyber security measures and work from there (Shreeve et al., 2020).

AI tooling can provide decision-makers in cyber security with timely and relevant information. Cyber security decisions have multiple stakeholders. These stakeholders can have non-technical interests including compliance with laws and regulations, and ethical concerns. AI decision tools should take those interests into account. However, AI models can lack transparency and explicability. It is therefore difficult to ascertain whether AI decision tools are fair and unbiased (Flechais & Chalhoub, 2023).

Decision making in a Security Operations Centre can be supported by AI. As a specific example, the prioritization of cyber security alerts through reinforcement learning can withstand any attack using reinforcement learning (Shah et al., 2020). Generative AI solutions exist for a wide range of operational cyber security areas including intrusion detection, vulnerability detection, phishing detection, threat intelligence and malware classification (Ferrag et al., 2025). In the financial sector, the most anticipated AI enabled user facing security measures are insider threat detection, simulations of attacks and several measures related to training and awareness like tailoring and determining phishing susceptibility (Frank et al., 2025).

Incident response (IR) is an important organizational cyber security capability. It entails the planning and execution of a response to cyber security incidents. Shaked et al. (2023) proposed a concept that integrates the business operations with the IR process model. Cyber security defense interventions (e.g., IR) can have a significant negative impact on business operations.

Naghshvarianjahromi et al. (2023) stated, referring to Haykin (2019), that cognitive dynamic systems (CDSs) and not AI are appropriate to use within cyber security. CDSs are systems that can process information, receive continuous feedback from the environment and can learn from that experience (Hilal et al., 2023). However, machine learning can be described as learning, adapting to changes and improving from experience (Samoili et al., 2021). Samoili et al. explicitly considered the integration of perception, learning, action and interacting with the environment as a form of AI. Therefore, for this study CDSs will be considered a form of AI.

2.3 Conceptual Model

This study considers the practical implementation of state-of-the-art AI tooling in a cyber security context. Barredo Arrieta et al. (2020) described Responsible AI as a concept that requires the implementation of several AI principles when putting AI models in practice. The Responsible AI principles mentioned are fairness, privacy, accountability, ethics, transparency, and safety & security. This generic conceptual model is tangentially related to the research question, which is specifically about *acceptance* of AI tooling.

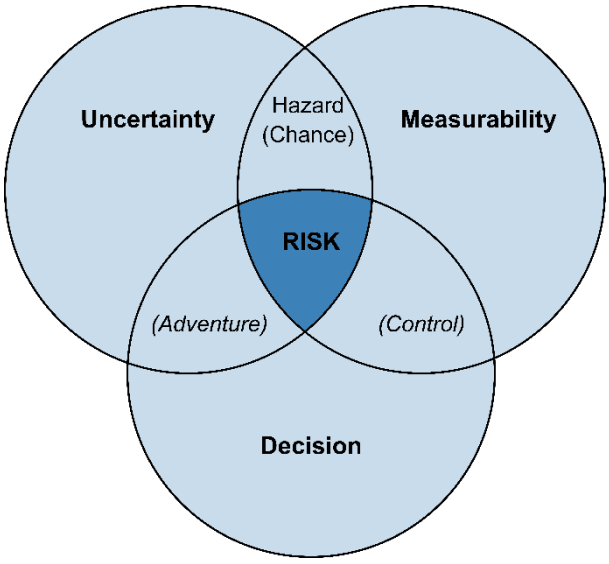
When focusing on adoption, technology acceptance models are more appropriate than the responsible AI concept. The mission critical technology acceptance model (MCTAM) extends the technology acceptance model (TAM) (Weger et al., 2023). The rationale for applying a mission critical model in this study is that a successful cyber-attack, especially a ransomware attack, can have a large negative impact on an organization and even on critical infrastructure. In some cases a ransomware attack has led to bankruptcy (Shaked et al., 2023; Yuryna Connolly & Borrion, 2022). AI-powered attacks could result in loss of trust in organizations and ultimately to systemic failures (Guembe et al., 2022). AI cyber defense tooling should help in preventing, detecting, and responding to such cyber-attacks. Note that Weger et al. (2023) conflated trust with acceptance of technology, even though trust is one of the factors influencing technology acceptance in their conceptual model. For example: "[…] resulted in similar findings with reliability and perceived usefulness being of greatest concern for a change in user trust […]".

Additionally, factors risk and context can influence the trust in automation (Liehner et al., 2022). Liehner et al. specifically considered decision making under risk as shown in Figure 5: The outcome of a decision in an uncertain situation can be measured, and the decision maker directly influences the outcome.

The different contexts tested in the study of Liehner et al. are urban, medical and logistics (i.e., a warehouse). A study on acceptance of autonomous security robots extending the TAM found a positive influence of usefulness and trust on acceptance. Physical risk influenced acceptance only in low-risk situations, not in high risk situations where the security robot would physically intervene (Ye et al., 2024). As a follow-up to the study of Liehner et al., Brauner et al. (2023) found that the public expected that AI is likely to be hacked, which is an undesired outcome. This confirms the link between trust and safety/security expectancy in the MCTAM.

Figure 5

Risk

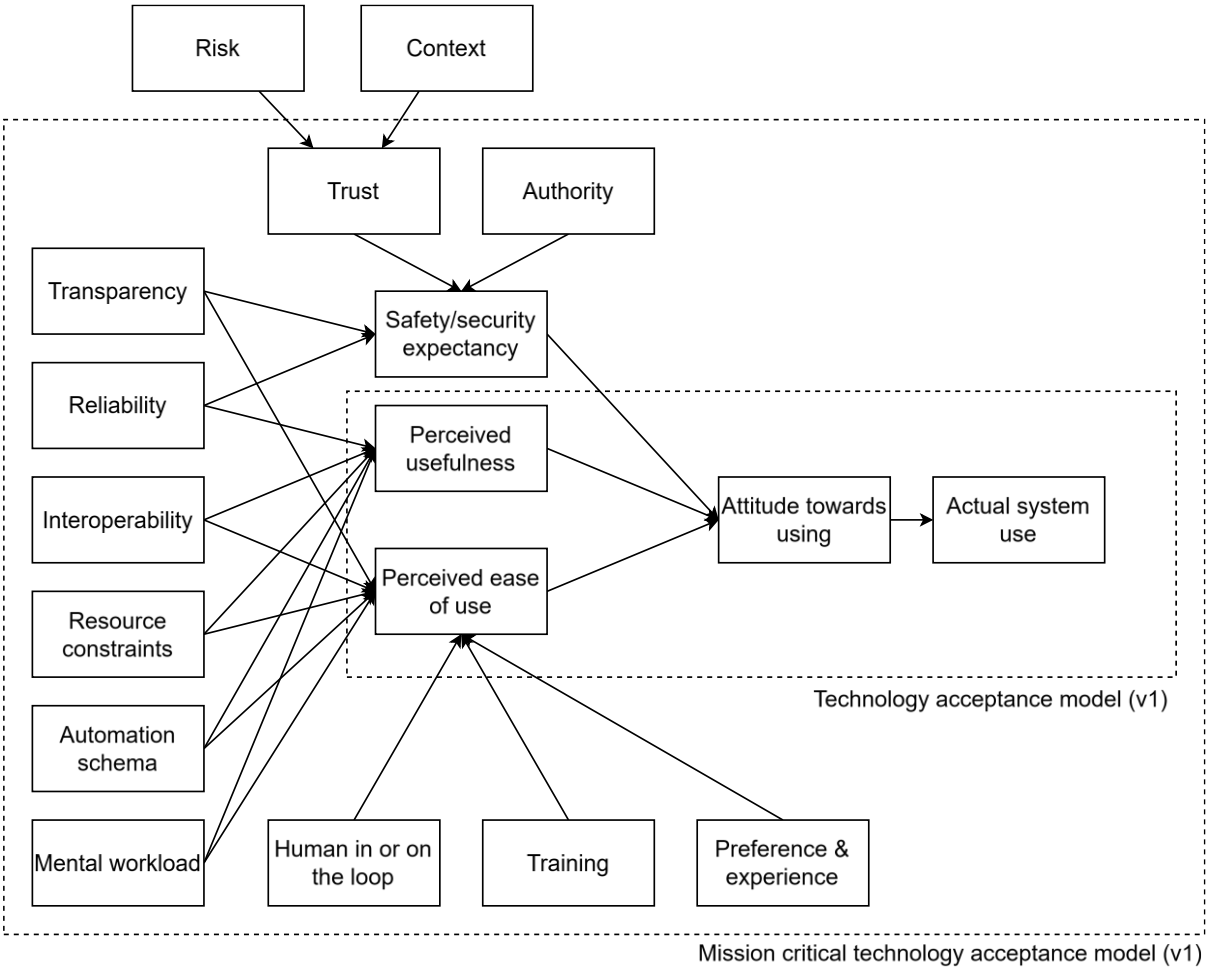


Note. From "Delegation of Moral Tasks to Automated Agents—The Impact of Risk and Context on Trusting a Machine to Perform a Task" by G. L. Liehner, P. Brauner, A. K. Schaar, and M. Ziefle, 2022, *IEEE Transactions on Technology and Society*, *3*(1) (https://doi.org/10.1109/TTS.2021.3118355).

The MCTAM includes the human in the loop concept. This study focuses on human *on* the loop, instead of human *in* the loop. For this study, the MCTAM has been modified accordingly, see Figure 6. Concepts *Context* and *Risk* as described by Liehner et al. have been added. This MMCTAM conceptual model has been used as a starting point for the literature review for SQ1 and SQ2: the different concepts (e.g., *transparency*) are used as potential factors for the Delphi study. The relations (i.e., arrows) between the aspects in the MMCTAM are not used during the literature review but are circled back to in the Discussion. The research method, including the literature review process and Delphi study, are explained in the next chapter. The results of the literature review for SQ1, SQ2 and SQ3 are presented in Chapter 4.

Figure 6

Modified Mission Critical Technology Acceptance Model (MMCTAM)



Note. Adapted from "Insight into User Acceptance and Adoption of Autonomous Systems in Mission Critical Environments" by K. Weger, L. Matsuyama, R. Zimmermann, B. Mesmer, D. Van Bossuyt, R. Semmens, and C. Eaton, 2023, *International Journal of Human-Computer Interaction*, *39*(7), 1423–1437 (https://doi.org/10.1080/10447318.2022.2086033).

3 Research Method

For this study a three round Delphi study was performed, see Figure 7.

First, a list of factors related to the acceptance of AI tooling within cyber security defense was extracted from the literature. The concepts of the MMCTAM - and not the relation between them - have been used as a starting point for the literature review for SQ1 and SQ2. To tailor to the expert panel, concepts related to *usefulness* were used and concepts related to *ease of use* were not. The literature review is the first step leading up to the Delphi study, see Figure 7.

Second, the list of factors was validated by interviewing one CISO from a Dutch financial institution.

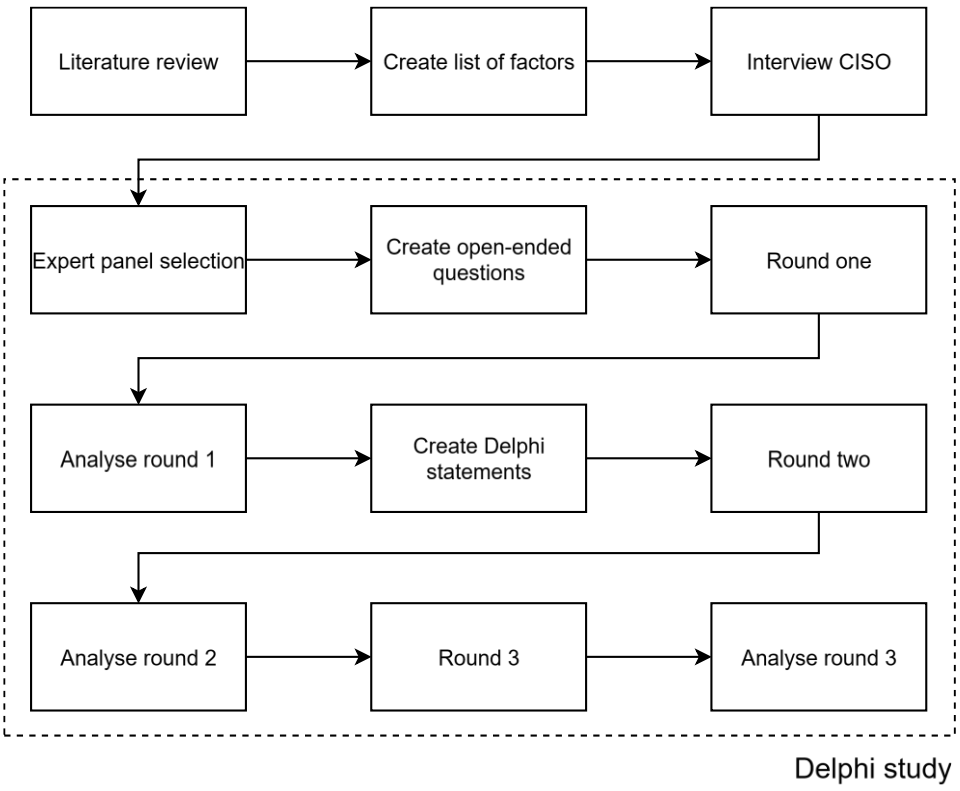
Third, chief information security officers (CISOs) were selected for the expert panel of the Delphi study, primarily because the CISO function takes both the business and IT perspective into account.

Next, a qualitative self-completion questionnaire was created based on the literature and feedback during the interview. The responses from the first round were input for the quantitative second round. All statements from the second round were reused in the third round. Both the second and third round consisted of a six-point Likert questionnaire.

Finally, the responses from the third round were analyzed to determine for which statements a consensus has been reached.

Figure 7

Research method



Note. Adapted from “The Delphi Method for Graduate Research” by G. J. Skulmoski, F. T. Hartman and J. Krahn, 2007, *Journal of Information Technology Education: Research, 6,* 001–021 (https://doi.org/10.28945/199).

3.1 Initial literature review

For Chapter 2, theoretical background, a literature review was carried out. The method has been included here, because some of those papers have been used in the remainder of the thesis.

Studies were retrieved from electronic databases Academic Search Complete and Business Source Premier (EBSCOhost) and from ACM. For EBSCOhost “Apply equivalent subjects” was disabled. This forces exact matches (*What Is the Apply Equivalent Subjects Expander?*, n.d.). Zotero was used as a reference manager.

A search was performed on terms ‘human computer interaction’ and ‘autonomy’, ‘human artificial intelligence interaction’ and ‘autonomy’, ‘human computer collaboration’ and ‘autonomy’, ‘human artificial intelligence collaboration’ and ‘autonomy’, ‘human artificial intelligence teaming’ and ‘autonomy’, or ‘human machine teaming’ and ‘autonomy’ in the title, abstract or keywords. The abbreviation ‘ai’ was used as a variation of ‘artificial intelligence’ in the search terms. ‘Autonomous’ was used as a variation of ‘autonomy’ in the search terms. The exact EBSCOhost advanced search syntax:

(TI ("human computer interaction" OR "human artificial intelligence interaction" OR "human computer collaboration" OR "human artificial intelligence collaboration" OR "human artificial intelligence teaming" OR "human machine teaming") OR AB ("human computer interaction" OR "human artificial intelligence interaction" OR "human computer collaboration" OR "human artificial intelligence collaboration" OR "human artificial intelligence teaming" OR "human machine teaming") OR KW ("human computer interaction" OR "human artificial intelligence interaction" OR "human computer collaboration" OR "human artificial intelligence collaboration" OR "human artificial intelligence teaming" OR "human machine teaming")) AND (TI (autonomy OR autonomous) OR AB (autonomy OR autonomous) OR KW (autonomy OR autonomous))

The EBSCOhost search was limited to dates ranging from 2019 to April 9, 2024, which is the date the query was run. This resulted in 96 papers, of which 95 are peer reviewed.

The exact ACM search syntax:

(keyword:("human computer interaction") OR keyword:("human artificial intelligence interaction") OR keyword:("human computer collaboration") OR keyword:("human artificial intelligence collaboration") OR keyword:("human artificial intelligence teaming") OR keyword:("human machine teaming")) AND (keyword:("autonomy") OR keyword:(autonomous))

The ACM search was limited to dates ranging from 2019 to April 16, 2024, which is the date the query was run. This resulted in 6 papers, which are all peer reviewed.

For ACM the content was filtered on research articles.

Papers that did not fit the context of the research question were disregarded. For example, papers that have been disregarded include research on tracking physical interaction between humans and computers, on perceptions of using consumer focused AI apps, on cyber-crime (i.e., fraud), on cyber-attacks on smart power grids, and on decision-making that did not apply or could not be transposed to cyber security automation.

Papers that did not reflect the state of the art in academic research have been disregarded: systematic reviews, papers without a research design or without newly performed research.

Papers that are not in English or that are not accessible due to a paywall were excluded as well. Papers that are not peer reviewed have only been included if they contributed were valuable. After applying the exclusion criteria, four papers from EBSCOhost and two from ACM were deemed valuable for the initial literature review.

For cyber security, a separate literature review was performed in an analogous way on search terms ‘cyber security’ and ‘autonomy’, or ‘cyber security’ and ‘decision-making’. The exact EBSCOhost advanced search syntax:

(TI ("cyber security") OR KW ("cyber security") OR AB ("cyber security")) AND (TI (autonomy OR "decision making") OR AB (autonomy OR "decision making") OR KW (autonomy OR "decision making"))

This search was limited to dates ranging from 2019 to April 23, 2024, which is the date the query was run. This resulted in 93 papers, of which 83 are peer reviewed.

The exact ACM search syntax:

(Abstract:("cyber security") OR Keyword:("cyber security") OR Title:("cyber security")) AND (Abstract:(autonomy) OR Abstract:("decision making") OR Keyword:(autonomy) OR Keyword:("decision making") OR Title:(autonomy) OR Title:("decision making"))

The ACM search was limited to dates ranging from 2019 to April 23, 2024, which is the date of the query. This resulted in 20 papers, which are all peer reviewed.

After applying the exclusion criteria, three papers from EBSCOhost and four from ACM were deemed valuable for the initial literature review.

Finally, backward and forward searches were done.

The summarized literature review process for the theoretical background:

- Records identified through database searching (n=195)
- Exclusion criteria:
    - No “equivalent subjects expander”
    - Language other than English
    - Full-text not available
    - Mismatched context (e.g., cybercrime)
    - Not peer reviewed unless it is a valuable addition
    - Not state of the art (e.g., systematic review)
- Full-text articles assessed for eligibility
- Studies included (n=13)
- Additional records identified through snowballing and for definition of concepts (n=43)

3.2 Literature

A list of factors related to the acceptance of AI tooling within cyber security defense was extracted from the literature via sub questions one, two and three: What factors are associated with acceptance of AI advising humans on which actions to take? What factors are associated with acceptance of the outsourcing of cyber security response (to external Security Operation Centers)? What requirements exist for AI usage within the Dutch financial sector?

The Modified Mission Critical Technology Acceptance Model (MMCTAM has been used as a starting point for the literature review for SQ1 (and SQ2. That is, the concepts from the MMCTAM relevant for SQ1 and SQ2 are used. Additionally, an exploratory literature review is performed for SQ1, SQ2 and SQ3.

Studies were retrieved from ACM, Hugo and Semantic Scholar. Hugo is an online tool provided by the HU University of Applied Sciences, based on Worldcat. Papers that were not in English, not accessible due to a paywall or did not fit the context of the research question were disregarded. For Hugo, duplicates were hidden. Zotero was used as a reference manager.

3.2.1 SQ1

The search for SQ1 was limited to dates ranging from 2019 to June 18, 2024, which is the date the query was run. It was performed via Hugo on terms (‘adoption’ OR ‘acceptance’) AND (‘artificial intelligence’ OR ‘human in the loop’), filtered on peer reviewed papers, and with duplicates hidden.

The exact Hugo query: ti:(("adoption" OR "acceptance") AND ("artificial intelligence" OR "human in the loop")). This resulted in 245 papers. Two out of the first 20 results were included for SQ1, because these two papers both reviewed literature on (critical success) factors related to adoption of AI (Kelly et al., 2023; Solaimani & Swaak, 2023).

On ACM, titles were searched on the term ‘human in the loop’. It was limited to dates ranging from 2019 to July 2, 2024, which is the date the query was run. For ACM the content was filtered on research articles. This resulted in 72 papers.

After applying the exclusion criteria, two papers from Hugo and two from ACM were deemed a valuable contribution for SQ1.

Finally, backward and forward searches were done.

The summarized literature review process for SQ1:

- Records identified through database searching (n=317)
- Exclusion criteria:
  - No “equivalent subjects expander”
  - Language other than English
  - Full-text not available
  - Mismatched context (e.g., cybercrime)
  - Not peer reviewed unless it is a valuable addition
- Full-text articles assessed for eligibility
- Studies included (n=4)
- Additional records identified through snowballing and for definition of concepts (n=17)

3.2.2 SQ2

Hugo was queried on ‘managed security services’, ‘managed security service provider’, ‘security operations center’ or ‘security operations centre’ between 2015 and 2024, because outsourcing was expected to be an older trend within IT, and therefore researched less the past 5 years. This assumption was wrong, see below.

The exact Hugo query:

ti:("managed security services" OR "managed security service provider" OR "security operations center" OR "security operations centre")

This results in 29 papers of which 13 are peer reviewed according to Hugo. Only 2 papers are from before 2019.

After applying the exclusion criteria, three papers were deemed valuable for SQ2.

Finally, backward and forward searches were done.

Due to the limited available literature on outsourcing security response specifically, any factors identified have been included even if outsourcing was not an explicit part of a study.

The summarized literature review process for SQ2:

- Records identified through database searching (n=29)
- Exclusion criteria:
  - No “equivalent subjects expander”
  - Language other than English
  - Full-text not available
  - Mismatched context (e.g., cybercrime)
  - Not peer reviewed unless it is a valuable addition
- Full-text articles assessed for eligibility
- Studies included (n=3)
- Additional records identified through snowballing and for definition of concepts (n=11)

3.2.3 SQ3

The laws and regulations (AI Act, GDPR, etc.) formed the bulk of the literature review for SQ3, because as stated in 1.5 the financial sector is heavily regulated, and SQ3 is meant to identify sector-specific factors based on those laws and regulations. Previously identified papers that addressed those laws and regulation have been used as well.

When the regulations required further guidance, a specific search was carried out: i) Semantic Scholar was searched on terms (‘GDPR’ ‘article 22’ ‘significant effects’) in October 2024 filtered on articles with a PDF and ii) Hugo was searched on (‘foundation models’ AND ‘general ‘purpose’). The first search resulted in 29 articles. The second search resulted in five articles. After applying the exclusion criteria, one paper from Semantic Scholar and one from ACM were deemed a valuable addition for SQ3.

3.3 Interview

The factors identified through the literature review from SQ1 and SQ2 were validated by interviewing one CISO from within the Dutch financial sector. See Appendix G for the transcript. CISOs of other financial institutions have not been involved in this step. This minimized the time that the other participants had to spend on this study, to limit attrition (Schmalz et al., 2021).

During the interview the factors derived from the first two sub questions were presented one by one. For each factor the interviewee was asked if the factor is relevant to the research, and if it is an important factor. The interviewer rated the response as *very high*, *high*, *medium*, or *low* importance. *Very high* was applied when the interviewee stressed that a factor was really important. *High* was applied when the interviewee answered with a clear affirmative, *low* was applied in case of a clear 'no' with regards to importance, and medium was applied when the interviewee was not very outspoken on the factor being important or not. During the interview, an additional importance level became apparent: very high priority. The interviewee did not explicitly mention the importance of all factors; in that case the interviewer assigned an importance based on the notes taken during the interview and transcription.

In the initial design of the research SQ3 was out of scope for the interview. The interviewer did, however, ask the following about the potential factor of general-purpose AI concentration risk:

1. In what way do you think concentration risk relates to accepting AI tooling? As an example, consider outsourcing autonomous AI to the same supplier as huge parts of IT and cyber defense are outsourced to?
2. What about cloud providers like Microsoft (Azure), Google (Cloud), Amazon (AWS) or Alibaba (Cloud)?

After all factors were discussed, two questions were asked: 1) Do you agree that these provide sufficient coverage for the sub questions? and 2) Have any potential factors related to the research question been overlooked? The list of factors was slightly adjusted based on the feedback received. The subsequent Delphi study started with a qualitative round, where the responses of participants could introduce additional factors. Therefor the validation of the list of factors identified via the literature review by a single expert was deemed to be sufficient. The following protocol was followed during the interview.

3.4 Interview Protocol

The interview was semi-structured. The interview protocol was adapted from Jacob and Furgerson (2015) and Bell et al. (2022, Chapter 20).

The description of clustered factors in Table B3 (Appendix B) was used as a glossary (for the interviewer) during the interview.

Microsoft Teams was used for one interview. After starting the call, recording was started in Microsoft Teams and transcription was enabled. After the interview, the automatic transcription was compared with the recording, checked on mistakes and amended where necessary. The transcription should reflect what was said verbatim; any deviations resulting from automatic transcription were corrected. Any reference to names of people or (information traceable to) firms were removed from the transcription. The recording itself was removed.

The interviewee was informed that answers to interview questions, the Delphi study and results of the study will be pseudonymized: name, email address and firm will be replaced by a single unique number. The reference table linking that number to the participants in both the interview and Delphi study are stored separately from the research data and shall never be shared or published. The data the participants shared can be traced back to them via the linking table, which according to the General Data Protection Regulation (GDPR) is pseudonymization, not anonymization.

The interviewer introduced the study, the sub questions that are in scope for the interview, and the identified factors. For each factor, the interviewer asked if this factor is relevant to the topic at hand and if it is of high or low importance. After all factors had been presented, the interviewer checked if the interviewee agreed that these factors cover the sub questions and by extension the research question, or if the interviewee identified any gaps.

Finally, the interviewer wrapped up the interview by thanking the interviewee and stressing that the interviewee could contact the interviewer in case of questions.

3.5 Delphi Study

A classic three round Delphi study was performed. Delphi studies can be used to research the risks, opportunities and critical success factors of new technologies (Moeuf et al., 2020). Delphi is a research method suited for situations with incomplete knowledge, for example anticipated future developments (Schmalz et al., 2021; Skulmoski et al., 2007). The goal of Delphi is to reach a consensus between a group of experts, unless opinions of experts diverge (Beiderbeck et al., 2021; C.-C. Hsu & Sandford, 2007).

Participants in a Delphi study remain anonymous (Beiderbeck et al., 2021; Belton et al., 2019; Franklin & Hart, 2007; C.-C. Hsu & Sandford, 2007; Skulmoski et al., 2007). Anonymity should lower the barriers to participation for professionals in the financial sector.

In this study the goal was to see if experts agree or disagree on future developments in cyber security regarding usage of AI.

3.5.1 Delphi study setup

Nowack et al. (2011) listed three possible functions in a Delphi: the idea generation, consolidation and judgment functions. The idea generation function entails identifying future trends. The consolidation function then reduces the number of trends. Both functions are part of scenario development. The judgment function evaluates the scenarios. Almost every Delphi study they reviewed had a judgment function. The idea generation and consolidation function were present in about half the studies.

Both qualitative and quantitative studies are possible with the Delphi method (Skulmoski et al., 2007). For this study a mixed-method approach was used. The first round consisted of a questionnaire with open-ended questions. The second round consisted of a quantitative questionnaire using a six-point Likert scale to rate the participants' agreement with each statement, namely: 6 = 'fully agree', 5 = 'agree', 4 = 'slightly agree', 3 = 'slightly disagree', 2 = 'disagree' and 1 = 'fully disagree' (Joia & Cordeiro, 2021; Petsani et al., 2024; Trevelyan & Robinson, 2015).

Delphi statements are projections about the future (Beiderbeck et al., 2021; Schmalz et al., 2021). Within literature the terms *statement* and *scenario* are used interchangeably. Example statements are "Fintechs have low operating costs, managing to offer services to clients at lower values than traditional financial institutions." (Joia & Cordeiro, 2021, p. 5); "Many managers are not aware of how blockchain works or what benefits it brings to their enterprise." (Saheb & Mamaghani, 2021, p. 7); "In 2035, passengers will increasingly be willing to pay more for eco-friendly door-to-door journeys." (Kluge et al., 2020, p. 5). Participants rate statements on a Likert scale in case of a quantitative round. All statements should be either positive *or* negative, not mixed (Schmalz et al., 2021). Lengthy or abstract statements lead to less variance in response and the responses will be more moderate: the extremes on the Likert scale will be chosen less (Markmann et al., 2021). In practice statements average between 10 to 20 words, but there is no evidence to support this (Andersen, 2023). Andersen (2023) composed recommendations on Delphi statements as shown in Table B1 (Appendix B).

For this study, a scenario makes a prediction about the near future. In the first round a scenario is accompanied by one to three questions. In the second and third round, either the scenario itself has been rated by the participants, or one to three statements about the scenario were presented and rated. In the latter case, the scenario itself was not rated. In the remainder of this study, a scenario refers to the top-level description, and a statement refers to a prediction that has been rated by the participants. A scenario can also be a statement, but a separate statement is always lower in hierarchy than the scenario it corresponds to.

The first, explorative round of this study consisted of open-ended questions about the presented scenarios. Open questions have an, albeit limited, idea generation function. The scenarios and statements for the second were developed based on a literature review (idea generation) and the expert input from the first round (consolidation). The scenarios and statements of the second round were reused in the third round. In the second and third round, the experts judged the presented statements.

3.5.2 Expert Panel Selection

The group of participants consisted of a panel of experts (C.-C. Hsu & Sandford, 2007; Skulmoski et al., 2007). As this study was scoped on the Dutch financial sector, the group of participants consisted

of cyber security professionals from the Dutch financial sector: CISOs. CISOs need to continually consider business processes and objectives (Maynard et al., 2018). CISOs take a business-driven approach to security (Kayworth T. & Whitten D., 2010). CISOs interpret the cyber security situation of their organizations and translate it to terms the business leaders understand (Da Silva & Jensen, 2022). Therefore, the participating CISOs should take both the business and IT perspective into account.

An expert panel consisting of CISOs resulted in a homogenous sample. In case of a homogeneous sample, 10 to 15 participants gives sufficient results (Delbecq et al., 1976; C.-C. Hsu & Sandford, 2007; Skulmoski et al., 2007). Beiderbeck et al. (2021) however suggest at least 15 to 20 participants. Belton et al. (2019) suggest a minimum of five. Due to the small number of organizations in the Dutch financial sector, the aim was 10 to 12 participants. Participants in a Delphi study remain anonymous and should be instructed to refrain from sharing personal information which would compromise their anonymity. Electronic means for taking questionnaires are preferred to maintain anonymity (Beiderbeck et al., 2021; Belton et al., 2019; Franklin & Hart, 2007; C.-C. Hsu & Sandford, 2007; Skulmoski et al., 2007). Anonymity should lower the barriers to participation for professionals in the financial sector.

The financial sector includes banks, non-bank financial institutions (NBFIs) and capital markets, stock markets and bond markets. NBFIs include among others insurance companies, real estate and pension funds. NBFIs are less regulated or unregulated in the EU. NBFIs can be referred to as shadow banking. However, insurance companies and pension funds are not included in the narrow definition of shadow banking (Apostoaie & Bilan, 2020; Busch & van Rijn, 2018; Davies & Killeen, 2018; Melecky & Podpiera, 2020; Pradhan et al., 2018; Shahzad et al., 2019). The European Systemic Risk Board and Financial Stability Board considers pension funds and insurance companies as a separate group from other financial institutions (Resti et al., 2021).

Dutch financial institutions participate in an information sharing and analysis center (FI-ISAC NL) (Dutch Banking Association, n.d.). Their members include banks and organizations that are part of the Dutch financial vital infrastructure (Nationaal Coördinator Terrorismebestrijding en Veiligheid, 2023). The CISOs of ±20 organizations participating in the FI-ISAC NL were initially invited to participate in this Delphi study. Seven CISO and one former information security officer (ISO) participated in all three rounds.

3.5.3 Questionnaire

Self-completion questionnaires (also called surveys) can be completed online. Without an interviewer present the likelihood of participants exhibiting the social desirability effect is lower and there is no interviewer variability. However, using open-ended questions in questionnaires is more challenging compared to interviews: the participant cannot be probed to elaborate on an answer (Bell et al., 2022, p. 231).

In Delphi research the exploratory step can consist of either a questionnaire with open-ended questions or narrow questions based on the literature (C.-C. Hsu & Sandford, 2007; Skulmoski et al., 2007). Open questions allow for more creative input (Nowack et al., 2011). Open questions tap into extant knowledge of the participants to address any gap between literature and recent developments (Franklin & Hart, 2007). One or two questions per Delphi statement are sufficient to gather satisfactory qualitative input, with a maximum of three open-ended questions (Beiderbeck et al., 2021). Open-ended questions can be asked after introducing a decision-making scenario (Hirschhorn, 2019). The first round consisted of open questions.

The second and third round consisted of a quantitative questionnaire using the six-point Likert scale. A free text input field accompanied each Delphi statement in the second round, which allowed participants to elaborate on their choice (Belton et al., 2019). The third round included a single free text input field at the end of the questionnaire.

3.5.4 Questionnaire in Microsoft Forms

The participants were informed that answers to the Delphi study and results of the study would be pseudonymized.

The questionnaires were created using Microsoft Forms. When the participants span multiple organizations Microsoft Forms can only send an ‘anonymous’ link. That is, a single link that is shared with all participants. Microsoft Forms has no option to send unique links per participant, except to participants in the same organization as the creator of the form.

In the first round the email addresses of participants were collected. For Round 2 and 3, the questionnaire was cloned to create a unique questionnaire instance per participant to work around the Microsoft Forms limitations mentioned previously. The participants received their unique link(s) and did not have to submit their email address in Round 2 or 3. The link-participant pair was tracked outside of Microsoft Forms. Summarized feedback about the remarks in the second round was provided in the third round. In the third round, for each statement, the questionnaire showed the median of that statement and each participant's choice in Round 2.

Participants were allowed to save their responses; this option was only available if they were logged on with a Microsoft account. After a questionnaire had been filled in, additional responses were still accepted. This allowed the participants to edit their response.

3.5.5 Round 1

The first round consisted of a questionnaire with open-ended questions. The Delphi scenarios for the first-round questionnaire were developed based on the literature review. These initial Delphi scenarios left more room for interpretation than the Delphi scenarios and statements in the subsequent two rounds. The questionnaire was reviewed and tested by a security professional who was not part of the participant group. As an additional review step, the questionnaire was sent to one participant first and feedback on the questionnaire was provided by them in a short meeting. After minor changes, seven other participants took part in the Delphi first round.

The participants received an email with a link to the questionnaire. They also received a plain text email to announce the questionnaire email, in case the email with link got flagged as spam or phishing. Both emails included an introduction to the study, an explanation of the process and contact details for questions. A reminder was sent two weeks after the initial email.

3.5.6 Analysis Round 1

The responses to the open-ended questions were thematically analyzed using Atlas.ti version 25. The resulting themes served as input for Round 2.

First, the responses were matched to the identified factors via deductive coding. Additional topics discovered were coded using an inductive approach (Bell et al., 2022, p. 529). Responses with a similar meaning were combined (von Briel, 2018).

Next, a thematic analysis was performed (Bell et al., 2022, Chapter 25). A *theoretical* thematic analysis approach was followed: the codes were based on previous research (i.e., the identified factors) (Braun & Clarke, 2006). Codes, both newly identified and relating to the previously identified factors, were categorized to add a layer of abstraction. Relations between codes and code categories were made. Firstly, this grounded the identified factors in the data collected. Secondly, categories and to lesser extent codes with a higher occurrence were considered as more important. The categories were grouped into themes. This resulted in a prioritized list of themes and categories as input for the second round.

The four scenarios and 10 questions covered 25 factors out of 36 factors. As shown in Table B5 (Appendix B), all three very high priority factors were included in the scenarios. All high priority factors except implementation and maintenance cost were included. See Delphi Questionnaire Round One in the appendix for the full questionnaire. The Delphi scenarios and statements for the second-round questionnaire were developed based on the first-round analysis (Schmalz et al., 2021; von Briel, 2018).

3.5.7 Round 2

The second round consisted of a quantitative questionnaire using a six-point Likert scale. Distribution was the same as in Round 1, to the same expert panel.

The optimal range for a Likert scale is between four and seven points. An even numbered scale was used to prevent issues surrounding the mid-point; these issues include neutral bias and unclarity whether the neutral option was chosen due to not knowing, not deciding or not having an opinion. A *No comment* or *No answer* option can be added for instances where participants either do not have an opinion or do not know (Petsani et al., 2024; Trevelyan & Robinson, 2015). This study used a six-point Likert scale without a neutral mid-point, with a *No comment* option.

A free text input field accompanied each Delphi scenario in the second round, which allowed participants to elaborate on their choice (Belton et al., 2019).

Six scenarios and 16 statements were developed based on the responses in the first round, guided by the prioritized list of factors in Table B4 (Appendix B) and the factor coverage in Table B5 (Appendix B). See Appendix F for the full questionnaire.

The six scenarios and 16 statements covered 23 out of 37 factors, see Table B7 (Appendix B). The number of statements was limited to 16, because the participants had been promised that their participation would take one hour in total at most. All factors that were recognized by the participants were included in the second round, as was the newly introduced factor *Ability to train AI*. Very high priority factor *Privacy* was not mentioned in Round 1 and was not included in Round 2. High priority factor *Job displacement* was not included for the same reason.

The second round was tested on one participant. In a short interview after filling in the questionnaire, they highlighted that fully autonomous interventions in 2028 can be expected but only in relatively low impact use cases. Their view is that autonomous response on a false positive will have more impact than not autonomously responding on a true positive. Statement 16 has been added to the questionnaire based on this feedback before distribution to the other seven participants.

The response of the first round seems to indicate that AI tooling in general is accepted, unless it impacts business processes, or it is not accurate or capable for its purpose. Therefore, for Round 2 a ‘baseline’ question has been added that asks about acceptance of AI tooling with a human in the loop.

3.5.8 Analysis Round 2

In a Delphi study a distinction should be made between consensus, agreement, and stability. For each statement, consensus is considered per round. Agreement entails agreeing with the Delphi statement. Stability concerns the (lack of) changes in responses between rounds (Beiderbeck et al., 2021; Saheb & Mamaghani, 2021; Trevelyan & Robinson, 2015; von der Gracht, 2012). For this study, the goal was to find consensus; participants agreeing or not agreeing to a Delphi statement is an equally valid result. Agreement was measured via a six-point Likert scale, ranging from *Fully disagree* to *Fully agree*, without a neutral option.

Consensus can be assessed by the central tendency and level of dispersion of the responses. The central tendency was measured via the median, which is preferred above the mean (Trevelyan & Robinson, 2015).

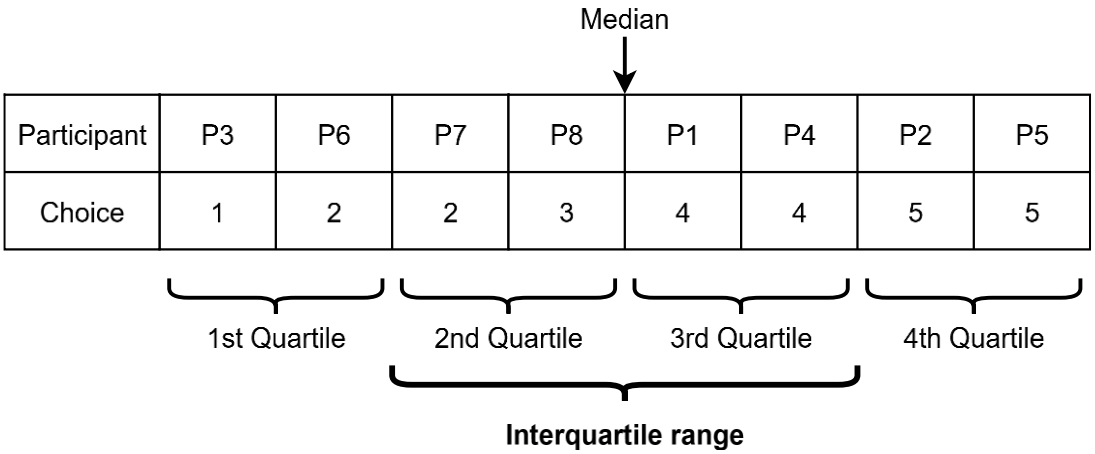
Level of dispersion was measured via the interquartile range (IQR). The IQR is the difference between the lower quartile (Q1) and the upper quartile (Q3) and corresponds with 50% of the responses, see Figure 8.

An IQR less than one means that 50% of the response are within 1 point distance on the Likert scale, which means 50% is clustered around one point (e.g., *Fully agree*) (Beiderbeck et al., 2021; C.-C. Hsu & Sandford, 2007; Schmalz et al., 2021; Trevelyan & Robinson, 2015; von der Gracht, 2012). An IQR ≤ 1 is more commonly used for Likert scales with less than 9 points (Schmalz et al., 2021; von der Gracht, 2012). However, Trevelyan and Robinson suggest an IQR ≤ 1.75.

For this study, considering the small sample size and a study design that includes a maximum of two quantitative rounds, the recommendation of Trevelyan and Robinson was followed: consensus has been reached for a statement if the IQR is lower than or equal to 1.75.

Figure 8

Median and IQR



Note. Median and IQR for Statement 10 in Round 3.

The median and IQR were calculated in Microsoft Excel using the R-8 formula as proposed by Hyndman and Fan (1996):

$$h = (N + 1/3)p + 1/3$$

$$Q_p = x\lfloor h \rfloor + (h - \lfloor h \rfloor)(x\lceil h \rceil - x\lfloor h \rfloor)$$

Where we "compute Q_p, the estimate for the p-quantile (the k-th q-quantile, where $p = k/q$) from a sample of size N by computing a real valued index h. When h is an integer, the h-th smallest of the N values, x_h, is the quantile estimate. Otherwise a rounding or interpolation scheme is used to compute the quantile estimate from h, $x\lfloor h \rfloor$, and $x\lceil h \rceil$" ("Quantile," 2025). In this study, $q = 4$ and the first and third quantile are calculated, for $p = 1/4$ and $p = 3/4$. In this study, the sample size N equals the number of responses to a statement.

The IQR was also calculated using Excel's built-in QUARTILE.EXC function which corresponds with IBM SPSS's default algorithm (i.e., HAVERAGE) (*Quantile Function*, n.d.). Both employ R-6. Excel and SPSS do not have built-in functionality for R-8. Additionally, the IQR using both R-6 and R-8 formulas, the median and quartiles have been calculated in the programming language R. To be clear, the language R on one hand, and R-6 and R-8 on the other, are not related. The R code is available in Appendix H.

The difference between R-6 and R-8 is in the way they calculate the value of a quartile, x_h, by calculating h differently. R-6 uses the following formula:

$$h = (N + 1)p$$

Regarding the calculation of the IQR, von der Gracht et al. (2012) refer to De Vet et al. (2005) who refer to Jones and Hunter (1995). Jones and Hunter, however, do not further define the IQR or quartiles. It is unclear which algorithm Trevelyan and Robinson (2015) employed, so the worst case was picked (i.e., the highest IQR).

3.5.9 Round 3

The third round consisted of the same quantitative questionnaire as the second round. A free text input field was added at the end of the questionnaire, not for every scenario. Distribution was the same as in first two rounds, to the same expert panel.

All Delphi statements of the second round were reused in the third round. Some studies however opt to only reuse the statements that did not reach a consensus (Trevelyan & Robinson, 2015). Each participant received customized feedback: Their previous response and the median were presented for each statement, as well as a summary of comments from all participants. The median and previous response were marked as shown in Figure F1 (Appendix F). Another option would have been a histogram as Schmalz et al. suggested. With the presented information the participants could ascertain how their responses related to those of the entire panel (Franklin & Hart, 2007; Schmalz et al., 2021).

Four participants left remarks. Based on their feedback, a clarification was added to statements two and three, see section F.2 in the appendix.

3.5.10 Analysis Round 3

Consensus was measured the same way as for the second round.

The stability between rounds can be measured after at least two quantitative rounds. Chi-square has been used to assess stability in previous Delphi studies but is advised against. For this study, the Wilcoxon matched pairs signed rank test was used with $p < .05$. Stability between rounds is achieved if there is no significant change, that is, we fail to reject the null hypothesis of the Wilcoxon matched pairs signed rank test (Trevelyan & Robinson, 2015).

The Wilcoxon matched pairs signed rank test requires at least six samples for a $p < .05$. For a small sample size, Wilcoxon signed rank test with the Pratt method for dealing with ties (also called Pratt's test) is recommended (Derrick & White, 2017). Pratt's test performs better in discrete uniform distributions (Conover, 1973). Pratt's test does not ignore pairs without a difference between rounds. In contrast, the original Wilcoxon test does ignore pairs where there is no difference between rounds.

4 Findings

This chapter comprises the findings of the four sub questions. The literature review of the first three sub questions served as input for SQ4 and was verified via an interview. Finally, a three round Delphi study was conducted to address SQ4.

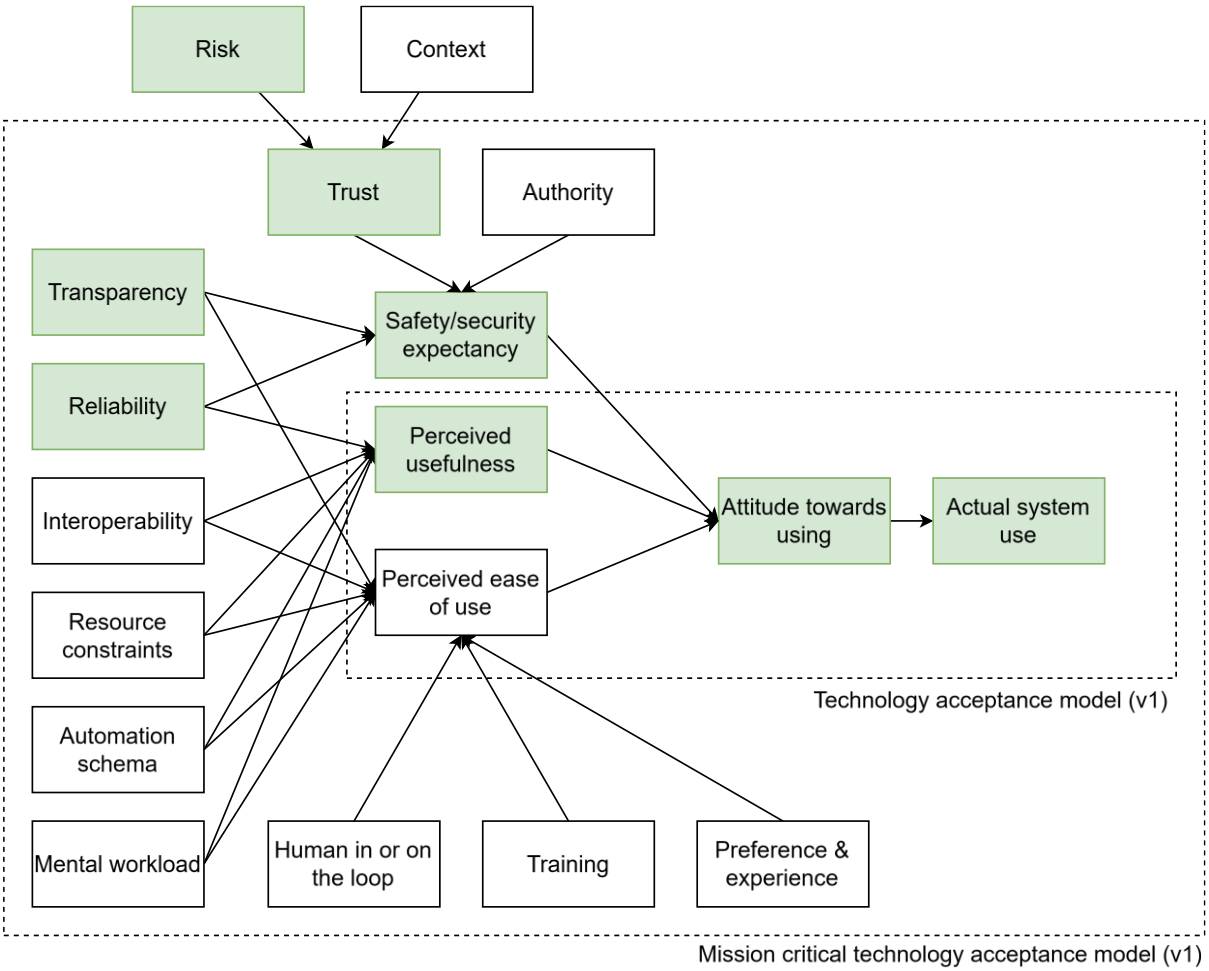
4.1 What Factors Are Associated With Acceptance of AI Advising Humans on Which Actions to Take? (SQ1)

The first sub question (SQ1) was investigated via two methods. First an exploratory literature review was performed, the result of which follows narratively below. Section 4.4 shows how the literature review of SQ1, SQ2, and SQ3 was verified by interviewing one CISO.

The MMCTAM has been used as a starting point for the literature review for SQ1 and SQ2. The following concepts from the MMCTAM are considered factors for SQ1: *risk*, *trust*, *transparency*, *safety/security*, *reliability* and *perceived usefulness*. They have been highlighted in Figure 9.

Figure 9

Accented Modified Mission Critical Technology Acceptance Model (MMCTAM)



Note. The factors in green are the most applicable for the target participant group of the Delphi study (CISOs).

Concepts from the MMCTAM that solely influence the *perceived ease of use* have not been included for SQ1. They are less applicable to the target participant group of the Delphi study (CISOs). CISOs focus on strategic objectives, not ease of use of specific implementations. *Interoperability* entails integration with other systems, which is related to system implementation as well. The concept of *resource constraints* entails training and available personnel for a specific AI system and therefore has not been included for SQ1. *Automation schema* from the conceptual model is referred to as *data error* in Weger et al. (2023); it concerns wrong input into an AI system. For this study the concept *reliability* covers inaccurate output of AI systems and incorrect input is not considered separately. *Mental workload* encompasses the mental workload of a human that is using an AI system, which has been considered a property of the implementation phase. The concept *Context* has not been used as this study is limited to the financial sector which corresponds to a single context when compared to the three contexts Liehner et al. used.

Kelly et al. (2023) performed a literature review on the acceptance of artificial intelligence. Most of the reviewed studies defined AI as Artificial Narrow Intelligence (ANI) and researched existing AI technology. Those papers are out of scope for SQ1 as they involve broadly accepted tooling: ANI systems operate within a predefined environment to perform a specific task and do not generalize (Samoili et al., 2021). The following papers however defined AI as either Artificial General Intelligence (AGI) or Artificial Super Intelligence (ASI): Meyer-Waarden and Cloarec (2022), Mohr and Kühl (2021), Prakash and Das (2021).

Kelly et al. referred to the Artificially Intelligent Device Use Acceptance model, which predicts consumer acceptance of AI devices (Gursoy et al., 2019). The Artificially Intelligent Device Use Acceptance involves customer emotion and social influence. Social influence means being influenced by the opinion of others, both socially as well as professionally (Prakash & Das, 2021).

Meyer-Waarden and Cloarec (2022) investigated adoption – by consumers – of autonomous vehicles. Autonomous vehicles operate without human intervention, which means full autonomy. Their results show that the performance-/effort expectancy, social recognition, well-being, hedonism, technology trust and security are enablers of autonomous vehicles. They found that privacy concerns are the main inhibitor. Social recognition is the elevation of one's social status when using a new product or technology. Performance expectancy is “the degree to which a technology is perceived as providing enhanced performance compared to the technologies currently adopted” (Rogers, 1983, p. 15; Solaimani & Swaak, 2023, p. 3).

Mohr and Kühl (2021) investigated the acceptance of AI in agriculture. They identified the significant factors to be perceived behavioral control and expectation of property rights over business data. Perceived behavioral control is the expectancy of the ability to perform a certain behavior (Sok et al., 2021). It is positively influenced by the innovativeness trait (Mohr & Kühl, 2021). A lack of control over the process of output being generated by agents is an inhibitor to acceptance (Kruse et al., 2022). The property right over business data concerns data sovereignty. Maintaining ownership of data increases the acceptance of AI systems.

Prakash and Das (2021) researched Intelligent clinical diagnostic decision support system (ICDDSS) adoption. An ICDDSS *assists* medical practitioners, although the participants of the study tend to believe that ICDDSSs can *replace* medical practitioners, specifically radiologists, in the future. On the automation-autonomy scale, this is HitL. They found that performance expectancy, effort expectancy, social influence, and initial trust in technology are key enablers for ICDDSSs. The main inhibitor is resistance to change, which is composed of inertia, perceived threat, medicolegal risk and performance risk. Medicolegal risk is specific to the medical context of their study.

The MMCTAM includes risk as a factor. The risks of operating LLMs include privacy violations of the users and of the information processed by the LLM, providing misinformation, bias and discrimination specifically, socioeconomic harm by job displacement and environmental harm caused by increased energy and resource consumption (Weidinger et al., 2022). When comparing decisions taken by AI and human expert, the decisions taken by AI on high-impact cases are considered less risky and more fair (Araujo et al., 2020). Therefore, risk is a factor influencing acceptance of AI and HitL.

Solaimani and Swaak (2023) investigated critical success factors for AI adoption and found that performance expectancy, top management support, and technical competencies and resources are the most critical. Technical competencies and resources refer to the expertise and know-how of

personnel, and the available IT infrastructure. Top management support refers to commitment and approval from the board for new initiatives and technologies (Abd Majid & Zainol Ariffin, 2021; Solaimani & Swaak, 2023).

The MMCTAM includes trust as a factor. Trust influences technology acceptance. Decisions are trusted more when a human is in the loop (Sele & Chugunova, 2024). Job displacement is an inhibitor to trust (Benedikt et al., 2020). However, people can overtrust unreliable AI in life-or-death decisions (Holbrook et al., 2024).

The MMCTAM includes human on the loop as a factor. When both AI agents and humans work together (either a human in or on the loop), agents and humans need to have a shared situational awareness. Humans need to understand what agents are up to, and agents do need to be aware of what humans are doing (Cleland-Huang et al., 2024; Fischer et al., 2021). AI agents having a high level of autonomy imposes a risk of deskilling, which is poignant considering the human operators need to be able to intervene and take over when the autonomous systems fail (Kosmyna et al., 2025; Veitch et al., 2024). Human oversight on algorithms requires: algorithmic transparency, accountability and explainability (Kyriakou & Otterbacher, 2023). Transparency and explainability can be achieved by rule-based cyber security solutions as opposed to black box models (Sarker et al., 2024). Accountability refers to the auditing of algorithms, data and design processes of AI systems (Barredo Arrieta et al., 2020). The question can be posed who bears the responsibility for the results of actions taken by autonomous agents, although humans are held to a higher moral responsibility than agents (Tolmeijer et al., 2022).

In the case of human in the loop, a human who can adjust recommendations made by an algorithm decreases decision accuracy (Sele & Chugunova, 2024). As mentioned before, LLMs can hallucinate. Hallucination introduces factual inaccuracy (Zhao et al., 2024).

The resulting list of potential factors has been summarized in Table B2 (Appendix B).

4.2 What Factors Are Associated With Acceptance of the Outsourcing of Cyber Security Response (to External Security Operation Centers)? (SQ2)

The second sub question was investigated via two methods. First an exploratory literature review was performed, the result of which follows narratively below. Next, the literature review of SQ1, SQ2 and SQ3, was verified by interviewing one CISO, see 4.4.

The response speed of outsourced security operations centers (SOCs) is lower than internal SOCs due to limited visibility, and delays after handing over remediating actions to the SOCs client (Kokulu et al., 2019). A faster detection of an attacker that gained access to the IT infrastructure will limit the time of the attack and reduce its impact (Brewer, 2019).

SOARs help in standardizing procedures (via playbooks), communication and reporting, which improves consistency of tickets. The accuracy/quality of tickets created using a SOAR is lower than those created without a SOAR. A SOAR can be dependent on a reliable internet connection. The efficiency of a SOC during investigations of security incidents is improved when using a SOAR (Bridges et al., 2023).

There is a shortage of qualified security personnel for SOCs (Brewer, 2019; Danquah, 2020; Vielberth et al., 2020). Top management support is a critical success factor for SOCs (Abd Majid & Zainol Ariffin, 2021; Vielberth et al., 2020). Forsberg and Frantti proposed a framework of metrics to measure the performance of a SOC (Forsberg & Frantti, 2023). The metric ‘technical accuracy of the analysis’ consists of incident classification, prioritization, escalation and accuracy of the conclusion of the investigation.

Vielberth et al. (2020) listed several factors which influence the choice of outsourcing a SOC: cost of implementing and maintaining versus outsourcing, time to set up a SOC internally, regulations including privacy regulations, availability, management support (referring to Abd Majid and Zainol Ariffin), data loss concerns and expertise. Note that data leakage is a risk for GenAI too (Rehberger, 2024). Cost, management support, and consistency in analysis and reporting are success factors for the establishment of a SOC (Abd Majid & Zainol Ariffin, 2019).

In their paper on the security assurance of cloud services Shukla et al. (2023) applied the security requirements of SINTEF (Bernsmed et al., 2015). They grouped requirements in the following categories: data storage requirements, data processing requirements, data transfer requirements,

access control requirements, security procedure requirements, incident management requirements, privacy requirements and layered services requirement.

In their paper on outsourcing information security operations, Cezar et al. (2017) looked into competitive externality. Competitive externality comes into play when one of a set of competitors within a sector experiences a security breach. If the advantage of being the only firm *not* being breached is larger than the loss when being the only firm *being* breached, the competitive externality is positive. When competitive externality is positive, outsourcing the detection function of information security operations is beneficial. When competitive externality is negative, outsourcing the prevention function is beneficial to a firm.

Risk interdependence is the likelihood of a security breach at one firm leading to a breach of a different firm (Cezar et al., 2017). They found that risk interdependence should be taken into consideration when outsourcing security observations. A positive risk interdependence exists when firms share information and establish trusted network connections. The compromise of one of the firms will affect the other(s). A negative risk interdependency exists when there is no spillover from a security breach at another firm; in that case investing in security measures will lower the likelihood of a breach because the other firm will become an easier target for hackers (Feng et al., 2021). Thus, we can infer that for firms in different sectors and different countries the risk interdependency converges to zero. Competitors will not all outsource to the same managed security services provider (MSSP) if the risk interdependency between the competitors will increase when outsourcing (Cezar et al., 2017). In case of both positive and negative risk interdependency a long-term relational contract with an MSSP could mitigate the risks of a double moral hazard. Double moral hazard refers to both the MSSP and the firm putting in less effort than stipulated in the contract (Feng et al., 2021).

Wu et al. (2024) extended the study of Cezar with the topic of information leakage. If the risk of information leakage is high, there is a lower tendency to outsource security operations.

The resulting list of potential factors has been summarized in Table B2 (Appendix B).

4.3 What requirements exist for AI usage within the Dutch financial sector? (SQ3)

The third sub question was investigated via two methods. First a literature review of regulations, professional and scientific literature was performed, the result of which follows narratively below. Next, the literature review of SQ1, SQ2 and SQ3, was verified by interviewing one CISO, see 4.4.

The regulations in the EU for cyber security have been evolving. Several laws and regulations applicable to the Dutch financial sector have or will be implemented between 2024 and 2027 as shown in Figure 10. These regulations, and the GDPR, have been reviewed.

Figure 10

NIS2, DORA, CRA, AI Act comparison

NIS2	DORA	CRA	AI Act
What is it: The Network and Information Security Directive (NIS2) replaces the original NIS. It aims to improve cyber security & resilience within the EU.	**What is it:** The Digital Operational Resilience Act (DORA) strengthens IT security of financial organizations.	**What is it:** The Cyber Resilience Act (CRA) applies security requirements for hardware and software products with a digital element.	**What is it:** The Artificial Intelligence (AI) Act applies risk-based rules & requirements to AI systems and their usage.
When will it apply: Each organization within scope of NIS2 must adhere to its requirements by Q4 2024.	**When will it apply:** Each organization within scope of DORA must adhere to its requirements by January 17, 2025.	**When will it apply:** Products within scope of CRA must adhere to its requirements by ~2027.	**When will it apply:** AI systems must adhere to to the requirements by around August 2, 2026, depending on the classification.
Who is in scope: All operators of critical infrastructure and essential services in the EU. NIS2 has 15 sectors of business & industry in scope.	**Who is in scope:** Financial entities in the EU such as banks, insurance companies, investment companies, and their third-party (IT) service providers.	**Who is in scope:** All producers and manufacturers of digital software and hardware products that are based in the EU or sell into the EU.	**Who is in scope:** All providers and operators of AI systems in the EU or providing services (using AI) within the EU.

Note. From “EU NIS2, DORA, CRA and AI Act” by Kirtley, 2024, *Threat-Modeling.Com* (https://threat-modeling.com/eu-nis2-dora-cra-and-ai-act/).

4.3.1 NIS2 and Cyber Resilience Act

NIS2 is a directive that needs to be transposed to national law (Paschou, 2024). Within the Netherlands NIS2 has been implemented as the Wbni act. Banking and financial market infrastructures are defined as sectors of high criticality in NIS2 (Directive 2022/2555 (NIS2), 2022). The definition of banking is broad; it applies to any business taking deposits or granting funds (Regulation 575/2013 (Capital Requirements Regulation), 2013). For financial institutions, both NIS2 and DORA apply, and DORA takes precedence (Paschou, 2024).

The Cyber Resilience Act (CRA) applies to the manufacturers and/or suppliers of digital software and hardware.

4.3.2 EU AI Act

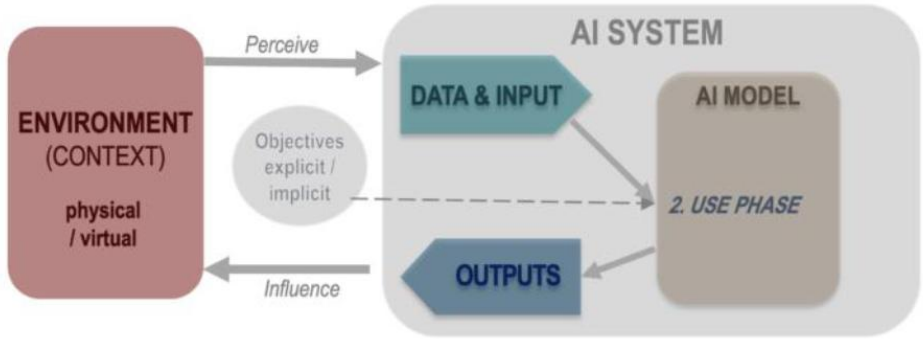
The EU AI Act is applicable to financial institutions within the EU. The AI Act defines AI *systems* along the lines of the Organisation for Economic Co-operation and Development (OECD) definition. The AI Act does not explicitly define 'AI' or 'AI models'; it only defines general purpose artificial intelligence (GPAI) models. The OECD defines AI as follows: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." (OECD, 2024, p. 4). In Article 3 of the AI Act, AI systems are defined as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments". AI systems and general-purpose AI *models* are distinctly defined (Joshi & Khalfi-Lanoux, 2024; Regulation 2024/1689 (AI Act), 2024, p. 46). An AI *model* is part of an AI *system* as shown in Figure 11.

The AI Act covers prohibited AI practices in Article 5; high-risk AI systems in Article 6; transparency requirements for AI systems that interact directly with natural persons, recognize emotions or categorize biometrics, or generate audio, image, video or text in Article 50. Other (low risk) applications of AI are out of scope of the AI Act. Applications that are not covered by the AI Act will fall under existing legislation (Parente, 2024; Passador, 2024).

Inference is defined in the AI Act as "the process of obtaining the outputs, such as predictions, content, recommendations, or decisions, which can influence physical and virtual environments, and to a capability of AI systems to derive models or algorithms, or both, from inputs or data" (Regulation 2024/1689 (AI Act), 2024, p. 4).

Figure 11

Updated OECD definition of AI system



Note. From "Explanatory memorandum on the updated OECD definition of an AI system", 2024, *OECD Artificial Intelligence Papers (8)*, p. 7 (https://doi.org/10.1787/623da898-en).

AI systems used within a single entity in the financial sector are not part of the critical digital infrastructure (Directive 2022/2557, 2022). Such AI systems scoped to the cyber security of the organization (and not its clients) are not considered high-risk AI systems per Article 6 and none of the areas listed in Annex III of the AI Act are applicable: biometrics, critical infrastructure (as explained above), education, employment, law enforcement, migration or justice/democratic processes (Regulation 2024/1689 (AI Act), 2024).

The AI Act has less strict requirements for open-source models, excluding GPAI models (Liesenfeld & Dingemanse, 2024; Regulation 2024/1689 (AI Act), 2024). Open-source is defined within the AI Act, but that definition is not used consistently by others; many LLMs are not (fully) open-source (Gibney, 2024).

GPAI according to the AI Act includes models of over a billion parameters and/or multi-modal models, for example generative AI (Parente, 2024; Passador, 2024; Regulation 2024/1689 (AI Act), 2024). Multi-modal models are models that support multiple modalities, for example text, images and/or audio as input and/or output. Obligations for GPAI include transparency and technical documentation according to Recital 101 of the AI Act. When a GPAI model is built by an organization, is solely used for internal processes and does not affect the rights of natural persons, the obligations for GPAI however do not apply according to Recital 97. However, if a firm is using an externally sourced AI system as part of cyber security AI tooling, then the obligations for GPAI apply. From Article 97: “It should be understood that the obligations for the providers of general-purpose AI models should apply once the general-purpose AI models are placed on the market.”.

Recital 100 states an AI system is considered a GPAI *system* when a GPAI model is integrated into an AI system. A GPAI model that performs more than 10^{25} floating point operations is considered a GPAI model with systemic risk according to the currently defined thresholds in Article 51 of the AI Act. The consequences of such a classification, listed in Article 55, mainly entail stricter governance.

As stated by Recital 105, copyright restrictions do apply on training material for GPAI models (Regulation 2024/1689 (AI Act), 2024).

4.3.3 GDPR

The GDPR is applicable within the EU. Article 22 states that “the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her” (Regulation 2016/679 (General Data Protection Regulation), 2016, p. 46). Article 22 applies if the decision making has significant effects on the data subject. That is, if it affects the autonomy or (economic) freedom of the data subject, being a natural person (Wiedemann, 2022).

Automated decisions could be based on profiling the activities of people. For example, to detect suspicious activity of an employee’s computer, a baseline of regular activities would need to be established (Olabanji et al., 2024). An automated action performed by an AI tool could depend on both profiling of a data subject, and making an automated decision, which can be considered two separate steps (Wiedemann, 2022). Automated actions will generally be based on inferences. The regulation in the EU surrounding inferences is still under development (Silveira, 2023).

Hamon et al. (2021) argue that for more advanced machine learning models it is impossible to provide explanations of decisions that can be comprehended by a layman. Without an explanation that is understandable, such models do not meet the transparency requirement of Article 22 of the GDPR.

If Article 22 applies, WP29 suggests that the data subject should be provided with: i) the categories of data that have been or will be used in the profiling or decision-making process; ii) why these categories are considered relevant; iii) how any profile used in the automated decision process is developed, including any statistics used in the analysis; iv) why this profile is relevant to the automated decision process; and v) how it is used for a decision concerning the data subject (Silveira, 2023).

4.3.4 DORA

Critical third party service providers are defined in Article 31 (Regulation 2022/2554 (Digital Operational Resilience Act), 2022). Article 31 does not set specific limitations on the usage of these providers in Recital 67 but does identify a concentration risk. Considering the current GPAI providers (e.g., Microsoft/OpenAI, Google, Facebook, Anthropic), the GPAI marketplace can suffer from the same concentration risk. An example would be many financial institutions within the Netherlands

using IT services of Amazon, Google or Microsoft and their AI services (De Nederlandsche Bank, 2024; Dutch Authority for the Financial Markets, 2024). These large tech companies also invest in and provide GPAI services.

If GPAI models and systems are used within the cyber security defense of financial institutions, then the recommendations for the security of these systems that are set in Article 35(1) apply. These recommendations include rollouts of patches and updates. Regular patching of third-party AI systems could affect the cyber security tooling that uses/depends on these systems.

DORA also requires implementation of processes to test ICT systems in Recital 56, implementation of monitoring third party service providers and their ability to deliver secure services in Recital 68, implementation of policies to assess vulnerabilities and apply software patches in Article 9 (4). Both third-party AI systems and first party AI systems therefore require testing.

To illustrate the market concentration surrounding foundation models, the UK Competition and Markets Authority identifies Amazon, Apple, Google (Alphabet), Meta, Microsoft and Nvidia – abbreviated as GAMMAN – as being involved in development of foundation models (Competition and Markets Authority (UK), 2024). Triguero et al. (2024) listed LLMs from OpenAI (referring to Microsoft), Google and Meta as the most relevant. Foundational models are “trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks” (Bommasani et al., 2022, p. 3).

To summarize, regulation regarding automated decision-making using inferences and/or deep learning is still a developing field within the EU. The AI act only covers a subset of automated decision making, and AI-powered cyber security tooling falls outside its scope when customers are out of scope. The resulting list of potential factors has been summarized in Table B6 (Appendix B).

4.4 Interview

On March 6, 2025, a CISO of a Dutch financial institution has been interviewed. The hour-long interview did not identify new factors. The interviewee confirmed that the concentration risk related to AI tooling is relevant, with low importance. The interviewee considered one factor irrelevant: security externality.

The list of factors to be included in the Delphi study was prioritized based on the interview: A high importance corresponds to a high priority for inclusion in the Delphi study, see Table B4 (Appendix B).

4.5 Delphi First Round

In a thematic qualitative analysis, the following themes were identified: AI enabled cyber security tooling can have advantages; AI enabled cyber security tooling can have disadvantages; the human factor plays a role; requirements for AI tooling; and finally, there are moderating factors involved. There is significant overlap with the factors previously identified in this study, which will be illustrated below.

Participant 3 (P3) and P7 mentioned cost savings in response to Question 9 of scenario C, even though cost savings was not included as a factor.

4.5.1 Advantages of Autonomous AI Cyber Security Tooling

The accurate response by AI tooling is described by P5 as “in an unbiased fashion”. P1 notes an “analysis based on the full picture”. Regarding HitL, P3 states that *humans* can make biased decisions and P5 agrees. P5 also states that *humans* are error-prone implying AI tooling is not. However, P8 disagrees by stating “decisions from an AI system may be wrong due to bias, hallucinations, etc.”.

AI tooling improves detection and prevention of security events and incidents. AI tooling can prevent data being exfiltrated by threat actors according to P1 and P4, improve threat detection according to P3 and help detect data exfiltration according to P5 and P8. Autonomous action after detection is explicitly mentioned by P8: “autonomous blocking based on changing exfiltration patterns”. P4 mentions the AI tooling can detect a malicious insider and threat actor’s usage of AI.

Most participants agree that AI tooling will improve the response speed (i.e., respond quicker) and improve efficiency during investigations. The improved speed is called out by P1, P2, P3, P5, P6 and P8. P2 extends this to “AI can improve productivity”. One reason for improved efficiency is 24/7 availability of AI tooling as mentioned by P4, P5, P6 and P8. Another reason is being able to scale as P3 puts it “AI can analyze large volumes from multiple sources”.

Cost saving is mentioned by P3 and P7, and P4 sees it as a possibility. P6 and P8 note that AI tooling could help with the shortage of skilled cyber security personnel. Related, P4 and P7 mention that AI tooling can reduce the workload of humans, as P7 states “less repetitive and uninteresting work”.

4.5.2 Disadvantages of Autonomous AI Cyber Security Tooling

AI tooling lacking knowledge is a disadvantage. As P6 puts “The AI defence first needs to 'learn' new defensive techniques that are not yet in the model.” or 2) it has no information about an attack in its training data as mentioned by P2 and P7. P4 expands on this by stating AI tools of nation state actors might always be one step ahead of the AI tools of defenders.

A lack of accuracy can result in impact on the business processes. As P5 phrases it, “[a blackbox AI tool] can run down a complete critical process”.

4.5.3 Human Factor Concerning Autonomous AI Cyber Security Tooling

Four participants do not agree with the premise of this study that AI tooling can respond autonomously, or they require HitL in some cases. According to P1 there are “some cases where a human operator always needs to intervene”. P2 requires a human in the loop for critical production systems: “Ability to have a human in the loop”. P4 thinks that only “some action should or could be done automatically by AI tooling”. P8 states: “AI systems are not yet capable to provide the sufficient cyber defense quality to rely on. Human analysis remains necessary.”

The possibility for human operators to veto decisions of AI tooling has the downside of reducing the response speed and efficiency of the AI tooling, negating the advantage of quicker response and efficiency during investigations. P3 mentions “reduced efficiency”. P6 agrees: “Little advantage over normal SIEM use cases with regards to efficiency and speed”.

Another downside of HotL is lack of situational awareness. P3 recognizes a “potential for inconsistent or biased decision-making”. P6 expands by saying that “analysts might not be able to properly deduce why the AI agent detected a certain attack pattern”. P7 and P8 share this view.

4.5.4 Moderating Factors for Autonomous AI Cyber Security Tooling

Several participants suggest putting restrictions on the usage or scope of the AI tooling. P3 mentions AI tooling should have clear boundaries. P7 similarly states “making clear the mandate of the AI agents”.

The potential impact of autonomous actions on the business is highlighted by three participants. P5 mentions about critical production systems that “the operational impact of a decision must be defined in detail to maintain control”. P6 words it differently: “Ability to limit options to isolation or only monitoring due to potential business impact.” P7 is draws a clear line: “limiting and cutting down operations is not [ok]”.

Both P2 and P5 note that HotL would help in training the AI. P2 adds that being able to train the AI is a requirement to run it on production systems.

4.5.5 Requirements for Autonomous AI Cyber Security Tooling

Regarding the accountability cluster, all eight participants agree that the AI tooling or AI agent is not responsible for the actions taken. P2, P3 and P7 stress that all actions taking by AI tooling must be logged and monitored; P3 phrases this as auditability.

Usage of autonomous AI might be restricted in the Dutch financial sector due to regulations. P3 states “financial entities are subject to stringent regulations” and P8 states “Autonomous AI systems may be forbidden by law”.

4.5.6 Overlap With Identified Factors

The following previously identified factors were recognized by the participants:

- Accountability & Responsibility.
- Accuracy (hallucinations) & Technical accuracy and consistency of analyses.
- Auditing of services.
- Efficiency during investigations.
- Explainability, more specifically the lack thereof.
- Implementation and maintenance cost, including time spent.
- Lower response speed.
- Shared situational awareness between human operators and AI.
- Shortage of skilled personnel.
- Transparency.

4.5.7 Potential New Factors

The following categories could *not* be related back to the previously identified factors. Only categories that were mentioned by at least two participants have been considered.

- Ability to train AI tooling.

4.6 Delphi Second Round

All eight participants took part in the Delphi second round. Their responses are shown in Figure 12.

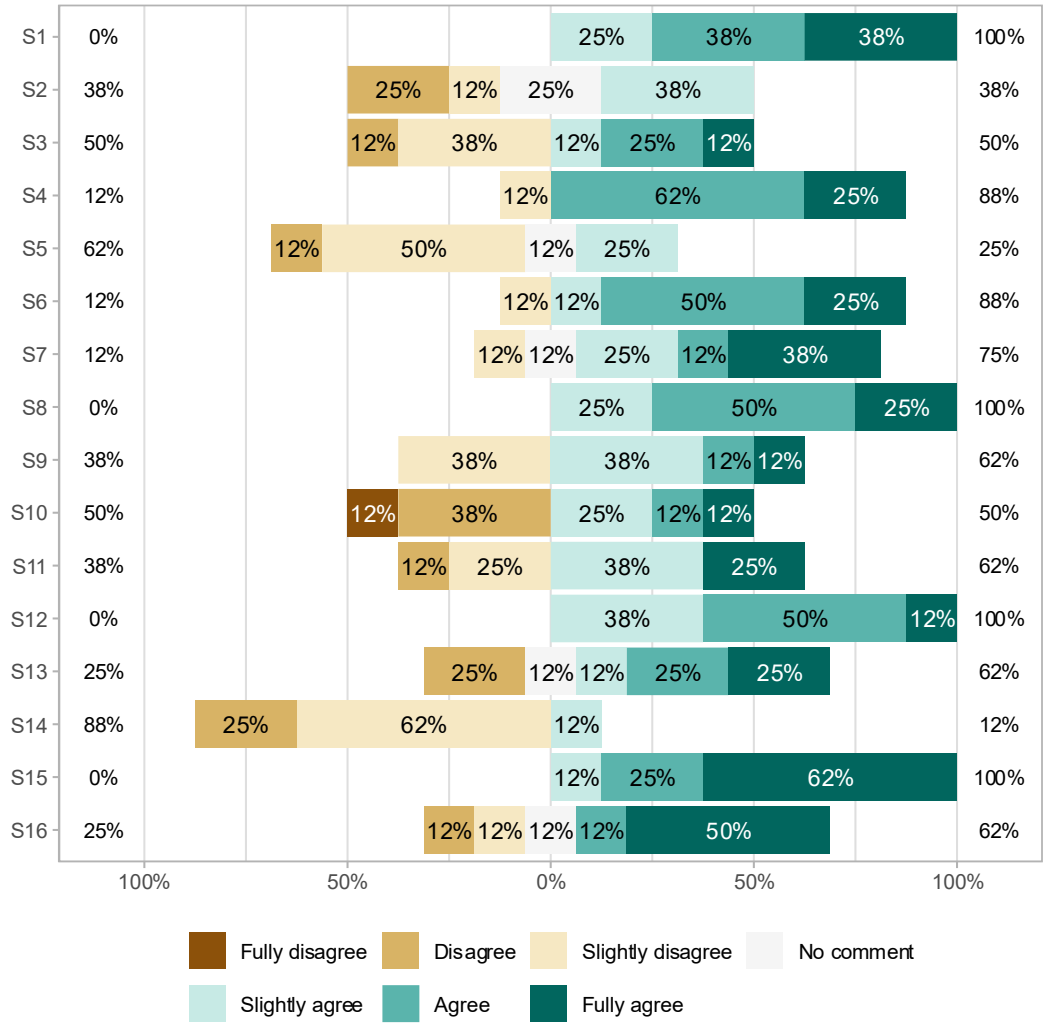
As shown in Table B8 (Appendix B), a preliminary consensus has been reached for nine out of 16 statements. Note that algorithm R-6 is the worst case. Since a consensus was not reached on all statements, a third round was performed.

P1 and P7 agree that AI can assist in decision making, because the term *support* used in Statement 1 rules out autonomous decision making. P5 adds that even though decisions are made by human operators, the output of the AI security tooling can indirectly cause impactful wrong decisions.

Regarding working effectively with AI tooling, Statement 4, P7 envisions humans focusing on the longer term. P1 thinks knowledgeable and experienced staff are needed to align AI actions with policies, and to review edge cases. P5 stresses that knowledge about AI is important and suggests requiring personnel to get certified. On Statement 12 P5 doubts if a human analyst is “equipped to verify the claims and logic” of an AI, reiterating that is exactly their point on Statement 15.

Figure 12

Round 2 responses



Note. The statements (S1-S16) are plotted on the y-axis. Agreement is plotted on the x-axis, from fully disagree to fully agree, centered around the *No comment* option. Responses to Statements 2, 5, 7, 13 and 16 included *No comment*. That option is not counted to the total of either agreement or disagreement.

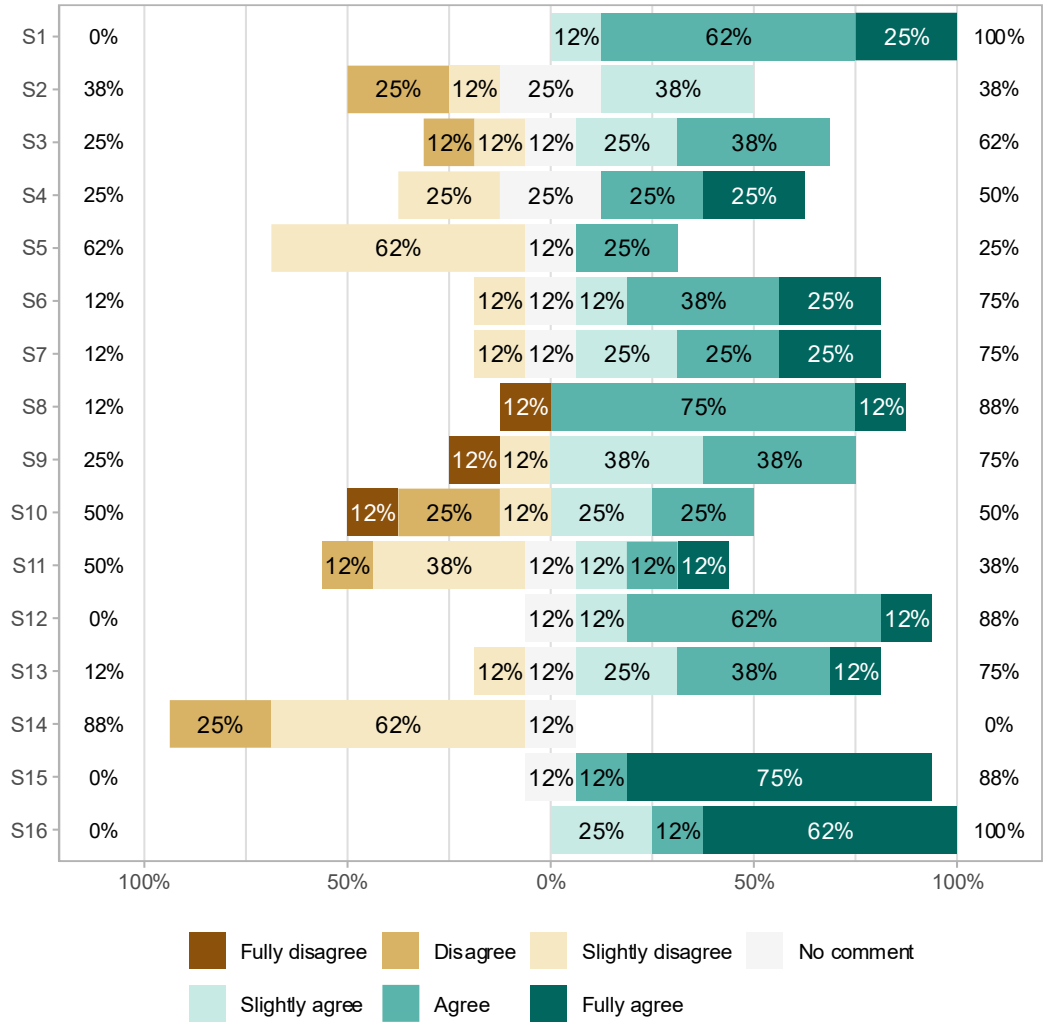
P3 disagrees with Statement 5 and 9, and agrees with Statement 13, by stating that “human interpretation is still key to make the final conclusion” and that AI should “speed up decision making and come with recommendations”. P1 repeats the need for HitL for Statements 8 and 9. P8 mentions on statement 13 that “major security incidents will always be analysed by humans.” This is in line with the response of P1, P2, P4 and P8 in Round 1. P5 has a different view on having to depend on fully autonomous AI including critical production systems, by extrapolating the speed AI is evolving: “I would not be surprised”. On statement 14 and 16, P5 states that only for mission critical impact human review is required.

4.7 Delphi Third Round

All eight participants took part in the Delphi third round. Their responses are shown in Figure 13.

Figure 13

Round 3 responses



Note. The statements (S1-S16) are plotted on the y-axis. Responses to Statements 2 to 7 and 11 to 15 included *No comment*. That option is not counted to the total of either agreement or disagreement.

The stability between the second and third round is measured using the Wilcoxon signed rank test with the Pratt method for dealing with ties. The Wilcoxon test needs at least 6 linked pairs for a $p < .05$. Statement 2 unfortunately has only five linked pairs, so the stability cannot be determined. As shown in Table 2, we fail to reject the null hypothesis for all other statements, therefore stability has been achieved for all statements except Statement 2.

As shown in Table 3, a consensus has been reached for eight out of 16 statements, based on the R-6 algorithm. That is one less statement than the previous round. Note that if Hyndman (R-8) were used, consensus would be achieved for 10 statements compared to nine in the previous round.

Table 2

Wilcoxon matched pairs ranked test on Round 2 versus Round 3

Statement	Z [a]	p
1	0.000	1
2	-	-
3	0.000	1
4	0.128	1
5	-1.732	0.25
6	-1.000	1
7	0.577	1
8	-0.410	1
9	0.152	1
10	0.000	1
11	1.414	0.5
12	-1.414	0.5
13	-0.686	0.75
14	1.000	1
15	-1.732	0.25
16	-1.410	0.5

[a] *Z* has been rounded to three decimal places.

Table 3

Round 3 median and IQR per statement

Statement	Median [a]	IQR (R-8) [b]	IQR (R-6) [b]
1. In 2025, AI-driven cyber security tooling can support the decision-making of cyber security personnel.	5	**0.58**	**0.75**
2. Laws and regulations block the implementation of autonomous AI-driven cyber security tooling in the Dutch financial sector. [c]	3.5 [c]	2.00 [c]	2.00 [c]
3. The implementation of AI-driven cyber security tooling needs to lead to operational cost savings.	4 (+0.5)	1.83	2.00
4. Cyber security personnel needs to be more knowledgeable and experienced to work effectively with AI-driven cyber security tooling.	5	3.00	3.00
5. The analysis of AI-driven cyber security tooling is more accurate than the analysis of cyber security personnel.	3	**1.67**	2.00
6. AI-driven cyber security tooling needs to be continuously trained on the organization's business and IT information to be effective.	5	**1.67**	2.00
7. The AI-driven cyber security tooling has an inherent information leakage risk.	5	1.83	2.00
8. In 2028, your organization has to depend on autonomous AI-driven cyber security tooling to handle the volume (i.e., scale) of cyber attacks.	5	**0.00**	**0.00**
9. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on production systems in your organization.	4	**1.58**	**1.75**
10. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on critical production systems, like payment and banking.	3.5 (+0.5)	2.58	2.75
11. A review of autonomous response actions by a human SOC analyst is always required.	3 (-1)	1.83	2.00
12. Less than 25% of systems and services are out of bounds for autonomous AI-driven cyber security tooling.	5	**0.00**	**0.00**
13. The analysis of major security incidents is performed by AI-driven cyber security.	5	**1.00**	**1.00**
14. The AI shift is capable to perform all detection and response activities fully autonomously.	3	**0.83**	**1.00**
15. Human SOC analysts must be able to understand the decisions of the AI tooling.	6	**0.00**	**0.00**
16. In 2028, response actions that can disrupt business operations require human approval at all times (24/7).	6	**1.58**	**1.75**

[a] Difference compared to Round 2 between brackets if the median changed.

[b] IQR ≤ 1.75 in bold.

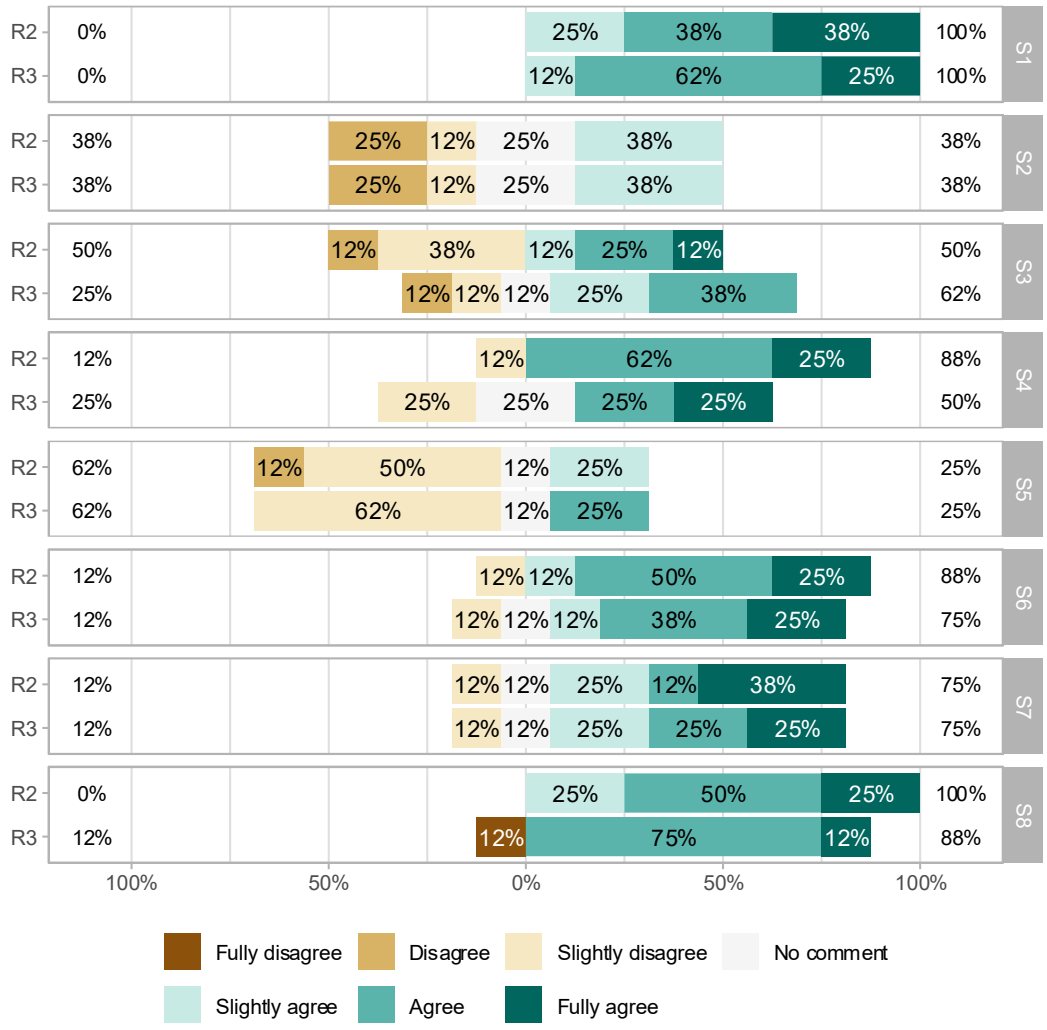
[c] Stability cannot be tested for Statement 2.

There are no big shifts from agreement to disagreement or vice versa between the rounds, see Figure 14 and Figure 15. The only exception is Statement 11 where the median changes from *Slightly agree* to *Slightly disagree*. However, no consensus has been reached on Statement 11.

The number of *No comment* entries rises from six to 13. The main contributor to this is P5, who chose *No comment* 11 times in the third round, see their remark above.

Figure 14

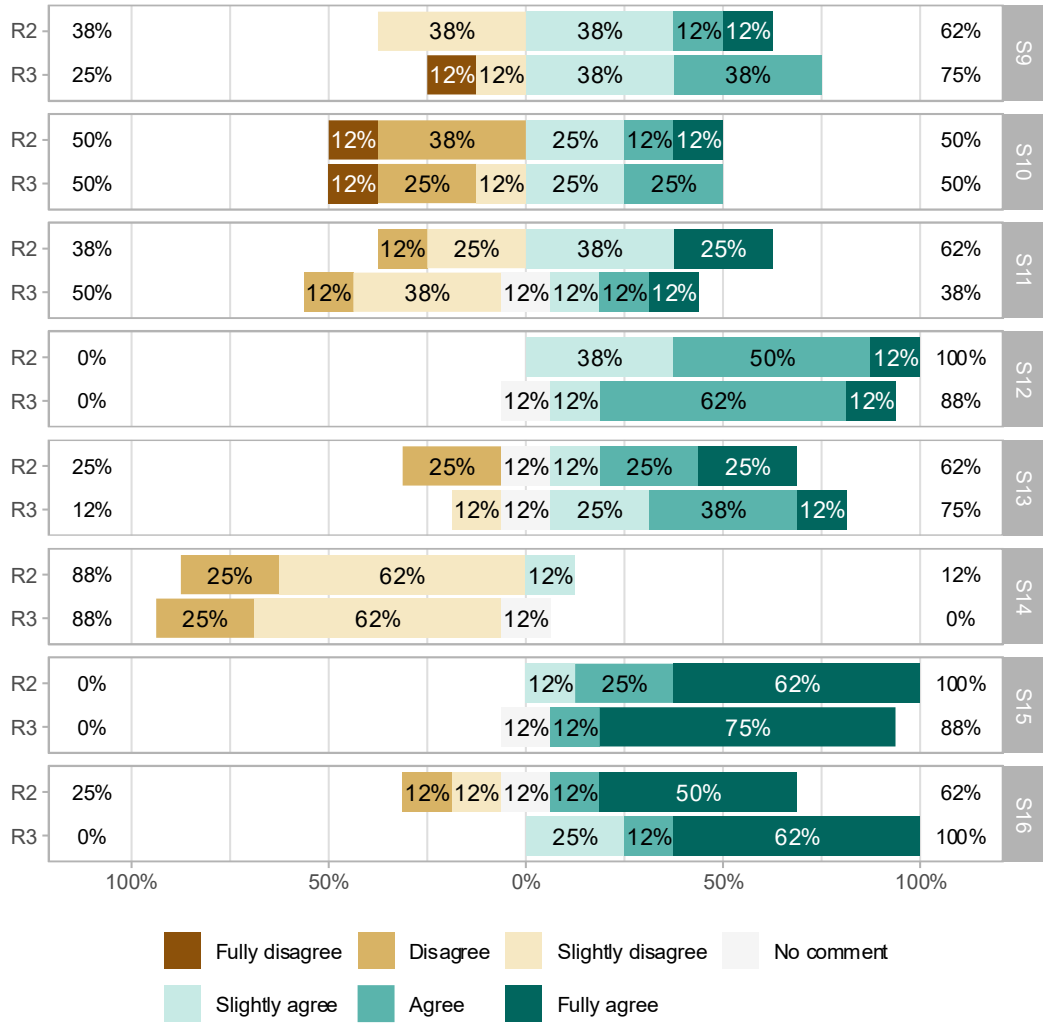
Round 2 and 3 comparison Statements 1 to 8



Note. The rounds (R2, R3) and statements (S1-S8) are plotted on the y-axis. Responses to Statement 2 do not appear to change, but as Figure C2 (Appendix C) shows several participants did rate this statement differently in the third round.

Figure 15

Round 2 and 3 comparison Statements 9 to 16



Note. The rounds (R2, R3) and statements (S9-16) are plotted on the y-axis.

All Delphi statements have been positively worded. However, some statements are positive about the application of AI in cyber defense and its capabilities, while others are not, see Table B9 (Appendix B).

Comparing the medians of statements across the three attitudes (positive, neutral, negative) towards AI as shown in Table 4, no clear pattern emerges. The participants agree the most with a neutral attitude towards AI. However, this is merely an observation. No statistics were calculated for this, except the average of the medians including and excluding statements where no consensus was reached. Since there is only one statements with a negative attitude towards AI with consensus, a comparison based on that single statistic will not be performed.

Table 4

Delphi statements attitude towards AI in cyber defense

Value	Positive statements [a]	Median	Neutral statements [a]	Median	Negative statements [a]	Median
	1.	5	3.	4	2.	3.5
	5.	3	4.	5	7.	5
	9.	4	6.	5	11.	3
	10.	3.5	**8.**	5	**16.**	6
	12.	5	**15.**	6		
	13.	5				
	14.	3				
Average		4.07		5		4.38
Average for IQR ≤ 1.75		4.4		5.5		6

[a] IQR ≤ 1.75 in bold.

Zooming in on the statements where stability and consensus was reached, the consensus is skewed towards agreement, see Table 5. This includes statements with a neutral or negative attitude towards AI. The only statements where full agreement is achieved on are a statement with neutral and negative attitude towards AI.

Table 5

Statements with consensus and their attitude towards AI

Statement	Consensus	Attitude towards AI
1. In 2025, AI-driven cyber security tooling can support the decision-making of cyber security personnel.	Agree	Positive
8. In 2028, your organization has to depend on autonomous AI-driven cyber security tooling to handle the volume (i.e., scale) of cyber attacks.	Agree	Neutral
9. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on production systems in your organization.	Slightly agree	Positive
12. Less than 25% of systems and services are out of bounds for autonomous AI-driven cyber security tooling.	Agree	Positive
13. The analysis of major security incidents is performed by AI-driven cyber security.	Agree	Positive
14. The AI shift is capable to perform all detection and response activities fully autonomously.	Slightly disagree	Positive
15. Human SOC analysts must be able to understand the decisions of the AI tooling.	Fully agree	Neutral
16. In 2028, response actions that can disrupt business operations require human approval at all times (24/7).	Fully agree	Negative

5 Discussion

First, the considerations for selecting relevant factors from the literature that were later used in the Delphi round are presented. Second, the findings of the Delphi study are discussed.

5.1 Relevant factor selection

This section describes the considerations for including or excluding factors from the findings of SQ1, SQ2 and SQ3. The resulting list of potential factors has been summarized in Table B2 (Appendix B).

5.1.1 SQ1

Discrimination has been considered part of *privacy concerns*.

The Artificially Intelligent Device Use Acceptance involves customer emotion and social influence and therefore is not a better fit than the MMCTAM for this study, as this study focuses on cyber security defense in a business context. *Social influence* is a better fit than social recognition in a business context. Well-being and hedonism are specific to the consumer context, and therefore out of scope for this study.

5.1.2 SQ2

The expertise of SOC personnel has been considered part of factor *shortage of skilled personnel*. Information leakage has been considered part of the *data loss risk* factor.

The security requirements of SINTEF that are applicable to SQ2 are *isolation: customer data should be isolated from other customers; auditing of services; security logging, detection, response and recovery*.

5.1.3 SQ3

DORA covers the regulation for AI in cyber security, NIS2 adds no additional requirements. The risks involving software and hardware from the CRA are already covered by DORA's requirements for third party management. From the DORA regulation, *concentration risk* could apply to cyber security AI tooling. Requirements *regular patching* and *regular testing* are focused on IT operations and have not been included.

This study focuses on autonomy and not on automation, therefore we shall assume that (some of) the AI systems that support autonomous AI tools and applications have the capability to infer, and the AI Act applies.

The transparency requirements from Article 50 of the AI Act could be applicable to autonomous AI tooling for cyber security**.** Externally sourcing AI systems for cyber security could add additional requirements from the AI Act. The *additional requirements for general-purpose AI systems with systemic risk* could apply to cyber security AI tooling, as well as *Copyright restrictions*. It was unclear as of March 2025 to what extent the AI Act would apply to cyber security tooling.

The GDPR *requirement to explain automated decisions* could apply to cyber security AI tooling. Cyber security AI tooling used by financial institutions will not directly use customer data, as opposed to fraud monitoring. Therefore, Article 22 of the GDPR should not apply.

5.2 Delphi Study

This study finds that there *is* trust in AI technology. There are, however, doubts about its reliability. P5 stated that they are worried about the irrelevant or false statements [Gen]A produces. Participants slightly disagreed with AI performing more accurate analyses than cyber security personnel. An expected lack of reliability inhibits acceptance within systems where it could disrupt business processes. This corresponds with a risk-based approach that is common in the financial sector. There is no complete trust in AI yet, which is understandable considering the pace of its developments in the past two to three years. P5 stated that they think the developments of AI in the past few months have been faster than anticipated. Note that there are two months between the second and third round participation of P5. The lack of complete trust is in line with Sele and Chugunova (2024) who found that trust is higher when a human is *in* the loop.

The highlighted factors in the MMCTAM are relevant to the acceptance of autonomous AI-driven cyber security tooling in the Dutch financial sector as shown in Table B7 (Appendix B), except *risk*. However, as stated before, Liehner et al. (2022) found that automation is task-dependent: there is less reliance on automation in a higher risk context. Participants P5, P6 and P7 stated in Round 1 of the Delphi study that impact on operations is not accepted, which was confirmed in the Round 3 by full agreement to Statement 16 (see Table 3 for the statements).

No consensus was reached on whether a human SOC analyst needs to review every autonomous intervention in a human on the loop scenario. However, the expectation is that autonomous response will be needed in the next three years due to a rise in the volume and scale of attacks. If that expectation becomes reality, there would not be enough human capacity to review all autonomous actions. As P3 and P6 stated in the first round, human review reduces efficiency and speed. Their observation is in line with Brewer (2019) stating that faster detection will reduce the impact of an attack. Based on Brewer (2019) it could have been expected that the participants would disagree with the statement that human SOC analyst needs to review every intervention. One reason for lack of consensus could be that Statement 11 should have been worded differently and focused on low impact autonomous actions. As P1 stated for Round 2, fully autonomous interventions in 2028 can be expected but only in relatively low impact use cases. Another potential reason is the beforementioned doubts about reliability. Participants may have associated ‘AI’ primarily with GenAI. Fully securing and ensuring robustness of GenAI systems and services is challenging (Beurer-Kellner et al., 2025; Ferrag et al., 2025). For cyber security, rule based AI might indeed be a safer approach for impactful decisions (Sarker et al., 2024). On the detection side GenAI is still useful though: intrusion detection, vulnerability detection, phishing detection, threat intelligence and malware classification (Ferrag et al., 2025)

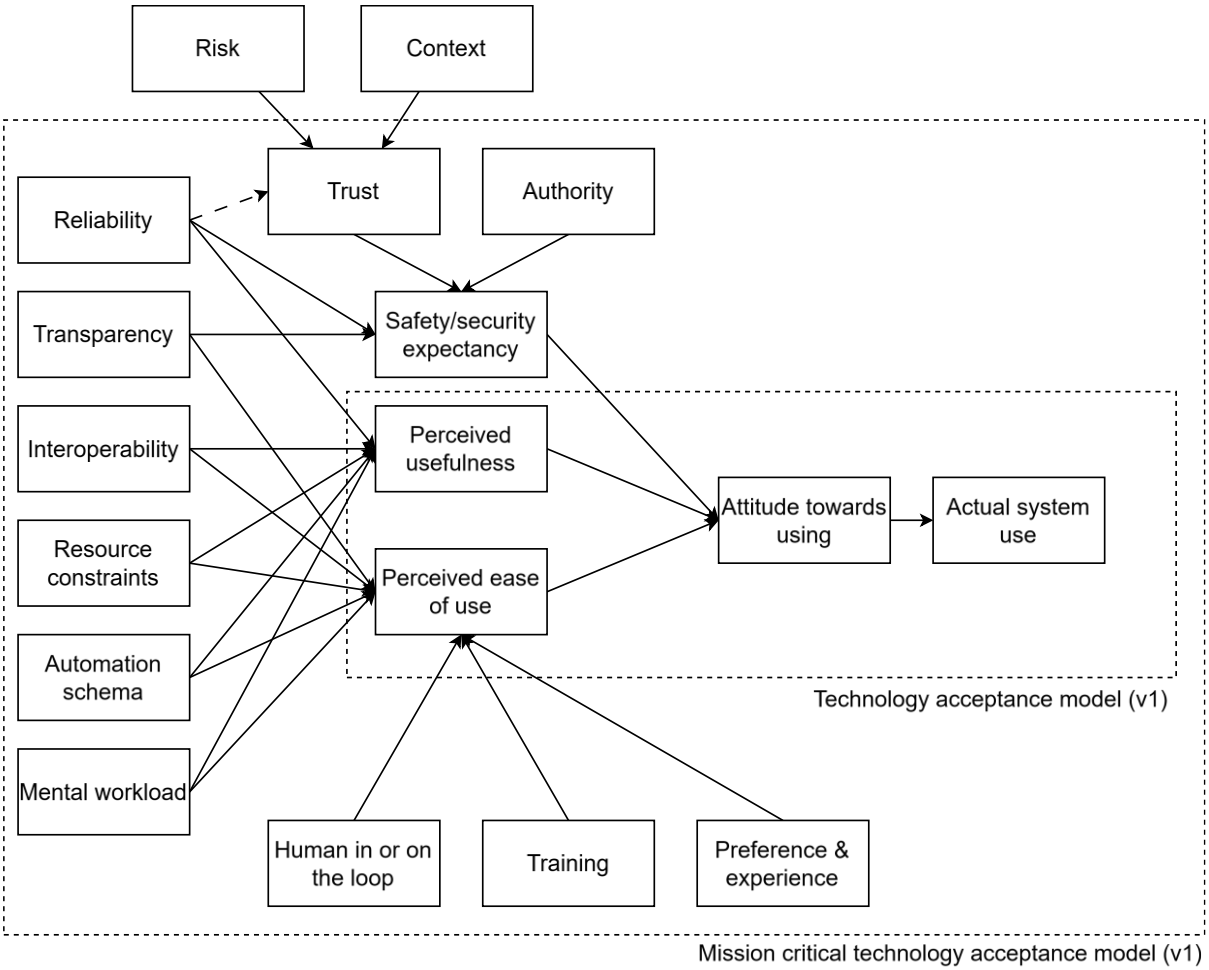
One of the key findings of this study is that humans must be able to understand the decisions of AI tooling, which is in line with Shreeve et al. (2023) and Kyriakou & Otterbacher (2023). The interpretability of GenAI systems is an active area of research (Ameisen et al., 2025; Kuiper et al., 2022). However, will GenAI systems reach a point where its results are explainable? Rudin (2019) might be right: If GenAI systems remain black boxes, we should not try to explain them but create and use models that are interpretable instead. That would rule out GenAI for autonomous decisions, but that would also rule out autonomous cyber defense agents as described by Kott (2023). (Note that Kott uses defense in a military context, but the concept of such an agent is transposable to cyber security.) P1, P3, P5 and P8 referred to the need for unbiased decisions by AI tooling. It is difficult to ascertain whether AI decision tools are fair and unbiased (Flechais & Chalhoub, 2023). This adds another perspective to the required explainability of AI systems. Financial institutions will need qualified personnel to be able to assess the explainability, interpretability and bias of internally developed or sourced AI systems. Practical guidelines should be developed, based on the extant literature and additional research.

According to the MMCTAM there is less reliance on autonomous AI in a higher risk context: a higher risk lowers trust. This counters the finding of Weger et al. (2023) that the level of autonomy does not influence acceptance of autonomous systems. It confirms the finding of Hauptman et al. (2023) that a higher level of autonomy is accepted when autonomous actions have less impact. This is in line with the typical risk-based approach in the financial sector: lower risk situations require fewer (security) controls. However, CISOs in the financial sector might always think risk-first and never opportunity-first, which would deviate from the findings of Shreeve et al. (2020). One of the opportunities for AI tooling is cost savings. This was mentioned by P3, P4 and P7 in Round 1, but no consensus was reached in Round 3. This supports the observation that CISOs think risk-first.

This study finds that the reliability, especially the accuracy, of autonomous AI tooling is being questioned according to Statement 5. As this Delphi study uses future scenarios, these doubts about reliability are not related to a specific product. It is a perception about AI tooling in general, which influences the trust in AI-powered technology in general. Therefore, the MMCTAM should be amended, see Figure 16: *reliability* influences *trust*.

Figure 16

Modified Mission critical technology acceptance model (MMCTAM) final version



Note. The dashed arrow from *Reliability* to *Trust* is the only difference with the original MMCTAM in Figure 6. *Reliability* and *Transparency* have switched position for legibility.

P3 expects that in the next three to five years a human will be required to be in the loop due to regulatory pressure, but this was not confirmed in the results. Note that the AI Act might impose additional requirements to *sourced* general purpose AI (GPAI) systems. GPAI systems could also impose a concentration risk within the Dutch financial sector, although this study found no significant result on that.

Even though ethics is one of the Responsible AI principles (Barredo Arrieta et al., 2020), none of the participants referred to ethics explicitly. At first this is surprising. However, this study finds that black-box autonomous AI systems are not accepted in cyber security: The decisions of AI tooling need to be transparent and explainable to humans. The ethical principle of explicability encompasses transparency and explainability (Formosa et al., 2021). My tentative conclusion is that ethics as a term was not top of mind for the participants, but they have incorporated ethical principles in their professional life. An area for future research is the ethics of AI *for* cyber security, building on research from Formosa et al. and Barredo Arrieta et al. among others.

Privacy is another Responsible AI principle (Barredo Arrieta et al., 2020). Privacy concerns are an inhibitor to acceptance – by consumers - of autonomy (Meyer-Waarden & Cloarec, 2022). Privacy concerns were categorized as very high priority during the interview. However, privacy was not mentioned in Round 1 and was therefore not included in Round 2. LLMs do have privacy issues: they can leak personal data from their AI system or the model's training data (Zhang et al., 2025). Privacy might be an important topic, but my tentative conclusion is in the same line as for ethics: When your mindset is attuned to autonomous response, privacy is not top of mind and could be part of future research on the ethics of AI *for* cyber security.

Future research could design guidelines for implementing AI-powered tooling based on the identified factors, or a subset of those factors (e.g., explainability and by extent also interpretability). What additionally should be investigated is how opinions will shift in the next two to three years, as during the six months of the Delphi part of this study, opinions have already started to shift.

To summarize, the key findings are that the factors that are associated with acceptance of autonomous interventions in cyber security defense are 1) the potential impact on business operations and 2) humans being able to understand these autonomous interventions. Although this was researched specifically for the Dutch financial sector, these findings might be relevant to other sectors as well. Most organizations want to prevent operational impact by autonomous AI and understand its decisions. However, organizations in other sectors might have a different risk-appetite, meaning the level of operational impact they tolerate differs from the financial sector and therefore accept autonomous AI more quickly than in the Dutch financial sector.

5.3 Recommendations

Reflecting on the above, you could say the CISOs can't have their cake and eat it too. Autonomous AI is accepted if the decisions are explainable, unbiased and do not have any operational impact. Autonomous AI is not fully trusted yet, because the reliability and accuracy are not perfect. Conversely, the expectation is that autonomous AI will be needed in cyber security defense. Therefore, financial organizations' capability and trust need to grow: Start by implementing autonomous AI in low-risk and low-impact scenarios (Nyre-Yu et al., 2019). As P3 stated, have a "validation and testing process". Consider staying away from GenAI based systems, because explaining their output is not yet possible. Build up the required skill and expertise and involve users early on (Cabitza et al., 2021; Hauptman et al., 2023; Solaimani & Swaak, 2023). As P5 suggested, require personnel to get certified. Then, the most difficult part, every organization will need to start allowing autonomous AI in higher-impact scenarios, driven by a risk-based approach. Because after all, the threat actors will keep on coming and they will eventually figure out how to use the full capabilities of AI to attack any of us. This advice is based on an interpretation of the findings of this study and the current cyber security and AI landscape in practice, and therefore further research is required to make any scientific claims.

6 Conclusion

6.1 Conclusion

This study aimed to find what factors are associated with acceptance of AI-driven autonomous interventions in cyber security defense, specifically by organizations within the Dutch financial sector. Potential factors were extracted from the literature via SQ1, SQ2 and SQ3. These 36 factors were verified by interviewing a CISO of a Dutch financial institution. Via a mixed-method three round Delphi study the factors that are associated with acceptance of autonomous AI tooling were determined. The first round identified an additional factor. The second and third round covered 23 out of 37 factors, driven by the prioritized list resulting from the interview. Eight (chief) information security officers from Dutch financial institutions took part in the Delphi study. The goal of a Delphi study is to reach consensus on statements, usually statements about the future.

This study confirms that autonomous AI-driven cyber security tooling is expected to be needed in the next three years to be able to withstand the growing scale of cyber-attacks. Human *on* the loop is accepted in some but not all scenarios. When human on the loop is not accepted, human *in* the loop supporting cyber security personnel is accepted. The key factors that are associated with acceptance of autonomous interventions are 1) the potential impact on business operations and 2) humans being able to understand these autonomous interventions. Explainability and transparency are requirements for human understanding, as is humans and AI having a shared situational awareness.

The other factors that are associated with acceptance of autonomous AI tooling in the financial sector are auditing of services, efficiency during investigations and a quicker response, perceived usefulness, performance/effort expectancy, safety/security expectancy, technical accuracy and consistency of analyses.

This study confirmed that the Mission Critical Technology Acceptance Model (MCTAM) by Weger et al. (2023) fits the cyber security context. This study found that *reliability* influences *trust*. The trust in AI tooling is hampered, because reliability, especially the accuracy, of autonomous AI tooling is being questioned. Based on the literature the MCTAM was modified: *risk* and *context* influence *trust*; however, the results of this study do not confirm this modification of the MCTAM. The resulting MMCTAM is shown in Figure 16.

This study finds that black-box autonomous AI systems are not accepted in cyber security: The decisions of AI tooling need to be transparent and explainable to humans. If GenAI systems remain black boxes, we should not try to explain them but create and use models that are interpretable instead (Rudin, 2019). For cyber security, rule based AI might be a safer approach for impactful decisions (Sarker et al., 2024). On the detection side GenAI is still useful though: intrusion detection, vulnerability detection, phishing detection, threat intelligence and malware classification (Ferrag et al., 2025).

We conclude that, yes, autonomous interventions in cyber security defense can be accepted by organizations within the financial sector. Autonomous AI *is* accepted if the decisions are explainable, unbiased and do not have any operational impact. Autonomous AI is not fully trusted yet, because the reliability and accuracy are not perfect. Conversely, the expectation is that autonomous AI will be needed. Therefore, financial organizations' capability and trust need to grow: Start by implementing autonomous AI in low-risk and low-impact scenarios (Nyre-Yu et al., 2019). Build up the required skill and expertise and involve users early on (Cabitza et al., 2021; Hauptman et al., 2023; Solaimani & Swaak, 2023). Then, the most difficult part, every organization will need to start allowing autonomous AI in higher-impact scenarios, driven by a risk-based approach. Because after all, the threat actors keep on coming and they will eventually figure out how to use the full capabilities of AI to attack any of us. This advice is based on an interpretation of the findings of this study, and the current cyber security and AI landscape in practice; therefore, further research is required.

Looking into the future, I predict that by the end of 2026 we will know if the integration of GenAI in cyber security products (and beyond) was hyped, or that explainability and transparency, and reliability have been achieved to an acceptable level.

6.2 Limitations

This study provides new insight into the factors related to organizations accepting AI tooling to take autonomous actions in a cyber security context. The participants of the Delphi study, chief information security officers, have a strategic perspective; they will not interact with autonomous AI in cyber security defense directly. This study therefore has a strategic perspective. Another topic for research could be the acceptance of (specific implementations of) AI enabled autonomous response by operational teams, focusing on ease of use instead of usefulness.

This exploratory study however had a small sample size: it was restricted to a small, homogenous group of experts from the financial sector in the Netherlands. The views of a small group may not be representative, which impairs generalization (Skulmoski et al., 2007). A larger follow-up study within the context of the Dutch financial sector is needed to validate the conclusions of this study, preferably with a heterogenous group of experts. A larger sample reduces the group error (Skulmoski et al., 2007). Different sectors will (have to) start using autonomous AI as well. Future research in different sectors or countries is needed for further generalization (Skulmoski et al., 2007).

6.3 Reflection

Reflecting on the journey of this thesis, it would have been nice to have stumbled on good examples of structured literature reviews at the beginning of this journey (Page et al., 2021; Xiao & Watson, 2019). My focus and energy have been on applying the Delphi method, see the section below. Even though the literature reviews performed for this thesis were exploratory, a next learning objective for me would be researching and writing a full-on structured review paper following the extant guidelines. Reflecting on the courses preparing for the thesis, more guidance could have been given on the topic of doing and documenting structured literature reviews.

6.3.1 Delphi method

Completing this study presented several small challenges in applying the Delphi method, not in the least because preparing for something you did not do (to this extent) before will have gaps.

I am very grateful for the time and effort the participants put in. In hindsight though, the number of participants should have been significantly higher at least at the start. The pool of potential participants was defined too narrow, with at most about 20 participants. Starting by inviting 20 participants, then requesting participation in three rounds, leads to difficulty reaching sufficient sample size at the final round. Participants have limited time available or might lose interest. Keeping the initial participants engaged took effort and time, because there was no room for attrition. Chasing the participants took almost a month for the second and more than a month for the third round. For the second round, this risked other participants dropping out before the third round started.

Perhaps an online Delphi would have been better? There were two months between the rounds on average. Five months passed between the interview up to the final response to the third Delphi round. During this time, from March to August 2025, the perception of safe usage of GenAI changed, see Chapter 5. However, it would have been difficult to get CISOs of financial institutions to commit to participate at the same time in a real-time Delphi study. Buy-in upfront was not guaranteed, therefore a digital, but not real-time, Delphi study was the safer choice.

Providing a *No comment* option in the quantitative rounds prevented calculating the stability between Round 2 and 3 for one statement. If a neutral mid-point would have been available, the participants might have chosen that instead of *No comment*. However, as shown in 3.5.7 there are issues with the neutral midpoint. I stand by this choice, supported by literature, as it might lead to a more realistic result. By realistic I mean that if a participant does not know, they can use the *No comment* escape instead of being neutral.

A Delphi study might be difficult to manage for a thesis, especially if it is part-time and includes participants from different companies. A Delphi study might in some ways be more difficult than having interviews or performing a questionnaire. Is the additional effort, including learning how to perform Delphi, of value? This journey did teach me a lot, so I tend to agree.

On the other hand, because this Delphi study took almost half a year, it was possible to capture the changing opinions with respect to (Gen)AI. A single interview per participant, or a single survey can never capture this.

Next, I learned a few lessons:

The Delphi first round used scenarios. The research proposal for this study did not include scenario generation but only mentioned that open questions would be used. Scenario generation is an integral part of a Delphi study, which should have been clear in the proposal phase.

In hindsight, the term ‘lower response speed’ is confusing and even confused me at times; it means that it takes longer to respond, as in: It lowers the response speed. This would be a disadvantage.

Finally, based on my experience with a small-scale Delphi study I have taken the liberty to suggest some considerations in Table 6 for others attempting a Delphi study as part of their thesis.

Table 6

Considerations for Delphi studies

Qualitative rounds
Why limit scenarios to two or three open questions? It takes mental effort to switch scenarios. However, a new scenario can break thought patterns, for example by introducing a new perspective.
Quantitative rounds
Start with at least 12 participants (and 24 in case of a heterogenous group of experts). The absolute minimum for a stability test between rounds is the same six participants in both rounds. Chasing participants to limit attrition will add delays, and delays can cause participants to drop off in the next round. Prefer a real-time online Delphi if scheduling allows. Opinions of participants can change between rounds if there are weeks or months in between them, making reaching stability more difficult. A delay of one to two months between subsequent rounds might capture changing opinions. To conclude anything with statistical significance about this change in opinions is not possible with a Delphi study with two quantitative rounds. However, two separate surveys might be a better instrument when opinions between rounds/surveys are expected to diverge a lot. It is unclear how Delphi studies in (peer reviewed) papers calculated the IQR. Always specify the algorithm you use. Do add a *No comment* choice. This gives participants the option to disqualify themselves for a statement if it is not their area of expertise. If the *No comment* choice is left out, it forces participants to choose an option on the Likert scale in every round. However, for each participant that chooses *No comment*, the stability between rounds cannot be calculated for that participant.

7 References

Abd Majid, M., & Zainol Ariffin, K. A. (2019, October 1). *Success Factors for Cyber Security Operation Center (SOC) Establishment*. Proceedings of the 1st International Conference on Informatics, Engineering, Science and Technology, INCITEST 2019, 18 July 2019, Bandung, Indonesia. https://eudl.eu/doi/10.4108/eai.18-7-2019.2287841

Abd Majid, M., & Zainol Ariffin, K. A. (2021). Model for successful development and implementation of Cyber Security Operations Centre (SOC). *PloS One*, *16*(11), e0260157. https://doi.org/10.1371/journal.pone.0260157

Abnormal Security. (2025, January 23). *How GhostGPT Empowers Cybercriminals with Uncensored AI*. Abnormal. https://abnormalsecurity.com/blog/ghostgpt-uncensored-ai-chatbot

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. (2025). *Circuit Tracing: Revealing Computational Graphs in Language Models*. Transformer Circuits. https://transformer-circuits.pub/2025/attribution-graphs/methods.html

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Andersen, P. D. (2023). Constructing Delphi statements for technology foresight. *FUTURES & FORESIGHT SCIENCE*, *5*(2), e144. https://doi.org/10.1002/ffo2.144

Apostoaie, C.-M., & Bilan, I. (2020). Macro determinants of shadow banking in Central and Eastern European countries. *Economic Research-Ekonomska Istraživanja*, *33*(1), 1146–1171. https://doi.org/10.1080/1331677X.2019.1633943

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Aubley, C., Bowen, E., Frank, W., Golden, D., Morris, M., & Norton, K. (2021, December 7). *Cyber AI: Real defense*. Deloitte Insights. https://www2.deloitte.com/us/en/insights/focus/tech-trends/2022/future-of-cybersecurity-and-ai.html

Baltariu, I. A., & Puscasu, E. (2024, January 18). *Stream-Jacking 2.0: Deep fakes power account takeovers on YouTube to maximize crypto-doubling scams*. Bitdefender Labs. https://www.bitdefender.com/blog/labs/stream-jacking-2-0-deep-fakes-power-account-takeovers-on-youtube-to-maximize-crypto-doubling-scams/

Baron, H., Buker, J., Gifford, R., Heide, S., Kaluza, A., & Yeoh, J. (2024). *State of AI and Security Survey Report*. https://cloudsecurityalliance.org/artifacts/the-state-of-ai-and-security-survey-report

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

BBC. (2024, January 30). *UPS to cut 12,000 jobs after "disappointing" year*. https://www.bbc.com/news/business-68144738

Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L., & Schweitzer, V. M. (2021). Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. *MethodsX*, *8*, 101401. https://doi.org/10.1016/j.mex.2021.101401

Belanger, A. (2025, August 21). *Bank forced to rehire workers after lying about chatbot productivity, union says*. Ars Technica. https://arstechnica.com/tech-policy/2025/08/bank-forced-to-rehire-workers-after-lying-about-chatbot-productivity-union-says/

Bell, E., Bryman, A., & Harley, B. (2022). *Business research methods* (Sixth edition). Oxford University Press.

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting and Social Change*, *147*, 72–82. https://doi.org/10.1016/j.techfore.2019.07.002

Benedikt, L., Joshi, C., Nolan, L., Henstra-Hill, R., Shaw, L., & Hook, S. (2020). Human-in-the-loop AI in government: A case study. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 488–497. https://doi.org/10.1145/3377325.3377489

Berman, D., Buczak, A., Chavis, J., & Corbett, C. (2019). A Survey of Deep Learning Methods for Cyber Security. *Information*, *10*(4), 122. https://doi.org/10.3390/info10040122

Bernsmed, K., Meland, P. H., & Jaatun, M. G. (2015). *Cloud Security Requirements—A checklist with security and privacy requirements for public cloud services*. SINTEF Digital / Software Engineering, Safety and Security. https://www.sintef.no/en/publications/publication/1334380/

Beurer-Kellner, L., Creţu, B. B. A.-M., Debenedetti, E., Dobos, D., Fabian, D., Fischer, M., Froelicher, D., Grosse, K., Naeff, D., Ozoani, E., Paverd, A., Tramèr, F., & Volhejn, V. (2025). *Design Patterns for Securing LLM Agents against Prompt Injections* (No. arXiv:2506.08837). arXiv. https://doi.org/10.48550/arXiv.2506.08837

Bisen, A., McConlogue, M., & Perumal, Z. (2023, December 1). *Real-time Insights from the Open, Social and Deep Web with AI & LLMs*. https://www.overwatchdata.ai/blog/real-time-insights-from-the-open-social-and-deep-web-with-ai-and-llms

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (No. arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bosch, K., Fransen, F., de Graaf, P., Hehanussa, D., van der Kleij, R., te Paske, B. J., Vetjens, B., & Wolthuis, R. (2023). *Checkmate, cyber security?* TNO. https://www.tno.nl/nl/newsroom/insights/2023/10/toekomst-cybersecurity-autonoom-systeem/

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

Brauner, P., Hick, A., Philipsen, R., & Ziefle, M. (2023). What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1113903

Brewer, R. (2019). Could SOAR save skills-short SOCs? *Computer Fraud & Security*, *2019*(10), 8–11. https://doi.org/10.1016/S1361-3723(19)30106-X

Bridges, R. A., Rice, A. E., Oesch, S., Nichols, J. A., Watson, C., Spakes, K., Norem, S., Huettel, M., Jewell, B., Weber, B., Gannon, C., Bizovi, O., Hollifield, S. C., & Erwin, S. (2023). Testing SOAR tools in use. *Computers & Security*, *129*. https://doi.org/10.1016/j.cose.2023.103201

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G., Steinhardt, J., Flynn, C., hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., & Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*.

Busch, D., & van Rijn, M. B. J. (2018). Towards Single Supervision and Resolution of Systemically Important Non-Bank Financial Institutions in the European Union. *European Business Organization Law Review*, *19*(2), 301–363. https://doi.org/10.1007/s40804-018-0107-5

Cabitza, F., Campagner, A., & Simone, C. (2021). The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. *International Journal of Human - Computer Studies*, *155*. https://doi.org/10.1016/j.ijhcs.2021.102696

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 104:1-104:24. https://doi.org/10.1145/3359206

Cassetto, O. (2024, June 10). *AI SOC: The Definition and Components of AI-Driven SOC*. Radiant Security. https://radiantsecurity.ai/learn/ai-driven-soc/

Cato Networks. (2025, March 18). The Rise of the Zero-Knowledge Threat Actor: New LLM Jailbreak Technique Discovered by Cato Networks Enables Easy Creation of Password-Stealing Malware. *Cato Networks*. https://www.catonetworks.com/news/the-rise-of-the-zero-knowledge-threat-actor/

Cezar, A., Cavusoglu, H., & Raghunathan, S. (2017). Sourcing Information Security Operations: The Role of Risk Interdependency and Competitive Externality in Outsourcing Decisions. *Production & Operations Management*, *26*(5).

Chang, O., Liu, D., & Metzman, J. (2024, November 20). Leveling Up Fuzzing: Finding more vulnerabilities with AI. *Google Online Security Blog*. https://security.googleblog.com/2024/11/leveling-up-fuzzing-finding-more.html

Check Point. (2023, April 27). *Global Cyberattacks Continue to Rise with Africa and APAC Suffering Most*. https://blog.checkpoint.com/research/global-cyberattacks-continue-to-rise/

Check Point Research. (2025, January 10). *FunkSec – Alleged Top Ransomware Group Powered by AI*. Check Point Research. https://research.checkpoint.com/2025/funksec-alleged-top-ransomware-group-powered-by-ai/

Chen, Y., Xie, H., Ma, M., Kang, Y., Gao, X., Shi, L., Cao, Y., Gao, X., Fan, H., Wen, M., Zeng, J., Ghosh, S., Zhang, X., Zhang, C., Lin, Q., Rajmohan, S., Zhang, D., & Xu, T. (2024). Automatic Root Cause Analysis via Large Language Models for Cloud Incidents. *Proceedings of the Nineteenth European Conference on Computer Systems*, 674–688. https://doi.org/10.1145/3627703.3629553

Chu, M., Zong, K., Shu, X., Gong, J., Lu, Z., Guo, K., Dai, X., & Zhou, G. (2023). Work with AI and Work for AI: Autonomous Vehicle Safety Drivers' Lived Experiences. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3544548.3581564

Ciancaglini, V., & Sancho, D. (2024, May 8). *Back to the Hype: An Update on How Cybercriminals Are Using GenAI*. https://www.trendmicro.com/vinfo/gb/security/news/cybercrime-and-digital-threats/back-to-the-hype-an-update-on-how-cybercriminals-are-using-genai

Cleland-Huang, J., Chambers, T., Zudaire, S., Chowdhury, M. T., Agrawal, A., & Vierhauser, M. (2024). Human–machine Teaming with Small Unmanned Aerial Systems in a MAPE-K Environment. *ACM Transactions on Autonomous and Adaptive Systems*, *19*(1), 3:1-3:35. https://doi.org/10.1145/3618001

Coker, H., Jr. (2024, September 4). *Service for America: Cyber Is Serving Your Country | ONCD*. The White House. https://www.whitehouse.gov/oncd/briefing-room/2024/09/04/service-for-america-cyber-is-serving-your-country/

Columbus, L. (2023, January 3). Defensive vs. offensive AI: Why security teams are losing the AI war. *VentureBeat*. https://venturebeat.com/security/defensive-vs-offensive-ai-why-security-teams-are-losing-the-ai-war/

Competition and Markets Authority (UK). (2024). *AI Foundation Models Update paper*. https://www.gov.uk/government/news/cma-outlines-growing-concerns-in-markets-for-ai-foundation-models

Conover, W. J. (1973). On Methods of Handling Ties in the Wilcoxon Signed-Rank Test. *Journal of the American Statistical Association*, *68*(344), 985–988. https://doi.org/10.2307/2284536

Criddle, C., & Quinio, A. (2024, June 30). *Financial services shun AI over job and regulatory fears*. https://www.ft.com/content/0675e4d9-62a1-4d6c-9098-a8cb0d1e32ed

Cuprik, R. (2025, August 27). *ESET researcher discovers the first known AI-written ransomware: I feel thrilled but cautious*. https://www.eset.com/blog/en/business-topics/threat-landscape/the-first-known-ai-written-ransomware/

Cyber Security Raad. (2023, December 22). *Informerende brief van de Cyber Security Raad over onderwijsversterking en kennisontwikkeling* [Brief]. Ministerie van Justitie en Veiligheid. https://www.cybersecurityraad.nl/documenten/brieven/2023/12/22/informerende-brief-van-de-cyber-security-raad-over-onderwijsversterking-en-kennisontwikkeling

Cyware. (2023, December 1). *SOAR and AI in Cybersecurity: Reshaping Security Operations*. https://cyware.com/security-guides/security-orchestration-automation-and-response/from-insight-to-action-how-ai-and-soar-are-reshaping-security-operations-13d9/

Da Silva, J., & Jensen, R. B. (2022). “Cyber security is a dark art”: The CISO as Soothsayer. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 365:1-365:31. https://doi.org/10.1145/3555090

Danquah, P. (2020). Security Operations Center: A Framework for Automated Triage, Containment and Escalation. *Journal of Information Security*, *11*(4), 225–240. https://doi.org/10.4236/jis.2020.114015

Darktrace. (2024, April 9). *Darktrace Transforms Security Operations and Improves Cyber Resilience with Launch of Darktrace ActiveAI Security Platform™*. https://ir.darktrace.com/press-releases/2024/4/9/ad92f587789affc79165e131f0e4d8752139a9b7d960c0c148a888da891b071d

Davies, R. B., & Killeen, N. (2018). Location decisions of non-bank financial foreign direct investment: Firm-level evidence from Europe. *Review of International Economics*, *26*(2), 378–403. https://doi.org/10.1111/roie.12336

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

De Lucia, M., Newcomb, A., & Kott, A. (2019). *Features and Operation of an Autonomous Agent for Cyber Defense*.

De Nederlandsche Bank. (2019). *General principles for the use of Artificial Intelligence in the financial sector*.

De Nederlandsche Bank. (2024). *Visie op Toezicht 2025-2028*. https://www.dnb.nl/nieuws-voor-de-sector/toezicht-2024/dnb-publiceert-visie-op-toezicht-2025-2028/

De Nederlandsche Bank & Dutch Authority for the Financial Markets. (2024). *De impact van AI op de financiële sector en het toezicht*. https://www.dnb.nl/algemeen-nieuws/nieuwsberichten-2024/financieel-toezichthouders-bereiden-zich-voor-op-ai-tijdperk/

De Vet, E., Brug, J., Nooijer, J., Dijkstra, A., & Vries, N. (2005). Determinants of forward stage transitions: A Delphi study. *Health Education Research*, *20*, 195–205. https://doi.org/10.1093/her/cyg111

de Visser, E. J., Phillips, E., Tenhundfeld, N., Donadio, B., Barentine, C., Kim, B., Madison, A., Ries, A., & Tossell, C. C. (2023). Trust in automated parking systems: A mixed methods evaluation: Transportation Research: Part F. *Transportation Research: Part F*, *96*, 185–199. https://doi.org/10.1016/j.trf.2023.05.018

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1976). Group techniques for program planning. *Group & Organization Studies*, *1*(2), 256–256. https://doi.org/10.1177/105960117600100220

Derrick, B., & White, P. (2017). Comparing Two Samples from an Individual Likert Question. *International Journal of Mathematics and Statistics*. https://www.semanticscholar.org/paper/Comparing-Two-Samples-from-an-Individual-Likert-Derrick-White/cc9f628b3c4282881ce211ac998ccd4910a9c25e

Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on Measures for a High Common Level of Cybersecurity across the Union, Amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and Repealing Directive (EU) 2016/1148 (NIS 2 Directive) (2022). https://eur-lex.europa.eu/eli/dir/2022/2555/oj

Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the Resilience of Critical Entities and Repealing Council Directive 2008/114/EC (2022). https://eur-lex.europa.eu/eli/dir/2022/2557/oj

Douer, N., & Meyer, J. (2020). The Responsibility Quantification Model of Human Interaction With Automation: IEEE Transactions on Automation Science & Engineering. *IEEE Transactions on*

Automation Science & Engineering, *17*(2), 1044–1060. https://doi.org/10.1109/TASE.2020.2965466

Dutch Authority for the Financial Markets. (2024). *Trend Monitor 2025*. https://www.afm.nl/en/over-de-afm/verslaglegging/trendzicht

Dutch Banking Association. (n.d.). *FI-ISAC*. Retrieved April 2, 2024, from https://www.nvb.nl/themas/veilig-bankieren/fi-isac/

ENISA. (2023). *ENISA Threat Landscape 2023* [Report/Study]. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

ENISA. (2024). *ENISA Threat Landscape 2024* [Report/Study]. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024

Fang, R., Bowman, D., & Kang, D. (2024). *Voice-Enabled AI Agents can Perform Common Scams* (No. arXiv:2410.15650). arXiv. https://doi.org/10.48550/arXiv.2410.15650

Feng, N., Zhang, S., Li, M., & Li, D. (2021). Contracting managed security service: Double moral hazard and risk interdependency. *Electronic Commerce Research and Applications*, *50*. https://doi.org/10.1016/j.elerap.2021.101097

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., & Debbah, M. (2025). *Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities* (No. arXiv:2405.12750). arXiv. https://doi.org/10.48550/arXiv.2405.12750

Fire, M., Elbazis, Y., Wasenstein, A., & Rokach, L. (2025). *Dark LLMs: The Growing Threat of Unaligned AI Models* (No. arXiv:2505.10066). arXiv. https://doi.org/10.48550/arXiv.2505.10066

Fischer, J. E., Greenhalgh, C., Jiang, W., Ramchurn, S. D., Wu, F., & Rodden, T. (2021). In-the-loop or on-the-loop? Interactional arrangements to support team coordination with a planning agent. *Concurrency and Computation: Practice and Experience*, *33*(8). https://doi.org/10.1002/cpe.4082

Flechais, I., & Chalhoub, G. (2023). Practical Cybersecurity Ethics: Mapping CyBOK to Ethical Concerns. *Proceedings of the 2023 New Security Paradigms Workshop*, 62–75. https://doi.org/10.1145/3633500.3633505

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical

Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Formosa, P., Wilson, M., & Richards, D. (2021). A principlist framework for cybersecurity ethics. *Computers & Security*, *109*, 102382. https://doi.org/10.1016/j.cose.2021.102382

Forsberg, J., & Frantti, T. (2023). Technical performance metrics of a security operations center. *Computers & Security*, *135*. https://doi.org/10.1016/j.cose.2023.103529

Frank, M., Brennecke, M., Hölzmer, P., Pocher, N., & Fridgen, G. (2025). *Potential of AI for User-Centric Cybersecurity in the Financial Sector*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2025.543

Franklin, K. K., & Hart, J. K. (2007). Idea Generation and Exploration: Benefits and Limitations of the Policy Delphi Research Method. *Innovative Higher Education*, *31*(4), 237–246. https://doi.org/10.1007/s10755-006-9022-8

Geer, D., Jardine, E., & Leverett, E. (2020). On market concentration and cybersecurity risk. *Journal of Cyber Policy*, *5*(1), 9–29. https://doi.org/10.1080/23738871.2020.1728355

Gibney, E. (2024). Not all ‘open source’ AI models are actually open: Here’s a ranking. *Nature*. https://doi.org/10.1038/d41586-024-02012-5

Gooding, S., Brown, O., & Zee, P. van der. (2025, September 27). *Nx npm Packages Compromised in Supply Chain Attack Weaponizi...* Socket. https://socket.dev/blog/nx-packages-compromised

Google Project Zero. (2024, November 1). Project Zero: From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code. *Project Zero*. https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html

Google Threat Intelligence Group. (2025, January 29). *Adversarial Misuse of Generative AI*. Google Cloud Blog. https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence*, *36*(1), 2037254. https://doi.org/10.1080/08839514.2022.2037254

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, *49*, 157–169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008

Hamon, R., Junklewitz, H., Malgieri, G., Hert, P. D., Beslay, L., & Sanchez, I. (2021). Impossible Explanations?: Beyond explainable AI in the GDPR from a COVID-19 use case scenario. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 549–559. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445917

Hauptman, A. I., Schelble, B. G., McNeese, N. J., & Madathil, K. C. (2023). Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior*, *138*. https://doi.org/10.1016/j.chb.2022.107451

Haykin, S. (2019). Artificial Intelligence Communicates With Cognitive Dynamic System for Cybersecurity. *IEEE Transactions on Cognitive Communications and Networking*, *PP*, 1–1. https://doi.org/10.1109/TCCN.2019.2930253

Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024). *Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects* (No. arXiv:2412.00586). arXiv. https://doi.org/10.48550/arXiv.2412.00586

Hilal, W., Gadsden, S. A., & Yawney, J. (2023). Cognitive Dynamic Systems: A Review of Theory, Applications, and Recent Advances. *Proceedings of the IEEE*, *111*(6), 575–622. https://doi.org/10.1109/JPROC.2023.3272577

Hirschhorn, F. (2019). Reflections on the application of the Delphi method: Lessons from a case in public transport research. *International Journal of Social Research Methodology*, *22*(3), 309–322. https://doi.org/10.1080/13645579.2018.1543841

Hofmans, T. (2024, August 11). *Voorlopig is AI voor cybercriminelen vooral een hulpmiddel*. Tweakers. https://tweakers.net/reviews/12368/voorlopig-is-ai-voor-cybercriminelen-vooral-een-hulpmiddel.html

Holbrook, C., Holman, D., Clingo, J., & Wagner, A. R. (2024). Overtrust in AI Recommendations About Whether or Not to Kill: Evidence from Two Human-Robot Interaction Studies. *Scientific Reports*, *14*(1), 19751. https://doi.org/10.1038/s41598-024-69771-z

Hsieh, P.-J. (2023). Determinants of physicians' intention to use AI-assisted diagnosis: An integrated readiness perspective. *Computers in Human Behavior*, *147*, 107868. https://doi.org/10.1016/j.chb.2023.107868

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi Technique: Making Sense of Consensus. *Practical Assessment, Research, and Evaluation*, *12*(1), Article 1. https://doi.org/10.7275/pdz9-th90

Hsu, D., Neu, M., Farrag, M., & Kindi, R. (2024, June 24). Leveraging AI for efficient incident response. *Engineering at Meta*. https://engineering.fb.com/2024/06/24/data-infrastructure/leveraging-ai-for-efficient-incident-response/

Hyndman, R., & Fan, Y. (1996). Sample Quantiles in Statistical Packages. *The American Statistician*, *50*, 361–365. https://doi.org/10.1080/00031305.1996.10473566

Insikt Group. (2024). *Adversarial Intelligence: Red Teaming Malicious Use Cases for AI*. https://www.recordedfuture.com/adversarial-intelligence-red-teaming-malicious-use-cases-ai

Ivanova, I. (2025, May 9). *As Klarna flips from AI-first to hiring people again, a new landmark survey reveals most AI projects fail to deliver*. Fortune. https://fortune.com/2025/05/09/klarna-ai-humans-return-on-investment/

Jacob, S., & Furgerson, S. (2015). Writing Interview Protocols and Conducting Interviews: Tips for Students New to the Field of Qualitative Research. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2012.1718

Joia, L. A., & Cordeiro, J. P. V. (2021). Unlocking the Potential of Fintechs for Financial Inclusion: A Delphi-Based Approach. *Sustainability*, *13*(21), 11675. https://doi.org/10.3390/su132111675

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ (Clinical Research Ed.)*, *311*(7001), 376–380. https://doi.org/10.1136/bmj.311.7001.376

Joshi, A., & Khalfi-Lanoux, N. (2024, July). *Top 10 operational impacts of the EU AI Act – Subject matter, definitions, key actors and scope*. https://iapp.org/resources/article/top-impacts-eu-ai-act-scope-subject-definitions-actors/

Kayworth T. & Whitten D. (2010). Effective information security requires a balance of social and technology factors. *MIS Quarterly Executive*, *9*(3), 163–175.

Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*. https://doi.org/10.1016/j.tele.2022.101925

Klarna. (2024, May 14). *90% of Klarna staff are using AI daily—Game changer for productivity*. https://www.klarna.com/international/press/90-of-klarna-staff-are-using-ai-daily-game-changer-for-productivity/

Kluge, U., Ringbeck, J., & Spinler, S. (2020). Door-to-door travel in 2035 – A Delphi study. *Technological Forecasting and Social Change*, *157*, 120096. https://doi.org/10.1016/j.techfore.2020.120096

Kokulu, F. B., Soneji, A., Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., & Ahn, G.-J. (2019). Matched and Mismatched SOCs: A Qualitative Study on Security Operations Center Issues. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 1955–1970. https://doi.org/10.1145/3319535.3354239

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* (No. arXiv:2506.08872). arXiv. https://doi.org/10.48550/arXiv.2506.08872

Kott, A. (2023). *Autonomous Intelligent Cyber-defense Agent: Introduction and Overview* (No. arXiv:2304.12408). arXiv. https://doi.org/10.48550/arXiv.2304.12408

Kruse, J., Connor, A. M., & Marks, S. (2022). Evaluation of a Multi-agent “Human-in-the-loop” Game Design System. *ACM Trans. Interact. Intell. Syst.*, *12*(3), 19:1-19:26. https://doi.org/10.1145/3531009

Kuiper, O., van den Berg, M., van der Burgt, J., & Leijnen, S. (2022). Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities. In L. A. Leiva, C. Pruski, R. Markovich, A. Najjar, & C. Schommer (Eds.), *Artificial Intelligence and Machine Learning* (pp. 105–119). Springer International Publishing. https://doi.org/10.1007/978-3-030-93842-0_6

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., … Chan, L. (2025). *Measuring AI Ability to Complete Long Tasks* (No. arXiv:2503.14499). arXiv. https://doi.org/10.48550/arXiv.2503.14499

Kyriakou, K., & Otterbacher, J. (2023). In humans, we trust. *Discover Artificial Intelligence*, *3*(1), 44. https://doi.org/10.1007/s44163-023-00092-2

Lee, H.-P. (Hank), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3706598.3713778

Liehner, G. L., Brauner, P., Schaar, A. K., & Ziefle, M. (2022). Delegation of Moral Tasks to Automated Agents—The Impact of Risk and Context on Trusting a Machine to Perform a Task. *IEEE Transactions on Technology and Society*, *3*(1). https://doi.org/10.1109/TTS.2021.3118355

Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1774–1787. https://doi.org/10.1145/3630106.3659005

Lin, Z., Cui, J., Liao, X., & Wang, X. (2024). Malla: Demystifying Real-world Large Language Model Integrated Malicious Services. In D. Balzarotti & W. Xu (Eds.), *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*. USENIX Association. https://www.usenix.org/conference/usenixsecurity24/presentation/lin-zilong

Lorenz, P., Perset, K., & Berryhill, J. (2023, September 18). *Initial policy considerations for generative artificial intelligence*. OECD. https://www.oecd.org/governance/initial-policy-considerations-for-generative-artificial-intelligence-fae2d1e6-en.htm

Madjar, T., Larson, S., & Proofpoint Threat Research Team. (2024, April 10). *TA547 Targets German Organizations with Rhadamanthys Stealer*. https://www.darkreading.com/threat-intelligence/ta547-uses-llm-generated-dropper-infect-german-orgs

Magramo, K. (2024, May 17). *British engineering giant Arup revealed as $25 million deepfake scam victim*. CNN. https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/index.html

Markmann, C., Spickermann, A., von der Gracht, H. A., & Brem, A. (2021). Improving the question formulation in Delphi-like surveys: Analysis of the effects of abstract language and amount of information on response behavior. *FUTURES & FORESIGHT SCIENCE*, *3*(1), e56. https://doi.org/10.1002/ffo2.56

Mayer, M. (2023). Trusting machine intelligence: Artificial intelligence and human-autonomy teaming in military operations. *Defense & Security Analysis*, *39*(4), 521–538. https://doi.org/10.1080/14751798.2023.2264070

Maynard, S., Onibere, M., & Ahmad, A. (2018). Defining the Strategic Role of the Chief Information Security Officer. *Pacific Asia Journal of the Association for Information Systems*, *10*(3). https://doi.org/10.17705/1pais.10303

Mayoral-Vilches, V., & Rynning, P. M. (2025). *Cybersecurity AI: Hacking the AI Hackers via Prompt Injection* (No. arXiv:2508.21669). arXiv. https://doi.org/10.48550/arXiv.2508.21669

Melecky, M., & Podpiera, A. M. (2020). Financial sector strategies and financial sector outcomes: Do the strategies perform? *Economic Systems*, *44*(2). https://doi.org/10.1016/j.ecosys.2020.100757

Meyer-Waarden, L., & Cloarec, J. (2022). "Baby, you can drive my car": Psychological antecedents that drive consumers' adoption of AI-powered autonomous vehicles. *Technovation*, *109*. https://doi.org/10.1016/j.technovation.2021.102348

Microsoft. (2023, October 11). *Microsoft Defender for Endpoint now stops human-operated attacks on its own*. https://www.microsoft.com/en-us/security/blog/2023/10/11/microsoft-defender-for-endpoint-now-stops-human-operated-attacks-on-its-own/

Microsoft Threat Intelligence. (2024, February 14). *Staying ahead of threat actors in the age of AI*. Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

Ministerie van Economische Zaken en Klimaat. (2024, May 15). *Onderzoeksrapportage Onderwijs en Arbeidsmarkt Cybersecurity* [Rapport]. Rijksoverheid; Ministerie van Algemene Zaken. https://www.rijksoverheid.nl/documenten/rapporten/2024/05/15/bijlage-onderzoeksrapport-adviesrapport-onderwijs-en-arbeidsmarkt-cybersecurity-ptvt

MIT Technology Review. (2021). *Preparing for AI-enabled cyberattacks*. MIT. https://www.technologyreview.com/2021/04/08/1021696/preparing-for-ai-enabled-cyberattacks/

Moeuf, A., Lamouri, S., Pellerin, R., Tamayo-Giraldo, S., Tobon-Valencia, E., & Eburdy, R. (2020). Identification of critical success factors, risks and opportunities of Industry 4.0 in SMEs.

International Journal of Production Research, *58*(5), 1384–1400. https://doi.org/10.1080/00207543.2019.1636323

Mohr, S., & Kühl, R. (2021). Acceptance of artificial intelligence in German agriculture: An application of the technology acceptance model and the theory of planned behavior. *Precision Agriculture : An International Journal on Advances in Precision Agriculture*, *22*(6), 1816–1844. https://doi.org/10.1007/s11119-021-09814-x

Moix, A., Lebedev, K., & Klein, J. (2025). *Threat Intelligence Report: August 2025*. Anthropic. https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf

Naghshvarianjahromi, M., Kumar, S., & Deen, M. J. (2023). Natural Intelligence as the Brain of Intelligent Systems: Sensors (14248220). *Sensors (14248220)*, *23*(5), 2859. https://doi.org/10.3390/s23052859

Nationaal Coördinator Terrorismebestrijding en Veiligheid. (2023, November 6). *Overzicht vitale processen*. Ministerie van Justitie en Veiligheid; Ministerie van Justitie en Veiligheid. https://www.nctv.nl/onderwerpen/vitale-infrastructuur/overzicht-vitale-processen

National Cyber Security Centre. (2024, January 24). *The near-term impact of AI on the cyber threat*. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat

NIST. (n.d.). *cybersecurity—Glossary*. Retrieved May 14, 2024, from https://csrc.nist.gov/glossary/term/cybersecurity

Nowack, M., Endrikat, J., & Guenther, E. (2011). Review of Delphi-based scenario studies: Quality and design considerations. *Technological Forecasting and Social Change*, *78*(9), 1603–1615. https://doi.org/10.1016/j.techfore.2011.03.006

Nyre-Yu, M., Gutzwiller, R. S. P., & Caldwell, B. S. P. (2019). Observing Cyber Security Incident Response: Qualitative Themes From Field Research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 437–441. https://doi.org/10.1177/1071181319631016

OECD. (2024). *Explanatory memorandum on the updated OECD definition of an AI system* (OECD Artificial Intelligence Papers No. 8). OECD Publishing. https://doi.org/10.1787/623da898-en

Olabanji, S. O., Marquis, Y. A., Adigwe, C. S., Ajayi, S. A., Oladoyinbo, T. O., & Olaniyi, O. O. (2024). AI-Driven Cloud Security: Examining the Impact of User Behavior Analysis on Threat

Detection. *Asian Journal of Research in Computer Science*, *17*(3), 57–74. https://doi.org/10.9734/ajrcos/2024/v17i3424

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

Osterman Research. (2024). *SOC efficiency 2024*. https://20705827.fs1.hubspotusercontent-na1.net/hubfs/20705827/Docs/Osterman%20Research/2024-10-03%20Osterman%20Research%2c%20SOC%20efficiency%202024%20-%20Radiant%20Security.pdf

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Paling, G. (2025, July 1). *Modern SOC Series - Augmented not Autonomous: The future of AI-assisted SOC (1/4)*. https://www.orangecyberdefense.com/global/blog/innovation/modern-soc-series-augmented-not-autonomous-the-future-of-ai-assisted-soc-1/3

Parente, F. (2024, February 21). *AI Act and its impacts on the European financial sector—European Union*. https://www.eiopa.europa.eu/publications/ai-act-and-its-impacts-european-financial-sector_en

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. https://doi.org/10.1145/3586183.3606763

Paschou, V. (2024, September 30). *NIS2 vs. DORA: Differences and common misconceptions*. activeMind.Legal. https://www.activemind.legal/guides/nis2-dora/

Passador, M. L. (2024). *AI Act and the ECB: Steering Financial Supervision in the EU* (SSRN Scholarly Paper No. 4881451). https://doi.org/10.2139/ssrn.4881451

Petsani, D., Santonen, T., Merino-Barbancho, B., Epelde, G., Bamidis, P., & Konstantinidis, E. (2024). Categorizing digital data collection and intervention tools in health and wellbeing living lab settings: A modified Delphi study. *International Journal of Medical Informatics*, *185*, 105408. https://doi.org/10.1016/j.ijmedinf.2024.105408

Potti, S. (2024, April 9). *Make Google part of your security team anywhere you operate, with defenses supercharged by AI*. Google Cloud Blog. https://cloud.google.com/blog/products/identity-security/make-google-part-of-your-security-team-supercharged-by-ai-next24

Pradhan, R. P., Arvin, M. B., Nair, M., Bennett, S. E., & Hall, J. H. (2018). The dynamics between energy consumption patterns, financial sector development and economic growth in Financial Action Task Force (FATF) countries. *Energy*, *159*, 42–53. https://doi.org/10.1016/j.energy.2018.06.094

Prakash, A. V., & Das, S. (2021). Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: A mixed-methods study. *Information & Management*, *58*(7), 103524. https://doi.org/10.1016/j.im.2021.103524

Quantile. (2025). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Quantile&oldid=1305273839

Quantile function. (n.d.). RDocumentation. Retrieved July 14, 2025, from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile

Ravichandran, H. (2023, August 30). *AI Is Supercharging The Crisis Of Online Crime: How Defensive AI Safeguards Cybersecurity*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2023/08/30/ai-is-supercharging-the-crisis-of-online-crime-how-defensive-ai-safeguards-cybersecurity/

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on Digital Operational Resilience for the Financial Sector and Amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011, EP, CONSIL, OJ L (2022). http://data.europa.eu/eli/reg/2022/2554/oj/eng

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (2024). http://data.europa.eu/eli/reg/2024/1689/oj

Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on Prudential Requirements for Credit Institutions and Investment Firms and Amending Regulation (EU) No 648/2012, 176 OJ L (2013). http://data.europa.eu/eli/reg/2013/575/oj/eng

Rehberger, J. (2024). *Trust No AI: Prompt Injection Along The CIA Security Triad* (No. arXiv:2412.06090). arXiv. https://doi.org/10.48550/arXiv.2412.06090

Resti, A., Onado, M., Quagliariello, M., & Molyneux, P. (2021, June). *Shadow Banking: What kind of Macroprudential Regulation Framework?* European Union. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662925/IPOL_STU(2021)662925_EN.pdf

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed.). Free Press.

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

Saheb, T., & Mamaghani, F. H. (2021). Exploring the barriers and organizational values of blockchain adoption in the banking industry. *Journal of High Technology Management Research*, *32*(2). https://doi.org/10.1016/j.hitech.2021.100417

Sakana AI. (2024, August 13). *Sakana AI*. https://sakana.ai/

Samoili, S., Lopez-Cobo, M., Delipetrev, B., Plumed, F., Gómez, E., & De Prato, G. (2021). *AI Watch. Defining Artificial Intelligence 2.0. Towards an operational definition and taxonomy of AI for the AI landscape*. https://ai-watch.ec.europa.eu/publications/ai-watch-defining-artificial-intelligence-20_en

Sangwan, R. S., Badr, Y., & Srinivasan, S. M. (2023). Cybersecurity for AI Systems: A Survey. *Journal of Cybersecurity and Privacy*, *3*(2), Article 2. https://doi.org/10.3390/jcp3020010

Sarker, I. H. (2024). LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling. *Discover Artificial Intelligence*, *4*(1), 40. https://doi.org/10.1007/s44163-024-00129-0

Sarker, I. H., Janicke, H., Ferrag, M. A., & Abuadbba, A. (2024). Multi-aspect rule-based AI: Methods, taxonomy, challenges and directions towards automation, intelligence and transparent cybersecurity modeling for critical infrastructures. *Internet of Things*, *25*, 101110. https://doi.org/10.1016/j.iot.2024.101110

Schafer, N. (2024, April 11). *The Rise of AI Phishing and What it Means for the Future of Scammers*. https://www.mailgun.com/blog/email/ai-phishing/

Schatz, D., Bashroush, R., & Wall, J. (2017). Towards a More Representative Definition of Cyber Security. *The Journal of Digital Forensics, Security and Law*. https://doi.org/10.15394/jdfsl.2017.1476

Schmalz, U., Spinler, S., & Ringbeck, J. (2021). Lessons Learned from a Two-Round Delphi-based Scenario Study. *Methods X*, *8*. https://doi.org/10.1016/j.mex.2020.101179

Schneier, B. (2017, March 21). *Security Orchestration for an Uncertain World*. https://securityintelligence.com/security-orchestration-for-an-uncertain-world/

Seetharaman, N., Clemente, G., & Más, C. (2023, October 26). *Adapt or Die: Generative AI & The Revolution of American Cyber Defense*. https://www.wraithwatch.com/post/adapt-or-die-generative-ai-the-revolution-of-american-cyber-defense

Sele, D., & Chugunova, M. (2024). Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *PLOS ONE*, *19*(2), e0298037. https://doi.org/10.1371/journal.pone.0298037

Shah, A., Sinha, A., Ganesan, R., Jajodia, S., & Cam, H. (2020). Two Can Play That Game: An Adversarial Evaluation of a Cyber-Alert Inspection System. *ACM Transactions on Intelligent Systems and Technology*, *11*(3), 32:1-32:20. https://doi.org/10.1145/3377554

Shahzad, S. J. H., Hoang, T. H. V., & Arreola-Hernandez, J. (2019). Risk spillovers between large banks and the financial sector: Asymmetric evidence from Europe. *Finance Research Letters*, *28*, 153–159. https://doi.org/10.1016/j.frl.2018.04.008

Shaked, A., Cherdantseva, Y., Burnap, P., & Maynard, P. (2023). Operations-informed incident response playbooks: Computers & Security. *Computers & Security*, *134*, N.PAG-N.PAG. https://doi.org/10.1016/j.cose.2023.103454

Shneiderman, B. (2020a). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *10*(4), 26:1-26:31. https://doi.org/10.1145/3419764

Shneiderman, B. (2020b). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shreeve, B., Gralha, C., Rashid, A., Araújo, J., & Goulão, M. (2023). Making Sense of the Unknown: How Managers Make Cyber Security Decisions. *ACM Transactions on Software Engineering and Methodology*, *32*(4), 83:1-83:33. https://doi.org/10.1145/3548682

Shreeve, B., Hallett, J., Edwards, M., Anthonysamy, P., Frey, S., & Rashid, A. (2020). "So if Mr Blue Head here clicks the link..." Risk Thinking in Cyber Security Decision Making. *ACM Transactions on Privacy and Security*, *24*(1), 5:1-5:29. https://doi.org/10.1145/3419101

Shukla, A., Katt, B., & Yamin, M. M. (2023). A quantitative framework for security assurance evaluation and selection of cloud services: A case study. *International Journal of Information Security*, 1–30. https://doi.org/10.1007/s10207-023-00709-8

Silveira, A. (2023). Automated individual decision-making and profiling [on case C-634/21—SCHUFA (Scoring)]. *UNIO – EU Law Journal*, *8*(2), 74–85. https://doi.org/10.21814/unio.8.2.4842

Simonovich, V. (2025, June 17). Cato CTRL™ Threat Research: WormGPT Variants Powered by Grok and Mixtral. *Cato Networks*. https://www.catonetworks.com/blog/cato-ctrl-wormgpt-variants-powered-by-grok-and-mixtral/

Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi Method for Graduate Research. *Journal of Information Technology Education: Research*, *6*, 001–021. https://doi.org/10.28945/199

Sok, J., Borges, J. R., Schmidt, P., & Ajzen, I. (2021). Farmer Behaviour as Reasoned Action: A Critical Review of Research with the Theory of Planned Behaviour. *Journal of Agricultural Economics*, *72*(2), 388–412. https://doi.org/10.1111/1477-9552.12408

Solaimani, S., & Swaak, L. (2023). Critical Success Factors in a multi-stage adoption of Artificial Intelligence: A Necessary Condition Analysis. *Journal of Engineering and Technology Management*, *69*. https://doi.org/10.1016/j.jengtecman.2023.101760

Tamari, S., Shustin, R., & Riancho, A. (2024, September 26). *Wiz Research Finds Critical NVIDIA AI Vulnerability Affecting Containers Using NVIDIA GPUs, Including Over 35% of Cloud Environments | Wiz Blog*. Wiz.Io. https://www.wiz.io/blog/wiz-research-critical-nvidia-ai-vulnerability

Tevet, I. (2024, September 11). *AI in Action: Top 4 Ways Security Teams Can Leverage AI Today*. Intezer. https://intezer.com/blog/alert-triage/ways-security-teams-leverage-ai/

The Hacker News. (2024, June 24). *Ease the Burden with AI-Driven Threat Intelligence Reporting*. The Hacker News. https://thehackernews.com/2024/06/ease-burden-with-ai-driven-threat.html

Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022). Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3491102.3517732

Tologonov, J., & Fokker, J. (2024, March 6). *The Dark Side of Innovation: Cybercriminals and Their Adoption of GenAI*. https://www.trellix.com/blogs/research/the-dark-side-of-innovation-cybercriminals-and-their-adoption-of-genai/

Trevelyan, E. G., & Robinson, P. N. (2015). Delphi methodology in health research: How to do it? *European Journal of Integrative Medicine*, *7*(4), 423–428. https://doi.org/10.1016/j.eujim.2015.07.002

Triguero, I., Molina, D., Poyatos, J., Del Ser, J., & Herrera, F. (2024). General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance. *Information Fusion*, *103*, 102135. https://doi.org/10.1016/j.inffus.2023.102135

Ushakov, A. (2024, July 25). *GXC Team Unmasked: The cybercriminal group targeting Spanish bank users with AI-powered phishing tools and Android malware*. Group-IB. https://www.group-ib.com/blog/gxc-team-unmasked/

Vantrepotte, Q., Chambon, V., & Berberian, B. (2023). The reliability of assistance systems modulates the sense of control and acceptability of human operators: Scientific Reports. *Scientific Reports*, *13*(1), 1–11. https://doi.org/10.1038/s41598-023-41253-8

Veitch, E., Dybvik, H., Steinert, M., & Alsos, O. A. (2024). Collaborative Work with Highly Automated Marine Navigation Systems: Computer Supported Cooperative Work: The Journal of Collaborative Computing. *Computer Supported Cooperative Work: The Journal of Collaborative Computing*, *33*(1), 7–38. https://doi.org/10.1007/s10606-022-09450-7

Vielberth, M., Bohm, F., Fichtinger, I., & Pernul, G. (2020). Security Operations Center: A Systematic Study and Open Challenges. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.3045514

Visa Payment Fraud Disruption. (2023). *Visa Payment Fraud Disruption Biannual Threats Report December 2023*.

von Briel, F. (2018). The future of omnichannel retail: A four-stage Delphi study. *Technological Forecasting and Social Change*, *132*, 217–229. https://doi.org/10.1016/j.techfore.2018.02.004

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting and Social Change*, *79*(8), 1525–1536. https://doi.org/10.1016/j.techfore.2012.04.013

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Weger, K., Easley, T., Branham, N., Tenhundfeld, N., & Mesmer, B. (2022). Individual Differences in the Acceptance and Adoption of AI-enabled Autonomous Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 241–245. https://doi.org/10.1177/1071181322661154

Weger, K., Matsuyama, L., Zimmermann, R., Mesmer, B., Van Bossuyt, D., Semmens, R., & Eaton, C. (2023). Insight into User Acceptance and Adoption of Autonomous Systems in Mission Critical Environments. *International Journal of Human-Computer Interaction*, *39*(7), 1423–1437. https://doi.org/10.1080/10447318.2022.2086033

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Weigman, A. (2025, September 2). Hexstrike-AI: LLM Orchestration Driving Real-World Zero-Day Exploits. *Check Point Blog*. https://blog.checkpoint.com/executive-insights/hexstrike-ai-when-llms-meet-zero-day-exploitation/

What is the Apply Equivalent Subjects expander? (n.d.). Retrieved September 11, 2025, from https://connect.ebsco.com/s/article/What-is-the-Apply-Equivalent-Subjects-expander?language=en_US

Wiedemann, K. (2022). Profiling and (automated) decision-making under the GDPR: A two-step approach. *Computer Law & Security Review*, *45*, 105662. https://doi.org/10.1016/j.clsr.2022.105662

Wise, J. (2023, June 30). *Imagine your child calling for money. Except it's not them – it's an AI scam*. The Guardian. https://www.theguardian.com/commentisfree/2023/jun/30/money-ai-scam-fraud-fraudsters-trick

Wu, T., Terry, M., & Cai, C. J. (2022). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *CHI Conference on Human Factors in Computing Systems*, 1–22. CHI '22: CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3517582

Wu, Y., Liu, Y., Dai, T., & Cheng, D. (2024). Managing partial outsourcing on information security in the presence of security externality. *Expert Systems With Applications*, *246*. https://doi.org/10.1016/j.eswa.2023.123003

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

Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., & Wang, H. (2024). *Large Language Models for Cyber Security: A Systematic Literature Review*. https://doi.org/10.48550/ARXIV.2405.04760

Ye, X., Jo, W., Ali, A., Bhatti, S. C., Esterwood, C., Kassie, H. A., & Robert, L. P. (2024). Autonomy Acceptance Model (AAM): The Role of Autonomy and Risk in Security Robot Acceptance. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 840–849. https://doi.org/10.1145/3610977.3635005

Yuryna Connolly, A., & Borrion, H. (2022). Reducing Ransomware Crime: Analysis of Victims' Payment Decisions. *Computers & Security*, *119*. https://doi.org/10.1016/j.cose.2022.102760

Zhang, R., Li, H.-W., Qian, X.-Y., Jiang, W.-B., & Chen, H.-X. (2025). On large language models safety, security, and privacy: A survey. *Journal of Electronic Science and Technology*, 100301. https://doi.org/10.1016/j.jnlest.2025.100301

Zhao, W., Goyal, T., Chiu, Y. Y., Jiang, L., Newman, B., Ravichander, A., Chandu, K., Bras, R. L., Cardie, C., Deng, Y., & Choi, Y. (2024). *WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries* (No. arXiv:2407.17468). arXiv. https://doi.org/10.48550/arXiv.2407.17468

Zhuhadar, L. P., & Lytras, M. D. (2023). The Application of AutoML Techniques in Diabetes Diagnosis: Current Approaches, Performance, and Future Directions. *Sustainability*, *15*(18). https://doi.org/10.3390/su151813484

Appendix A Generative AI disclosure

During this study, the generative AI tool Petal [petal.org] has been used to summarize a set of 10 initial papers (Amershi et al., 2019; Cabitza et al., 2021; Cai et al., 2019; Hauptman et al., 2023; Hsieh, 2023; Liehner et al., 2022; Nyre-Yu et al., 2019; Rudin, 2019; Shreeve et al., 2023; Weger et al., 2023). Petal extracted the research question, research problem, research function, conclusions, type of study, methodology and limitations. Petal summarized the following concepts in those papers, if present: transparency, technology acceptance, level of autonomy, trust, explainability, security, decision making, risk, human-computer interaction. Information provided by Petal was manually verified and corrected where necessary.

Additionally, ChatGPT has been used to suggest rephrasing during the research proposal stage. The actual rephrasing was done manually by hand-picking some but not all the suggested edits. The rephrasing has been limited to about 20 sentences.

Multiple GenAI tools have been used as inspiration for Delphi statements using this query: “Delphi statements are future projections about technology. Write 10 Delphi statements about AI usage in cyber security for the blue team.” These suggestions have not been used.

Mistral chat has been used to find a peer reviewed source for Figure 2. Both Mistal and ChatGPT have been used to help write R code.

Appendix B Tables

Table B1

Recommendations for constructing Delphi statements

Issues concerning complexity
Formulate most statements with 10–20 words.
If respondents [=participants] are familiar with the topic of the statement, the length of the statement must be shorter.
If respondents are less familiar with the topic of the statement, the length of the statement should increase, allowing greater explanation.
Issues concerning vagueness and ambiguity
Use simple English. Statements must be understandable for respondents without English as their first language.
Consider the use of technical jargon, acronyms, and abbreviations carefully.
Avoid abstract language, slang, and emotional terms.
Avoid “compound events” [a] in statements.
Use numbers to indicate the state of development such as cost reduction of technology or market share (e.g., 50% market share), and not vague formulations such as “widespread use.”
Include area of application and/or market for the technology in quest.
Technical standards can be useful in defining both technology and areas of application.
Avoid, or consider carefully, anchoring in the statements, for example, by indicating the present situation.

Note. From “Constructing Delphi statements for technology foresight” by P. D. Andersen, 2023, *FUTURES & FORESIGHT SCIENCE*, *5*(2), p. 11 (https://doi.org/10.1002/ffo2.144).

[a] Compound events are statements that participants can agree with one part but not another.

Table B2

Identified factors from SQ1 and SQ2

Sub question	Inhibitor (-) / enabler (+)	Factors
SQ1	-	Social influence/recognition (Gursoy et al., 2019; Meyer-Waarden & Cloarec, 2022; Prakash & Das, 2021)
SQ1	+	Performance/effort expectancy (Meyer-Waarden & Cloarec, 2022; Prakash & Das, 2021; Solaimani & Swaak, 2023)
SQ1	+	Trust (in technology) (Meyer-Waarden & Cloarec, 2022; Prakash & Das, 2021; Sele & Chugunova, 2024; Weger et al., 2023)
SQ1	+	Safety/security expectancy (Meyer-Waarden & Cloarec, 2022; Weger et al., 2023)
SQ1	+	Behavioral control; control of the process (Kruse et al., 2022; Mohr & Kühl, 2021)
SQ1	+	Data sovereignty (Mohr & Kühl, 2021)
SQ1	+	Technical competencies and resources (Solaimani & Swaak, 2023)
SQ1	-	Risk: bias including discrimination, environmental harm (Araujo et al., 2020; Weger et al., 2023; Weidinger et al., 2022)
SQ1	+	Transparency (Barredo Arrieta et al., 2020; Kyriakou & Otterbacher, 2023; Sarker et al., 2024; Weger et al., 2023)
SQ1	+	Explainability (Barredo Arrieta et al., 2020; Kyriakou & Otterbacher, 2023; Sarker et al., 2024)
SQ1	+	Algorithmic fairness (Araujo et al., 2020; Barredo Arrieta et al., 2020)
SQ1	+	Reliability (Weger et al., 2023)
SQ1	+	Perceived usefulness (Weger et al., 2023)
SQ1	+	Shared situational awareness (Cleland-Huang et al., 2024; Fischer et al., 2021)
SQ1	+	Accountability (Barredo Arrieta et al., 2020; Kyriakou & Otterbacher, 2023)
SQ1	o	Responsibility (Tolmeijer et al., 2022)
SQ1	-	User's resistance to change: inertia, perceived threat, performance risk (Prakash & Das, 2021)
SQ1	-	Privacy concerns (Meyer-Waarden & Cloarec, 2022; Weidinger et al., 2022)

Sub question	Inhibitor (-) / enabler (+)	Factors
SQ1	-	Job displacement (Benedikt et al., 2020; Weidinger et al., 2022)
SQ1	-	Deskilling (Kosmyna et al., 2025; Veitch et al., 2024)
SQ1	o	Accuracy, including inaccuracy caused by LLM hallucinations (Sele & Chugunova, 2024; Zhao et al., 2024)
SQ1 / SQ2	+	Top management support (Abd Majid & Zainol Ariffin, 2019, 2021; Solaimani & Swaak, 2023; Vielberth et al., 2020)
SQ2	-	Lower response speed and longer time before detection (Brewer, 2019; Kokulu et al., 2019)
SQ2	+	Isolation of customer (i.e. customer of AI tooling) data (Bernsmed et al., 2015; Shukla et al., 2023)
SQ2	+	Auditing of services (Bernsmed et al., 2015; Shukla et al., 2023)
SQ2	+	Security logging, detection, response and recovery (Bernsmed et al., 2015; Shukla et al., 2023)
SQ2	-	Risk interdependency between different FIs using same vendor (Cezar et al., 2017; Feng et al., 2021)
SQ2	+	Competitive/security externality when outsourcing to MSSP (Cezar et al., 2017; Feng et al., 2021)
SQ2	-	Information leakage / data loss risk (Rehberger, 2024; Vielberth et al., 2020; Y. Wu et al., 2024)
SQ2	-	Shortage of skilled personnel (Brewer, 2019; Danquah, 2020; Vielberth et al., 2020)
SQ2	+	Technical accuracy of the analysis and consistency between analyses (Abd Majid & Zainol Ariffin, 2019; Bridges et al., 2023; Forsberg & Frantti, 2023)
SQ2		Cost of implementing and maintain versus outsourcing (Abd Majid & Zainol Ariffin, 2019; Vielberth et al., 2020)
SQ2	-	Time to set up a SOC internally (Vielberth et al., 2020)
SQ2	-	Regulations including privacy regulations (Vielberth et al., 2020)
SQ2	o	Availability (Vielberth et al., 2020)
SQ2	-	Double moral hazard (Feng et al., 2021)
SQ2	+	Efficiency during investigations of security incidents (Bridges et al., 2023)
SQ2	-	Dependency on a reliable internet connection (Bridges et al., 2023)

Table B3

Clustered factors

Inhibitor (-), enabler (+), neither (o)	Cluster [a]	Factors
+/+/+	Transparency	Transparency. Explainability. Algorithmic fairness.
o/+/+	*Accountability*	Accountability. Responsibility. Auditing of services.
+/+/-	*Situational awareness*	Shared situational awareness between human operators and AI. Behavioral control; control of the process. Double moral hazard: gaps in who takes responsibility (AI/human).
+/+	*Data*	Data sovereignty. Data isolation (i.e. data of customers of AI tooling).
-	*Regulations*	Regulations including privacy regulations (e.g., EU AI Act, GDPR, DORA – see SQ3)
-/-/-	Risk	Risk: bias including discrimination and environmental harm. Privacy concerns. Data loss risk.
-/o	*Risk interdependency*	Risk interdependency (i.e., sharing information and establishing trusted connections) between different FIs using the same vendor/service. Security externality (i.e. improve security) when using the same vendor service.
+/+/-/+	Perceived usefulness	Perceived usefulness. Performance/effort expectancy. Lower response speed. Efficiency during investigations.
+/o/+/o/-	Reliability	Reliability, among others supported by security processes: logging, detection, response and recovery. Accuracy (hallucinations). Technical accuracy and consistency of analyses. Availability. Dependency on internet connection.
+	Safety/security expectancy	Safety/security expectancy, among others supported by security process: logging, detection, response and recovery.
-	*Cost*	Implementation and maintenance cost, including time spent.
+/-/-/-/-	*Human resources*	Competencies and resources. Job displacement. Shortage of skilled personnel. Resistance to change. Deskilling.
+	Trust	Trust (in technology).
+	*Management support*	Top management support.

Inhibitor (-), enabler (+), neither (o)	Cluster [a]	Factors
o	*Social influence*	Social influence (e.g., customers, employees or other firms in financial sector).

Note. Some adjacent clusters are related to each other; those relations are marked by shaded cells.

[a] Factors printed in italic are not part of the MMCTAM as defined in section 2.3. See section 5.2 for the final version of the MMCTAM.

Table B4

Factor priority based on interview

Factor	Priority	Mentioned
Accountability	high	yes
Accuracy (hallucinations)	low	yes
Algorithmic fairness	medium	yes
Auditing of services	high	yes
Availability	low	yes
Behavioral control; control of the process	low	no
Competencies and resources	high	yes
Concentration risk	low	yes
Data isolation (i.e. data of customers of AI tooling)	medium	yes
Data loss risk	very high	yes
Data sovereignty	medium	yes
Dependency on internet connection	low	yes
Deskilling	high	yes
Double moral hazard: gaps in who takes responsibility (AI/human)	medium	yes
Efficiency during investigations	high	yes
Explainability	very high	yes
Implementation and maintenance cost, including time spent	high	yes
Job displacement	high	yes
Lower response speed	high	yes
Perceived usefulness	high	yes
Performance/effort expectancy	high	yes
Privacy concerns	very high	yes
Regulations including privacy regulations	high	yes
Reliability, among others supported by security processes	low	yes
Resistance to change	low	yes
Responsibility	high	yes
Risk interdependency between FIs using the same vendor/service	low	no
Risk: bias including discrimination and environmental harm	low	yes
Safety/security expectancy	high	yes
Shared situational awareness between human operators and AI	low	yes
Shortage of skilled personnel	low	no
Social influence (e.g., customers, employees or other firms in finance)	low	yes
Technical accuracy and consistency of analyses	low	yes
Top management support	low	yes
Transparency	high	no

Factor	Priority	Mentioned
Trust (in technology)	medium	no

Table B5

Factor coverage in scenarios of Delphi Round 1

Factor	Priority	Scenario [a]
Accountability	high	C, D
Accuracy (hallucinations)	low	
Algorithmic fairness	medium	
Auditing of services	high	C
Availability	low	
Behavioral control; control of the process	low	
Competencies and resources	high	C
Concentration risk	low	*B*
Data isolation (i.e. data of customers of AI tooling)	medium	
Data loss risk	very high	B
Data sovereignty	medium	
Dependency on internet connection	low	
Deskilling	high	C, D
Double moral hazard: gaps in who takes responsibility (AI/human)	medium	D
Efficiency during investigations	high	*A*
Explainability	very high	B, C
Implementation and maintenance cost, including time spent	high	
Job displacement	high	C
Lower response speed	high	*A*
Perceived usefulness	high	*A*, C
Performance/effort expectancy	high	*A*, C
Privacy concerns	very high	*B*
Regulations including privacy regulations	high	*A*
Reliability, among others supported by security processes	low	C
Resistance to change	low	C
Responsibility	high	D
Risk interdependency between FIs using the same vendor/service	low	
Risk: bias including discrimination and environmental harm	low	
Safety/security expectancy	high	A
Shared situational awareness between human operators and AI	low	D
Shortage of skilled personnel	low	C
Social influence (e.g., customers, employees or other firms in finance)	low	B
Technical accuracy and consistency of analyses	low	C
Top management support	low	
Transparency	high	B

Factor	Priority	Scenario [a]
Trust (in technology)	medium	D

[a] Scenarios in italic were potentially relevant.

Table B6

SQ3 factors identified from regulations

Factor	Requirements
Applicability of the AI Act to cyber security applications/services	Additional requirements for general-purpose AI systems with systemic risk
	Transparency requirements limiting the usage of powerful GPAI models
	Additional requirements for the usage of external (to the firm) or non-EU approved AI systems
	Copyright, possibly limiting GPAI model usage
Usage of GPAI with systemic or provider concentration risk	Additional requirements for general-purpose AI systems with systemic risk
	GPAI provider concentration risk
(Im)possibility of using non-EU models	Additional requirements for the usage of external (to the firm) or non-EU approved AI systems
Copyright issues when using third party models	Copyright, possibly limiting GPAI model usage
Applicability of GDPR Article 22	Requirement to explain decisions (Hamon et al., 2021; Wiedemann, 2022)

Table B7

Factor coverage in scenarios of Delphi Rounds 2 & 3

Factor [a]	Priority	Statement [a, b]	Median [d]
Ability to train AI tooling	- [c]	6	
Accountability	high	*11*	
Accuracy (hallucinations)	low	5	
Auditing of services	high	**13**	5
Competencies and resources	high	4	
Data loss risk	very high	7	
Deskilling	high	4	
Efficiency during investigations	high	**13, 14**	5, 3
Explainability	very high	**15**	6
Implementation and maintenance cost, including time spent	high	3	
Lower response speed	high	**8**	5
Perceived usefulness	high	**1**, ***8***	5, 5
Performance/effort expectancy	high	***8***	*5*
Regulations including privacy regulations	high	2	
Reliability	low	*5, 6,* ***16***	6
Responsibility	high	11	
Safety/security expectancy	high	**14**	3
Shared situational awareness between human operators and AI	low	**15**	6
Shortage of skilled personnel	low	4	
Technical accuracy and consistency of analyses	low	5, 6, **16**	6
Transparency	high	**15**	6
Trust (in technology)	medium	**9**, 10, **12**	4, 5

[a] Factors and statements in bold have reached consensus and stability.

[b] Statements in italic were included implicitly as a cluster. E.g., *Perceived usefulness* via *Lower response speed*.

[c] Factor *Ability to train AI* was added after the first round. Priority of factors was defined during the interview; therefore, no priority was established for factor *Ability to train AI*.

[d] Median for statements that reached consensus and stability.

Table B8

Round 2 median and IQR per statement

Statement	Median	IQR (R-8) [a]	IQR (R-6) [a]
1. In 2025, AI-driven cyber security tooling can support the decision-making of cyber security personnel.	5	**1.58**	**1.75**
2. Laws and regulations block the implementation of autonomous AI-driven cyber security tooling in the Dutch financial sector.	3.5	2.00	2.00
3. The implementation of AI-driven cyber security tooling needs to lead to operational cost savings.	3.5	2.00	2.00
4. Cyber security personnel needs to be more knowledgeable and experienced to work effectively with AI-driven cyber security tooling.	5	**0.58**	**0.75**
5. The analysis of AI-driven cyber security tooling is more accurate than the analysis of cyber security personnel.	3	**0.83**	**1.00**
6. AI-driven cyber security tooling needs to be continuously trained on the organization's business and IT information to be effective.	5	**1.17**	**1.50**
7. The AI-driven cyber security tooling has an inherent information leakage risk.	5	2.00	2.00
8. In 2028, your organization has to depend on autonomous AI-driven cyber security tooling to handle the volume (i.e., scale) of cyber attacks.	5	**1.17**	**1.50**
9. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on production systems in your organization.	4	**1.58**	**1.75**
10. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on critical production systems, like payment and banking.	3	2.58	2.75
11. A review of autonomous response actions by a human SOC analyst is always required.	4	2.17	2.50
12. Less than 25% of systems and services are out of bounds for autonomous AI-driven cyber security tooling.	5	**1.00**	**1.00**
13. The analysis of major security incidents is performed by AI-driven cyber security.	5	3.50	4.00
14. The AI shift is capable to perform all detection and response activities fully autonomously.	3	**0.58**	**0.75**
15. Human SOC analysts must be able to understand the decisions of the AI tooling.	6	**1.00**	**1.00**
16. In 2028, response actions that can disrupt business operations require human approval at all times (24/7).	6	2.67	3.00

[a] IQR ≤ 1.75 in bold.

Table B9

Delphi statement attitude towards AI in cyber defense

Positive	Neither positive nor negative	Negative
1. In 2025, AI-driven cyber security tooling can support the decision-making of cyber security personnel.	3. The implementation of AI-driven cyber security tooling needs to lead to operational cost savings.	2. Laws and regulations block the implementation of autonomous AI-driven cyber security tooling in the Dutch financial sector.
5. The analysis of AI-driven cyber security tooling is more accurate than the analysis of cyber security personnel.	4. Cyber security personnel needs to be more knowledgeable and experienced to work effectively with AI-driven cyber security tooling.	7. The AI-driven cyber security tooling has an inherent information leakage risk.
9. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on production systems in your organization.	6. AI-driven cyber security tooling needs to be continuously trained on the organization's business and IT information to be effective.	11. A review of autonomous response actions by a human SOC analyst is always required.
10. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on critical production systems, like payment and banking.	8. In 2028, your organization has to depend on autonomous AI-driven cyber security tooling to handle the volume (i.e., scale) of cyber-attacks.	16. In 2028, response actions that can disrupt business operations require human approval at all times (24/7).
12. Less than 25% of systems and services are out of bounds for autonomous AI-driven cyber security tooling. 13. The analysis of major security incidents is performed by AI-driven cyber security. 14. The AI shift is capable to perform all detection and response activities fully autonomously.	15. Human SOC analysts must be able to understand the decisions of the AI tooling.	

Appendix C Figures

C.1 Comparison between Round 2 and Round 3

The following figures show the choices participants made in each round, only if they did not choose *No comment* in either round.

Figure C1

Statement 1 Round 2 versus Round 3

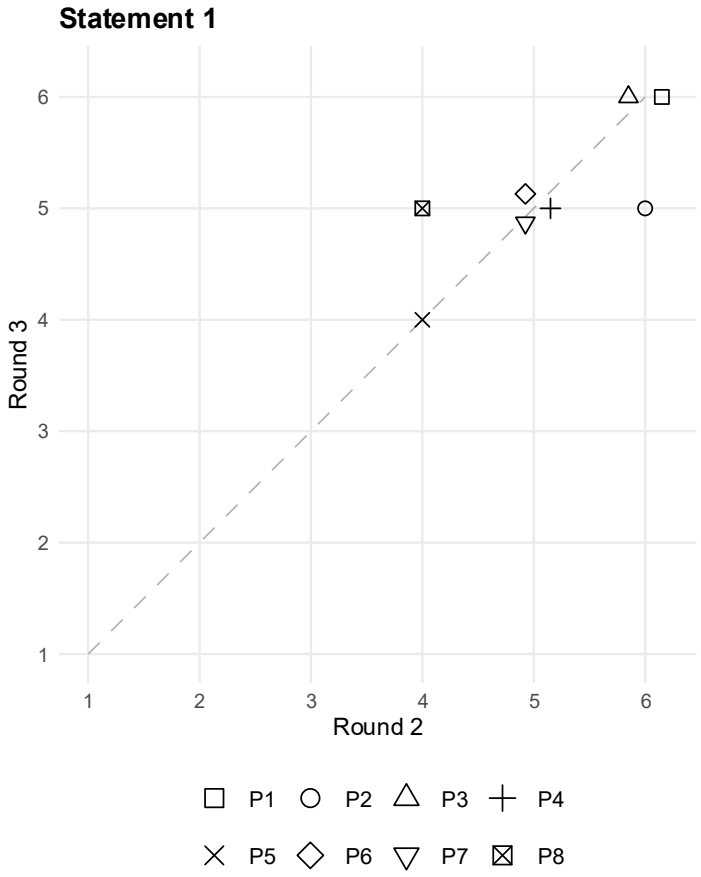


Figure C2

Statement 2 Round 2 versus Round 3

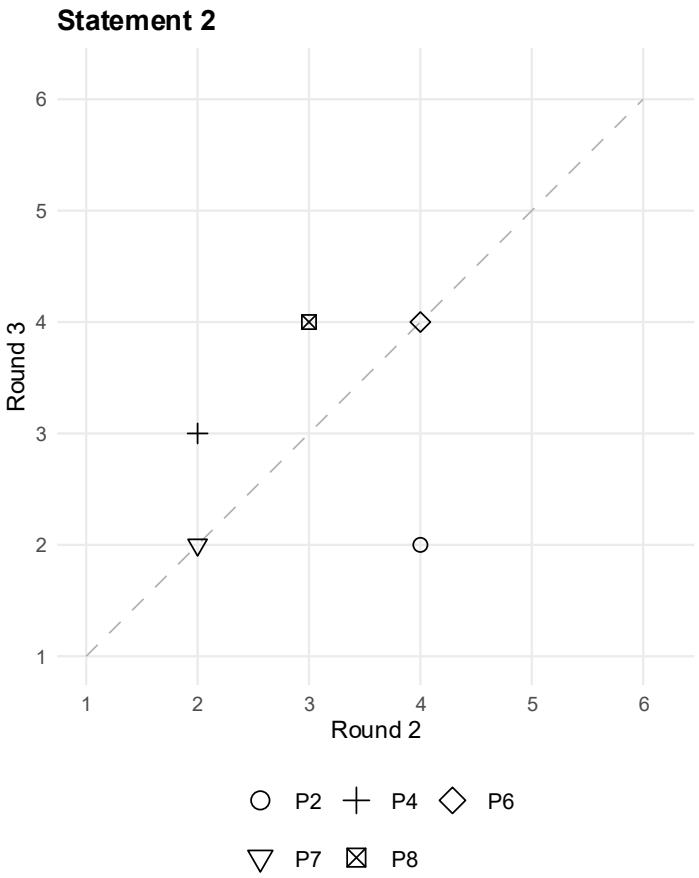


Figure C3

Statement 3 Round 2 versus Round 3

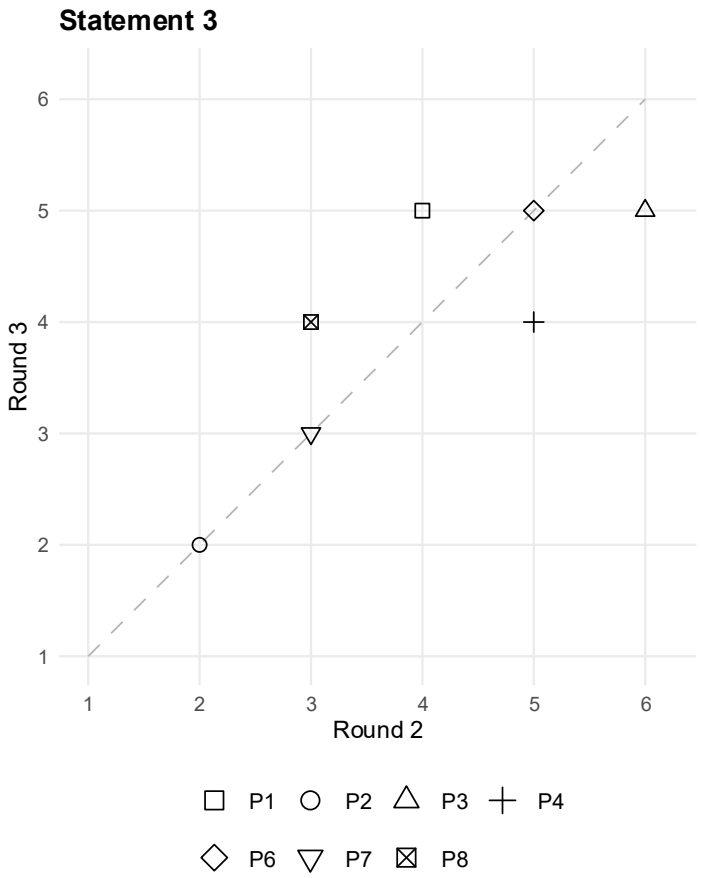


Figure C4

Statement 4 Round 2 versus Round 3

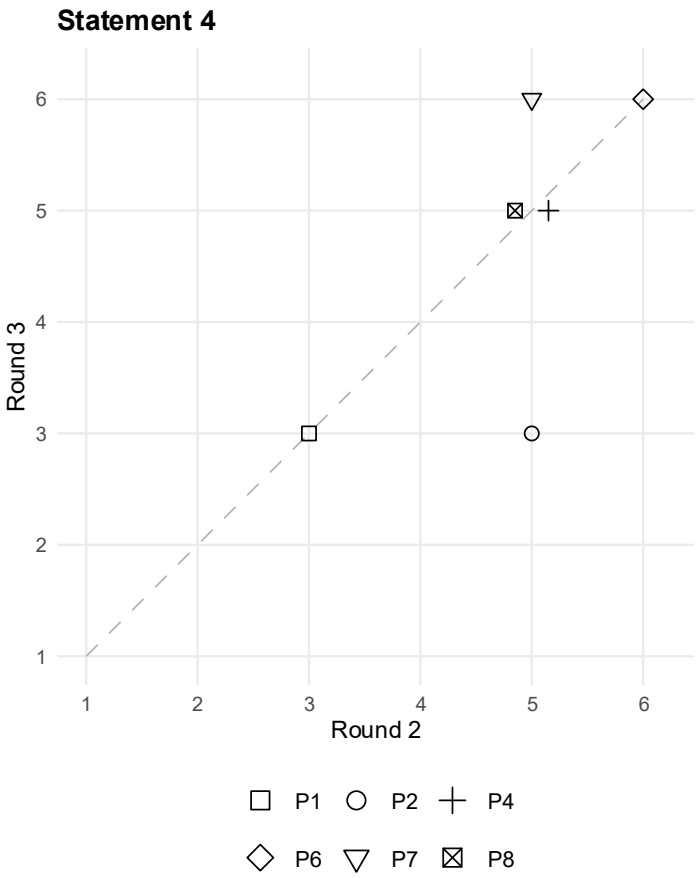


Figure C5

Statement 5 Round 2 versus Round 3

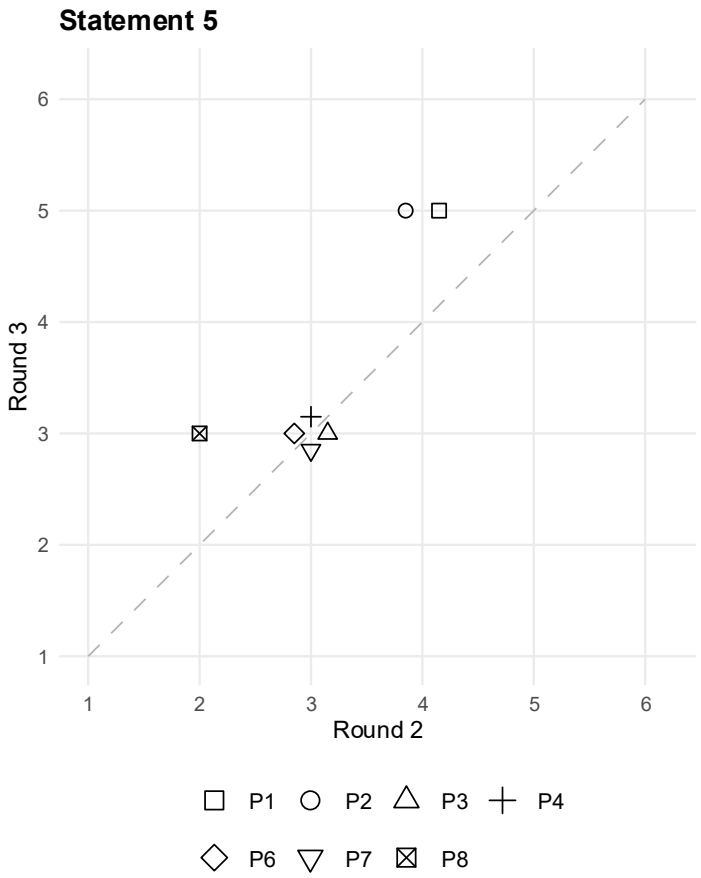


Figure C6

Statement 6 Round 2 versus Round 3

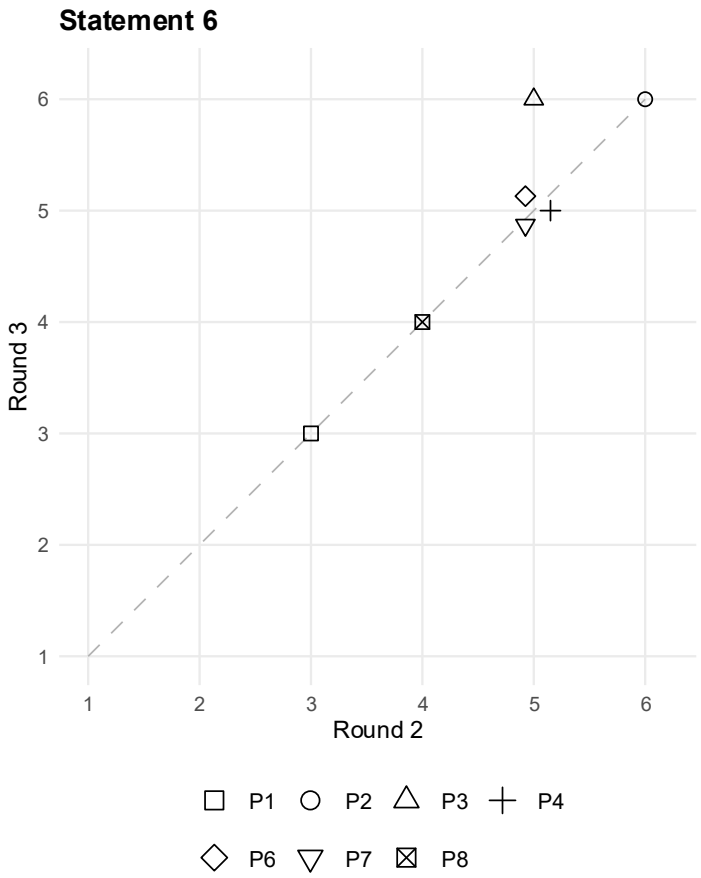


Figure C7

Statement 7 Round 2 versus Round 3

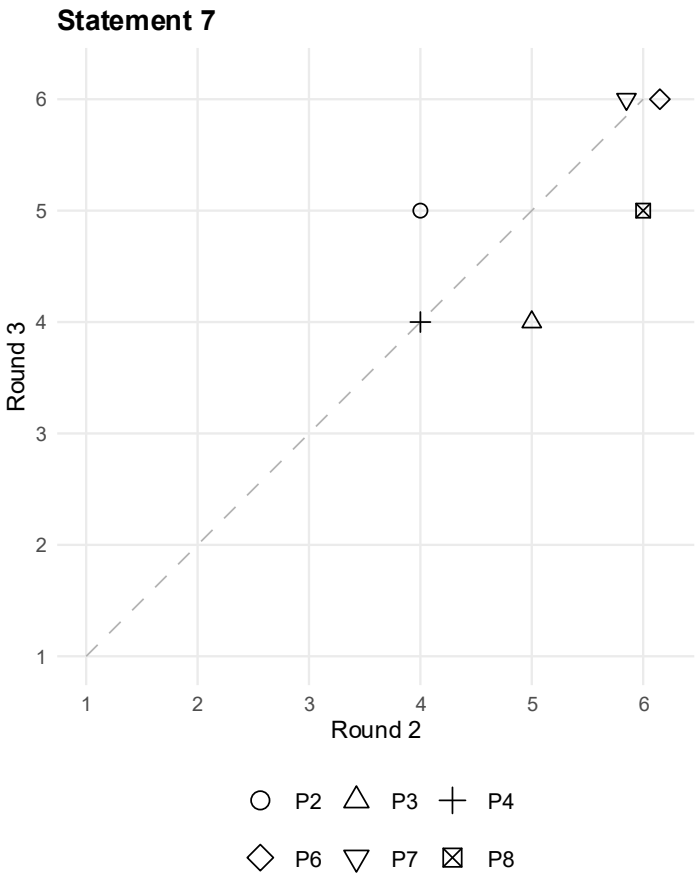


Figure C8

Statement 8 Round 2 versus Round 3

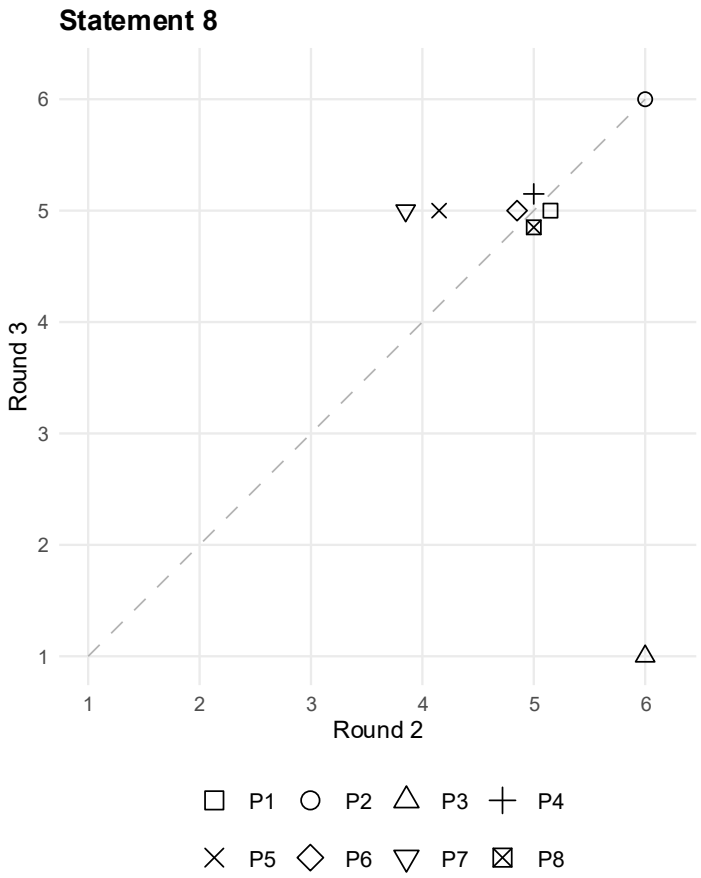


Figure C9

Statement 9 Round 2 versus Round 3

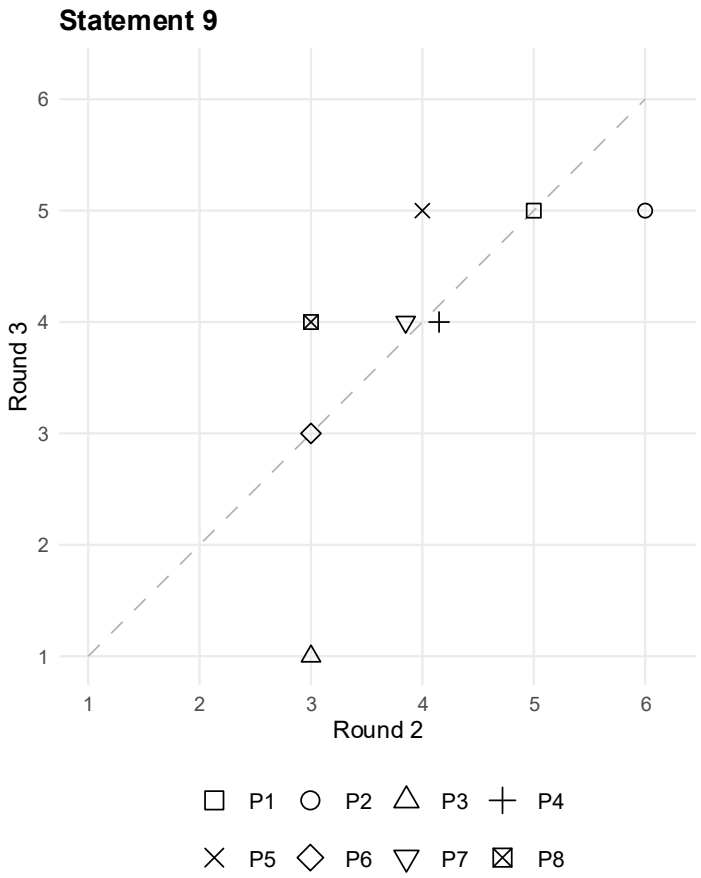


Figure C10

Statement 10 Round 2 versus Round 3

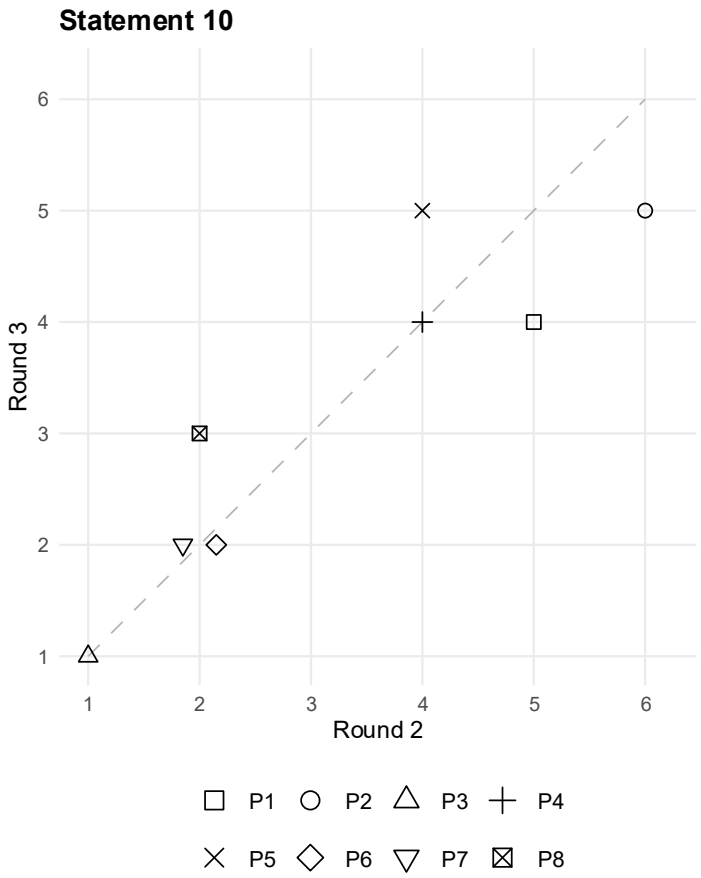


Figure C11

Statement 11 Round 2 versus Round 3

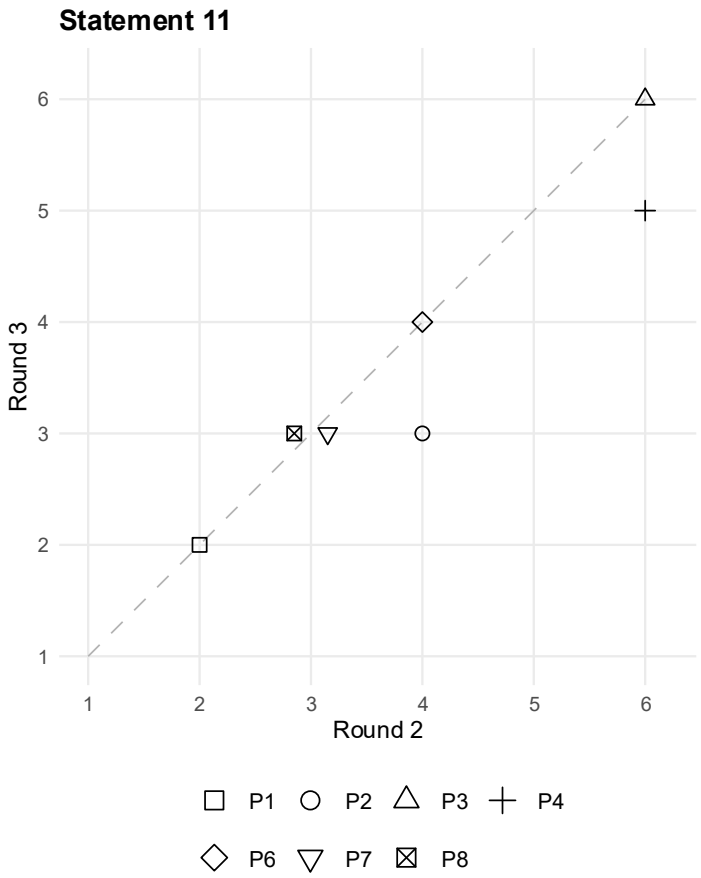


Figure C12

Statement 12 Round 2 versus Round 3

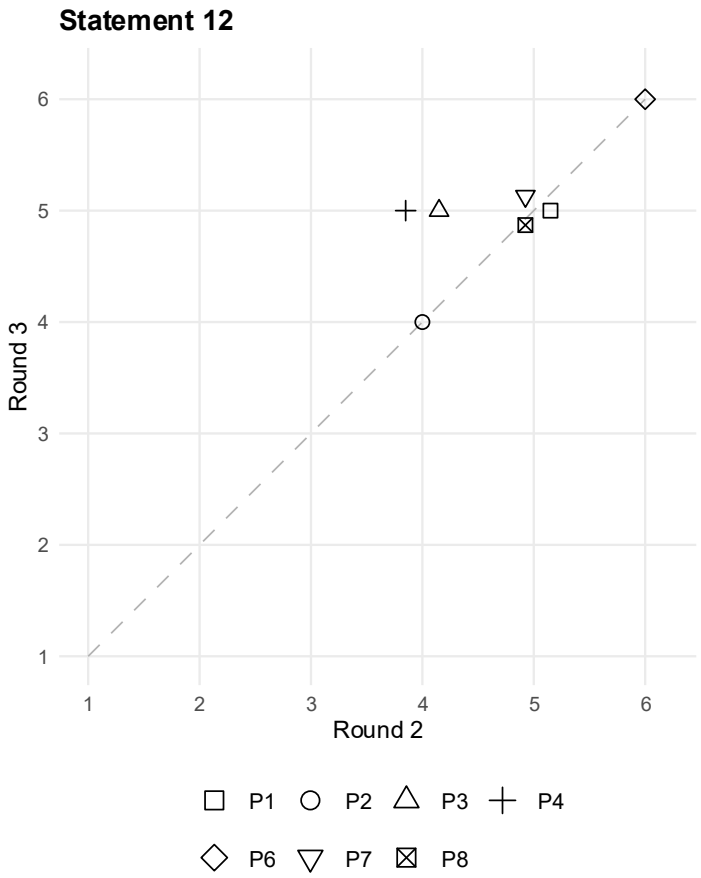


Figure C13

Statement 13 Round 2 versus Round 3

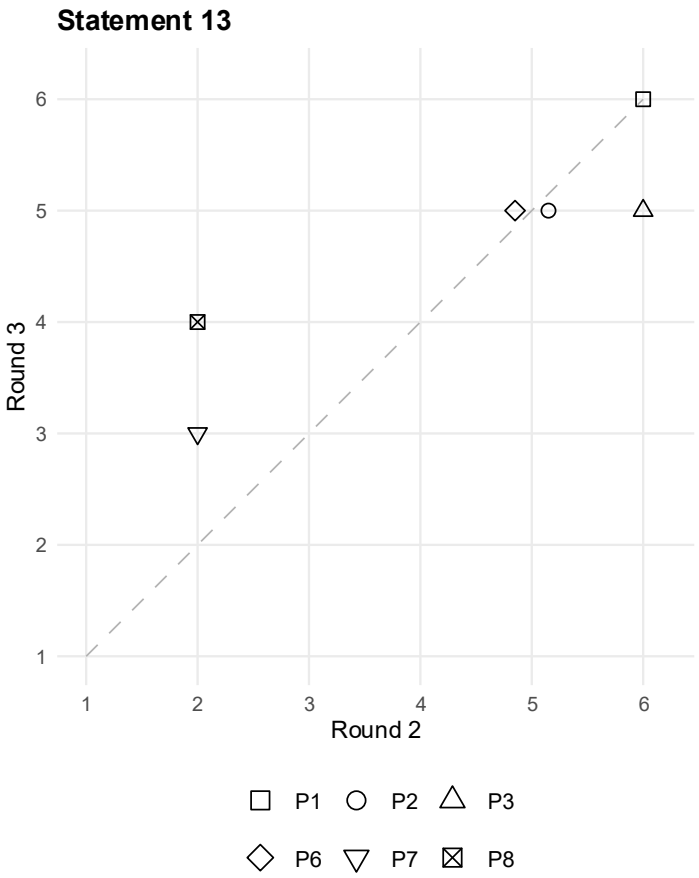


Figure C14

Statement 14 Round 2 versus Round 3

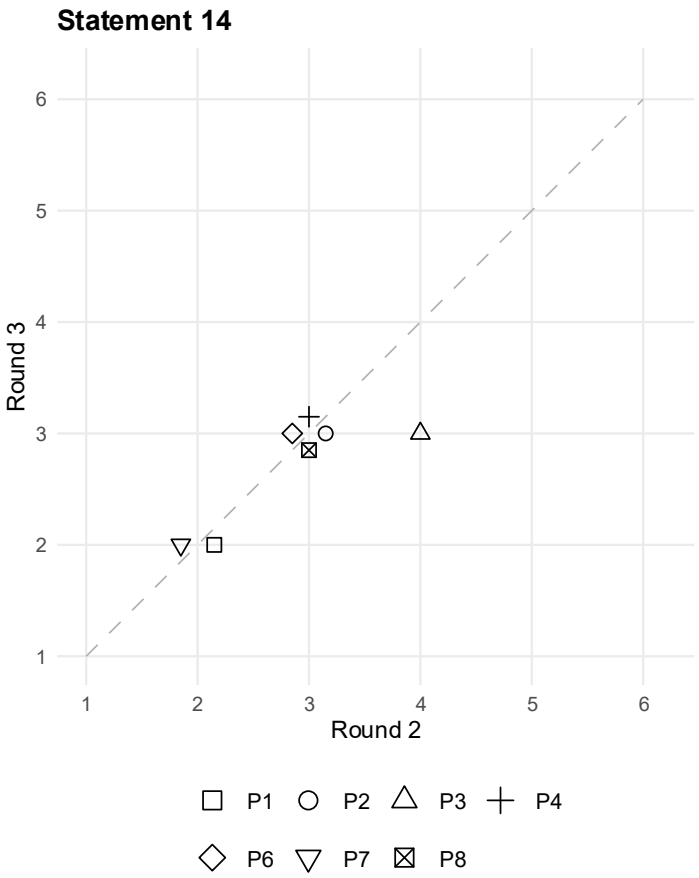


Figure C15

Statement 15 Round 2 versus Round 3

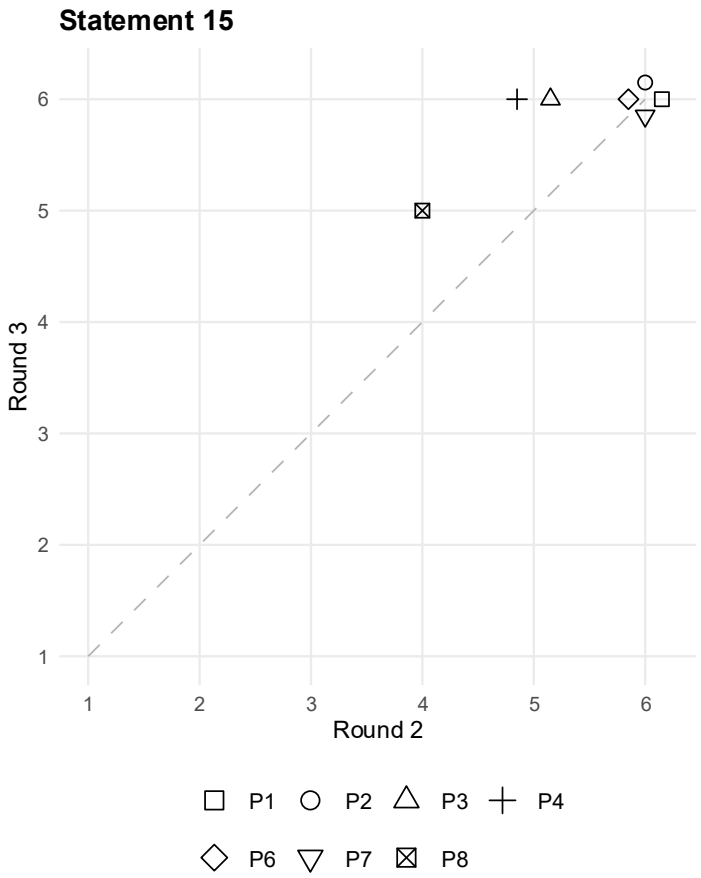
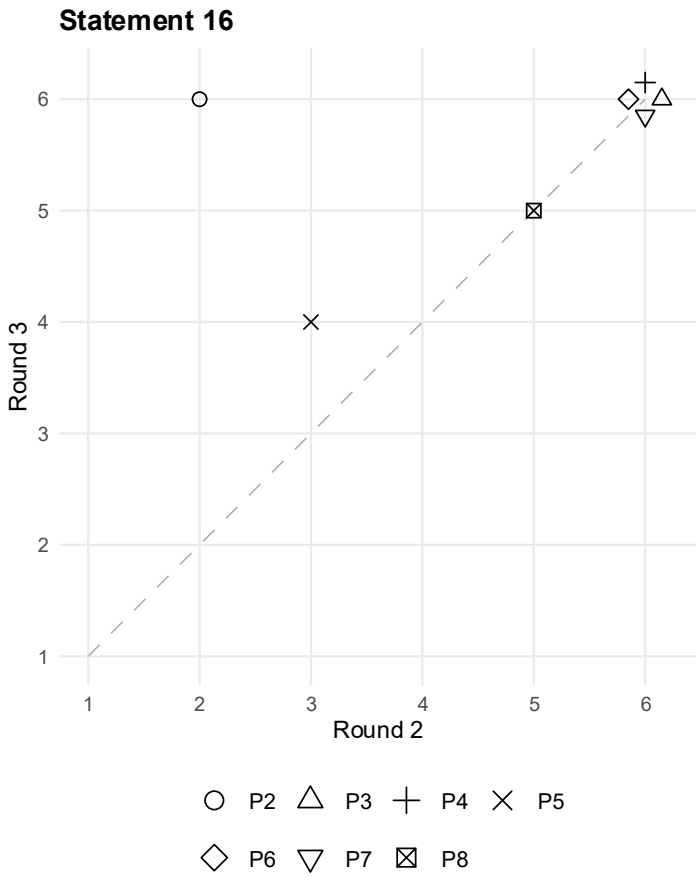


Figure C16

Statement 16 Round 2 versus Round 3



Appendix D Clustered Factors for Interview

The interviewer used the description of clustered factors below as a glossary during the interview.

D.1 Transparency

Transparency entails the visibility of the inner workings of a system or algorithm. Which steps does a system take to reach a conclusion or decide? The transparency of generative AI is currently lacking, making it difficult to judge whether GenAI systems are fair and unbiased. Decisions taken by AI are seen as fairer than those by human experts.

Explainability extends on transparency: are the inner workings and outcomes comprehensible to the stakeholders of the system?

The explicability of an AI system requires both comprehensibility and accountability. Accountability is the next factor.

D.2 Accountability and Situational Awareness

Accountability refers to the auditing and auditability of algorithms, data and design processes of AI systems and (AI-enabled) services. This is about governance: being in control of AI enabled processes or tools, and being able to support the claim of being in control in the financial sector, which is heavily regulated.

Besides the high-level, top-down view of governance, the autonomous actions themselves can be considered: Who bears the responsibility for the results of actions taken by autonomous agents? There could be gaps in who has or takes responsibility: human operator or AI?

Human operators and AI need to establish a shared situational awareness between them. That is, an overlap in knowledge of a situation is required, but this overlap in knowledge might blur the lines of responsibility. Human operators expect that an AI system can perform a certain behavior. In other words, human operators have a perception of control over the actions of agents.

D.3 Regulations

Within the financial sector complying with laws and regulations is paramount, regardless of using AI in an application or service. Regulations include AI and privacy regulations. Examples of current and upcoming regulations that might apply to autonomous AI in cyber security are the EU AI Act, GDPR, DORA.

D.4 Risk

Generative AI (GenAI) has inherent risks.

AI models can be trained on privacy sensitive information. That privacy sensitive information could leak. Information that is put into an AI system, for example by employees or customers, can be used to further train AI models leading to a virtuous circle.

Putting other confidential data in an AI system can lead to data loss.

Another risk associated with GenAI is that the training data is biased, or the AI model itself is biased due to other causes, which can lead to unfair and/or biased (for example discriminatory) output and unfair or biased decisions based on that output. Detecting this bias is difficult, because as noted under Transparency already, AI systems and especially GenAI lack transparency and explainability.

A final risk is the risk to the environment. This is mostly caused by energy usage and resource consumption; think huge data centers with up to millions of compute nodes (e.g., graphic cards) to train a model.

D.5 Risk Interdependency

Risk interdependence is the likelihood of a security breach at one firm leading to a breach at a different firm. Also, if other firms in the financial sector start using autonomous AI technology in cyber defense, the firms not using it could become a more likely target for hackers.

Firms using the same vendor or consulting firm for AI technology can raise risk interdependency, just as using the same managed security services provider could.

D.6 Data
Besides the loss of privacy information due to misuse of AI systems, concerns surrounding Data arise when using third-party AI systems. Firstly, data sovereignty: do you keep ownership of your business data? Secondly, data isolation: is data fully separated between different customers of those third-party AI systems?

D.7 Safety/security Expectancy
AI systems used for cyber security defense need themselves to be safe and secure to use, both hardware and software. The systems need to be resilient. The standard security processes of logging, detection, response and recovery support safe and secure usage.

D.8 Reliability
Analyses made by an AI system in a cyber security context need to be accurate and consistent. Due to the way large language models work, AI systems based on those LLMs can provide inaccurate information. Crudely put they generate the next word at each step. This inaccurate information includes hallucinations, also called confabulations.

A reliable AI system needs to have a high availability, or at least an availability consistent with its intended use. Availability usually depends among others on a reliable internet connection.

D.9 Perceived Usefulness
AI systems used for cyber security defense should be perceived as useful. The performance expected from such a tool should be a net positive, that is, it performs better than the systems already in place. For example, AI systems should respond quicker and increase efficiency during investigations of security events. The effort expectancy to start using or implementing such an AI system is a detractor: any new tool or system requires getting used to.

D.10 Human Resources
Technical competencies of the personnel using AI systems and human resources available for implementation could influence the successful implementation of AI systems. There is, however, a shortage of skilled personnel.

Introduction of autonomous AI can lead to job displacement. Employees might resist changes such as the introduction of autonomous AI. Introduction of autonomous AI can also lead to deskilling: human operators will be less experienced in the tasks that are performed by autonomous AI agents.

D.11 Cost
The cost of setting up a new system includes implementation cost, and maintenance cost. Also consider the time spent by engineers setting up such a system.

D.12 Trust
Trust in technology is crucial for the acceptance of that technology. People can have an initial, a priori, trust or distrust in technology. People can also overtrust AI systems even if those systems are unreliable.

D.13 Management Support
Top management support refers to commitment and approval from the board for new initiatives and technologies.

D.14 Social Influence
Customers, employees or other firms in the financial sector could influence the adoption of autonomous AI within a firm.

Appendix E Information letter

The following information letter was provided to participants. It is written in Dutch:

Informatiebrief “acceptatie van autonome ingrepen door AI in de cyber security-verdediging”

Oorspronkelijke titel van het onderzoek: Acceptance of AI-driven autonomous interventions in Cyber Security Defense.

E.1 Inleiding

U bent gevraagd om mee te doen aan een afstudeeronderzoek aan de opleiding Master of Informatics van de Hogeschool Utrecht. Het onderzoek vindt digitaal plaats. Er wordt één interview gehouden via Microsoft Teams; deze te interviewen deelnemer is al benaderd. Vervolgens worden online vragenlijsten ingevuld door alle deelnemers aan het onderzoek.

Meedoen is vrijwillig. Om mee te doen is wel uw schriftelijke toestemming nodig. Voordat u beslist of u wilt meedoen aan dit onderzoek, krijgt u uitleg over het onderzoek. Lees deze informatie rustig door en vraag de student om uitleg als u vragen heeft.

E.2 Algemene informatie

Dit onderzoek wordt uitgevoerd door de Hogeschool Utrecht.

E.3 Wat is het doel van het onderzoek?

Het doel van het onderzoek is het vinden van factoren die meespelen bij de acceptatie van autonoom acterende AI in de cyber security-verdediging (cyber security defense). Simpel gezegd, onder welke voorwaarden mag AI-gebaseerde security-applicatie of -systeem autonoom handelen? Het onderzoek richt zich op de financiële sector in Nederland.

E.4 Hoe wordt het onderzoek uitgevoerd?

Als vooronderzoek is een lijst opgesteld met factoren die mee zouden kunnen spelen. Deze factoren worden getoetst door één interview. Daarna wordt met een zogenaamde Delphi-studie geprobeerd om consensus te vinden - over óf factoren meespelen - in een kleine groep van experts. Participanten van de Delphi-studie vullen drie keer een vragenlijst in. De doorlooptijd van de Delphi-studie is grofweg drie maanden.

E.5 Waar u rekening mee dient te houden

Houd u er bij deelname aan de Delphi-studie rekening mee dat van u een tijdsinvestering van in totaal ongeveer een uur wordt verwacht. Het interview vraagt een additionele tijdsinvestering van ongeveer een uur van één deelnemer.

E.6 Wat gebeurt er als u niet wenst deel te nemen aan dit onderzoek?

U beslist zelf of u meedoet aan het onderzoek. Deelname is vrijwillig. Als u besluit niet mee te doen, kunt u dit aan ons laten weten. U hoeft niets te tekenen. U hoeft ook niet te zeggen waarom u niet wilt meedoen. Als u wel meedoet, kunt u zich altijd bedenken en toch stoppen. Ook tijdens het onderzoek.

De gegevens die tot dat moment over u zijn verzameld ten behoeve van het onderzoek, worden mogelijk gebruikt voor het onderzoek, de student zal hierover contact met u opnemen.

E.7 Einde van het onderzoek

Uw deelname aan het onderzoek stopt als:

- u zelf kiest om te stoppen
- uw bijdrage is afgerond en/of het einde van het onderzoek is bereikt.

E.8 Wat gebeurt er met uw gegevens en hoe wordt uw privacy geborgd?

Voor dit onderzoek worden uw persoonsgegevens (naam en emailadres) en uw onderzoeksgegevens verzameld en bewaard gedurende een periode van 7 jaar op de onderzoeklocatie om te voldoen aan wettelijk bepaalde bewaartermijnen. Indien het onderzoek gepubliceerd wordt als artikel in een wetenschappelijk tijdschrift, dan wordt de bewaartermijn verlengd tot: 10 jaar na publicatie van het artikel.

Uw gegevens worden gepseudonimiseerd. Tenminste het volgende wordt gedaan: persoonsgegevens (naam en emailadres) en verwijzingen naar de financiële instelling worden

vervangen door een uniek nummer; koppeling tussen dat nummer en deelnemer aan het onderzoek wordt los van de onderzoeksdata opgeslagen en wordt niet gepubliceerd.

Ook in rapportages zullen wij ervoor zorgen dat de gegevens gepseudonimiseerd zijn.

E.9 Heeft u vragen?

Bij vragen kunt u contact opnemen met de student.

Als u klachten of twijfels heeft over het onderzoek, kunt u dit bespreken met de student of een van de begeleiders. Wilt u dit liever niet, dan kunt u zich als het gaat om de verwerking van uw persoonsgegevens, wenden tot de Functionaris Gegevensbescherming via FG@hu.nl .

Als het om integriteit gaat, kunt u terecht bij de betreffende begeleiders van de student die het onderzoek uitvoert. U vindt de contactgegevens aan het einde van dit document.

E.10 Contactgegevens

1.	Student	Lennard van der Linden	<redacted: email address HU>
			<redacted: email address employer>
2.	Begeleiders	Peter Westerveld	<redacted: email address employer>
		Christian-Alexander Bunge	<redacted: email address HU>
3.	Functionaris Gegevensbescherming HU		<redacted: email address HU>

Voor meer informatie over uw rechten betreffende het gebruik van uw persoonlijke gegevens:

Autoriteit Persoonsgegevens (https://www.autoriteitpersoonsgegevens.nl/) of Privacybeleid van Hogeschool Utrecht (https://www.hu.nl/privacy).

Appendix F Delphi questionnaires

The following scenarios and statements were presented to the participants of the Delphi study.

F.1 Delphi Questionnaire Round One

The first round consisted of open questions.

F.1.1 Introduction of the Study

This study investigates the (technological) acceptance of autonomous AI tooling in the cyber security domain. The study is scoped to the Dutch financial sector.

This questionnaire contains four scenarios describing a hypothetical situation in the future. Each scenario has two to three open-ended questions. The study consists of three rounds in total, one questionnaire per round. The second and third questionnaire will use a Likert scale (from 'fully disagree' to 'fully agree').

Your responses will be depersonalized and solely used for research purposes. Your participation is voluntary, and you may withdraw at any time without any consequences.

1. I consent to data being collected as specified above and as specified in the provided information letter.
2. Email address
   Your email address will solely be used to send invitations for the next two rounds and provide customized feedback in the third round (i.e., your response in the second round).

F.1.2 Scenario A

A Nation State Actor has developed an advanced AI toolset. After successful initial compromise, this toolset enables them to automatically perform multiple steps on the cyber security kill chain including data exfiltration.

3. Can you name three reasons why defending against this actor **can** be handled autonomously - without human intervention - by AI-powered cyber defense?
4. Can you name two reasons why this **cannot** be handled autonomously by AI-powered cyber defense?

F.1.3 Scenario B

Another firm in the Dutch financial sector uses an AI tool that detects insider threats and prevents any suspicious sharing of data by autonomously blocking co-worker and non-personal accounts, and blocking their access to documents. The tool is a black box.

5. What reasons would you have to onboard this tool?
6. What reasons would you have to **not** onboard this tool?

F.1.4 Scenario C

Imagine that your security operations center (SOC) is detecting and responding fully autonomously by using AI agents. There is no human in the loop.

Think of an AI agent as an AI equivalent of a human SOC analyst.

7. Can you list three requirements for allowing AI agents to respond fully autonomously on production systems?

A production system is part of the live, operational environment that offers products and services to external and internal customers as opposed to a development or (user acceptance) test system.

8. Are there any additional requirements for critical production systems, like payment and banking?
9. What added value would an autonomous SOC provide? Can you list two benefits?

F.1.5 Scenario D

Now imagine a human SOC analyst is supervising these AI agents and can veto the autonomous actions of AI agents.

Vetoing means that a human is overseeing the decisions of AI agents and can correct mistakes when they happen. The human can revert a decision after the fact, but does not approve every decision upfront.

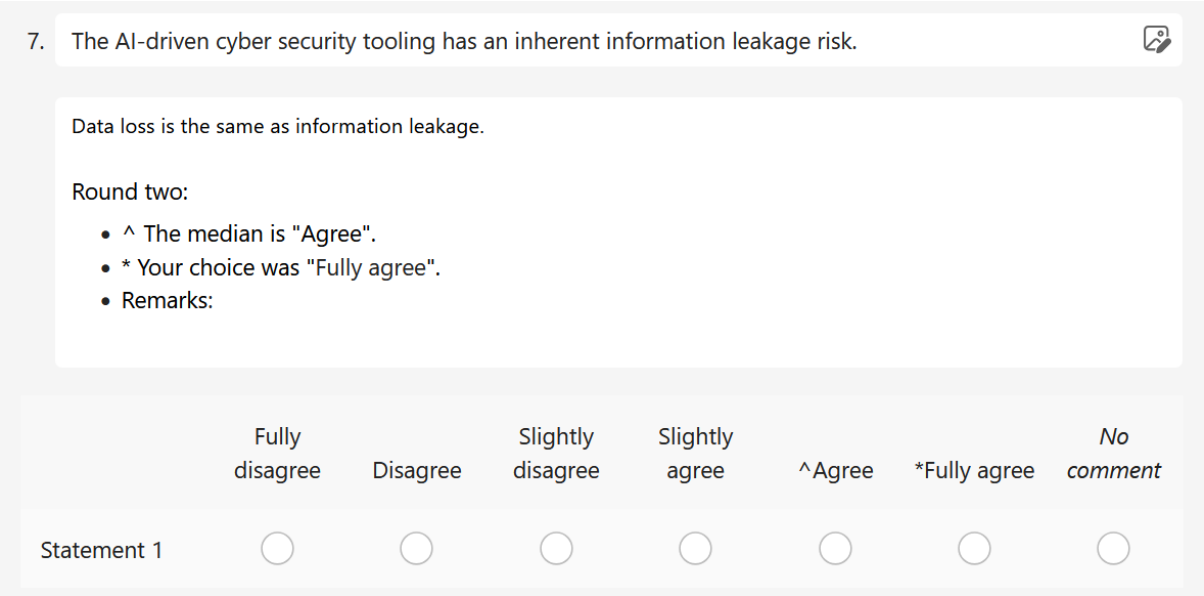
10. Can you list two benefits of having the possibility to veto?
11. Can you list two downsides of having the possibility to veto?
12. Who would be responsible for a specific action being taken? The SOC analyst, AI agent, or another option? Can you explain why in one or two sentences?

F.2 Delphi Questionnaire Round Two and Three

The second and third round questionnaires use the same scenarios and statements. Round three adds the median of the previous round, the participant's choice in the previous round, a clarification of the statement based on the remarks if needed, and a summary of the remarks, if applicable.

Figure F1

Example scenario and statement from questionnaire



The actual content of the questionnaire was as follows.

F.2.1 Introduction of the study – round 2

This questionnaire will present 6 scenarios. Each scenario has 1 to 3 multiple choice statements for a total of 16 statements. Please choose the option you think fits each statement best. If you have no opinion on the statement, please pick "no comment".

Each scenario has a "Remark" input field. This field is entirely optional and can be skipped. In case the scenario or statement(s) are unclear, or if you want to add context to your choice(s), please use the "Remark" field.

The questionnaire is designed to take less than 10 minutes to complete.

F.2.2 Introduction of the study – round 3

This questionnaire presents the same 6 scenarios and 16 statements as the previous round. Please choose the option you think fits each statement best. If you have no opinion on the statement, please pick "no comment". You can leave any remarks you have at the end of the questionnaire.

The following information from the previous round is shown for each statement:

- The most chosen option (i.e., the median) of the previous round. The '^' symbol will be used to mark the median. If the median is in between two answer options, both will be marked.
- Your choice in the previous round. The '*' symbol will be used to mark your previous choice.
- A clarification of the statement based on the remarks, if needed.
- A summary of the remarks, if applicable.

The questionnaire is designed to take less than 10 minutes to complete. It is the final round of this Delphi study.

F.2.3 Scenario 1

1. In 2025, AI-driven cyber security tooling can support the decision-making of cyber security personnel.

That means there is always a human in the loop. The AI is *not* autonomous.

Round two:

- ^ The median is "Agree".
- * Your choice was "".
- Remarks:
  - The word 'Support' is key here.
  - Enables quick analysis of data leading to quicker conclusions.
  - Wrong decisions [supported by AI] can have a high impact.

F.2.4 Scenario 2

Imagine that in 2026 your organization wants to implement fully autonomous AI-driven cyber security tooling.

2. Laws and regulations block the implementation of autonomous AI-driven cyber security tooling in the Dutch financial sector.

Clarification: 'Laws and regulations' includes what a regulator would or would not allow.

Round two:

- ^ The median is between "Slightly disagree" and "Slightly agree".
- * Your choice was "".
- Remarks:
  - EU legislations seem not unreasonable.

3. The implementation of AI-driven cyber security tooling needs to lead to operational cost savings.

[Clarification:] That is, the operational expenses after implementation.

Round two:

- ^ The median is between "Slightly disagree" and "Slightly agree".
- * Your choice was "".
- Remarks:
  - AI is very suitable for SOC/SIEM processes.

4. Cyber security personnel needs to be more knowledgeable and experienced to work effectively with AI-driven cyber security tooling.

Round two:

- ^ The median is "Agree".
- * Your choice was "".
- Remarks:
  - Personnel can focus on larger problems and trends.
  - More knowledgeable staff is needed to review edge cases and to monitor that AI actions are in line with policies.
  - Consider an AI body of knowledge and AI certification for personnel.

F.2.5 Scenario 3

Scenario 3 expands on scenario 2: Imagine that in 2026 your organization wants to implement fully autonomous AI-driven cyber security tooling.
For the following statements, assume that the AI tooling uses **generative AI** (i.e., large language models) under the hood.

5. The analysis of AI-driven cyber security tooling is more accurate than the analysis of cyber security personnel.

Potential causes for inaccuracy mentioned in the first round:

- Bias and hallucinations.
- The AI tooling is not aware of the latest techniques used by threat actors.

Round two:

- ^ The median is "Slightly disagree".
- * Your choice was "".
- Remarks:
  - AI has the potential to be more accurate, but difficult to predict a year in the future.
  - Human interpretation is still required for the final conclusion.

6. AI-driven cyber security tooling needs to be continuously trained on the organization's business and IT information to be effective.

In other words, a generic off-the-shelf tool is not sufficient. For example, a generic tool might not prevent business disruptions.

Round two:

- ^ The median is "Agree".
- * Your choice was "".
- Remarks:
  - Continuous training is key for contextual awareness and up-to-date knowledge.
  - Off-the-shelf tooling is expected to have strong protections in its own.

7. The AI-driven cyber security tooling has an inherent information leakage risk.

Data loss is the same as information leakage.

Round two:

- ^ The median is "Agree".
- * Your choice was "".
- Remarks:

F.2.6 Scenario 4

Imagine that in your organization AI-driven cyber security tooling can and does respond to security events and incidents fully autonomously.

8. In 2028, your organization has to depend on autonomous AI-driven cyber security tooling to handle the volume (i.e., scale) of cyber attacks.

Round two:

- ^ The median is "Agree".
- * Your choice was "".
- Remarks:
  - This would not surprise me with the speed AI is evolving.
  - There will be outliers that require human involvement.

9. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on **production systems** in your organization.

That means there is no human in the loop.

Round two:

- ^ The median is "Slightly agree".
- * Your choice was "".
- Remarks:
  - This would not surprise me with the speed AI is evolving.
  - Regulations require human involvement.

10. In 2028, AI-driven cyber security tooling responds (i.e., acts) fully autonomously on critical production systems, like payment and banking.

Round two:

- ^ The median is "Slightly disagree".
- * Your choice was "".
- Remarks:
  - This would not surprise me with the speed AI is evolving.
  - Regulations require human involvement.

F.2.7 Scenario 5

Imagine that in 2028 your organization has to depend on AI-driven cyber security tooling that is responding autonomously due to an increased volume and diversity of attacks.

11. A review of autonomous response actions by a human SOC analyst is always required.

The review happens after the response action has been taken by the AI-driven security tooling.

Round two:

- ^ The median is "Slightly agree".
- * Your choice was "".
- Remarks:
  - A follow up question is whether the analyst is capable to perform the review.
  - Not always, but this should happen on samples.

12. Less than 25% of systems and services are out of bounds for autonomous AI-driven cyber security tooling.

Round two:

- ^ The median is "Agree".
- * Your choice was "".
- Remarks:
  - I follow the 80/20 rule here. However, more data is needed to be more precise.

13. The analysis of major security incidents is performed by AI-driven cyber security.

Round two:

- ^ The median is "Agree".
- * Your choice was "".
- Remarks:
  - Major security incidents always require human analysis, which could be supported by AI.
  - AI tooling should speed up decision making and recommend actions.
  - The analysis process and output should be understandable and reproducible for a human.

F.2.8 Scenario 6

In 2028, your organization moves to a mixed AI/human SOC. On weekdays during office hours the human SOC analysts are on shift, outside that window the autonomous AI-driven cyber security tooling performs all detection and response activities.

14. The AI shift is capable to perform all detection and response activities fully autonomously.

Round two:

- ^ The median is "Slightly disagree".
- * Your choice was "".
- Remarks:
  - Mission critical impact always requires human review.
  - A human should always review based on samples and monitor the health of the AI 24/7.

15. Human SOC analysts must be able to understand the decisions of the AI tooling.

That is, the decisions of the AI tooling are explainable and transparent.

Round two:

- ^ The median is "Fully agree".
- * Your choice was "".
- Remarks:
  - Not only understand, but also be able reproduce the analysis.

16. In 2028, response actions that can disrupt business operations require human approval at all times (24/7).

Round two:

- ^ The median is "Fully agree".
- * Your choice was "".
- Remarks:

Appendix G Transcription

The transcription has been added as a separate file.

Appendix H Calculations in R

The IQR for Round 2 has been calculated using the following R code. Both type 6 (used by Excel and SPSS) and type 8 (Hyndman) have been calculated. The calculation for Round 3 used the same code, but different values as listed in H.2.

H.1 R Code Round 2

```
# Define a function to calculate quantiles and IQR
calculate_quantiles_and_iqr <- function(values) {
  # Calculate quantiles and IQR for type 6
  quantiles_6 <- quantile(values, probs = c(0.25, 0.5, 0.75), type = 6)
  IQR_6 <- IQR(values, type = 6)

  # Calculate quantiles and IQR for type 8 (Hyndman)
  quantiles_8 <- quantile(values, probs = c(0.25, 0.5, 0.75), type = 8)
  IQR_8 <- IQR(values, type = 8)

  # Return the results as a data frame row
  data.frame(
    IQR_T6 = IQR_6,
    Q1_T6 = quantiles_6[1],
    Median_T6 = quantiles_6[2],
    Q3_T6 = quantiles_6[3],
    IQR_R8 = IQR_8,
    Q1_R8 = quantiles_8[1],
    Median_R8 = quantiles_8[2],
    Q3_R8 = quantiles_8[3]
  )
}

# Define multiple sets of values
value_sets <- list(
c(6, 6, 6, 5, 4, 5, 5, 4),
c(4, 2, 4, 4, 2, 3),
c(4, 2, 6, 5, 3, 5, 3, 3),
c(3, 5, 5, 5, 6, 6, 5, 5),
c(4, 4, 3, 3, 3, 3, 2),
c(3, 6, 5, 5, 6, 5, 5, 4),
c(4, 5, 4, 3, 6, 6, 6),
c(5, 6, 6, 5, 4, 5, 4, 5),
c(5, 6, 3, 4, 4, 3, 4, 3),
c(5, 6, 1, 4, 4, 2, 2, 2),
c(2, 4, 6, 6, 4, 4, 3, 3),
c(5, 4, 4, 4, 5, 6, 5, 5),
c(6, 5, 6, 4, 5, 2, 2),
c(2, 3, 4, 3, 3, 3, 2, 3),
c(6, 6, 5, 5, 6, 6, 6, 4),
c(2, 6, 6, 3, 6, 6, 5)
)

# Initialize an empty data frame to store the results
results <- data.frame(matrix(ncol = 9, nrow = 0), stringsAsFactors = FALSE)
colnames(results) <- c("Seq", "IQR_T6", "Q1_T6", "Median_T6", "Q3_T6",
                       "IQR_R8", "Q1_R8", "Median_R8", "Q3_R8")

# Loop through each set of values and calculate quantiles and IQR
for (i in seq_along(value_sets)) {
  values <- value_sets[[i]]
  row_results <- calculate_quantiles_and_iqr(values)
  row_results$Seq <- i
  results <- rbind(results, row_results)
}

# Print the header
cat(sprintf("%-5s %-12s %-10s %-14s %-10s %-12s %-10s %-14s %-10s\n",
            "Seq", "IQR_T6", "Q1_T6", "Median_T6", "Q3_T6",
            "IQR_R8", "Q1_R8", "Median_R8", "Q3_R8"))

# Print the results table with each dataset on a single line but tabulated
for (i in 1:nrow(results)) {
  cat(sprintf("%-5d %-12.2f %-10.2f %-14.2f %-10.2f %-12.2f %-10.2f %-14.2f %-10.2f\n",
              results$Seq[i], results$IQR_T6[i], results$Q1_T6[i], results$Median_T6[i],
results$Q3_T6[i],
              results$IQR_R8[i], results$Q1_R8[i], results$Median_R8[i], results$Q3_R8[i]))
}
```

H.2 R value set of Round 3

```
# Define multiple sets of values
value_sets <- list(
c(6, 5, 6, 5, 4, 5, 5, 5),
c(2, 4, 3, 4, 2, 4),
c(5, 2, 5, 4, 5, 3, 4),
c(3, 3, 5, 6, 6, 5),
```

```
c(5, 5, 3, 3, 3, 3, 3),
c(3, 6, 6, 5, 5, 5, 4),
c(3, 5, 4, 4, 6, 6, 5),
c(5, 6, 1, 5, 5, 5, 5, 5),
c(5, 5, 1, 4, 5, 3, 4, 4),
c(4, 5, 1, 4, 5, 2, 2, 3),
c(2, 3, 6, 5, 4, 3, 3),
c(5, 4, 5, 5, 6, 5, 5),
c(6, 5, 5, 4, 5, 3, 4),
c(2, 3, 3, 3, 3, 2, 3),
c(6, 6, 6, 6, 6, 6, 5),
c(4, 6, 6, 6, 4, 6, 6, 5)
)
```

H.3 Likert-plot

The following code was used to plot second or third round separately.

```
library(data.table)
library(ggplot2)
library(ggstats)
library(bbplot)
library(svglite)
library(dplyr)

# change to R2 or R3
data <- data.frame(
   response = read.csv("R3.csv", header=TRUE, stringsAsFactors=FALSE)
 )
levels_order=c("Fully disagree","Disagree","Slightly disagree","No comment","Slightly
agree","Agree","Fully agree")
data <- data %>% mutate(across(everything(), ~factor(., levels = levels_order)))

#head(data)
names(data) <- gsub("response\\.", "", names(data))
names(data) <- gsub("X", "S", names(data))

p <- gglikert(
  #totals_hjust to fix vertical line through totolas)
  data,
  #totals_hjust = 0.17,
  #exclude_fill_values = "No comment",
  totals_include_center = FALSE,
  totals_size = 3,
  totals_fontface = "plain", # not bold
  symmetric=TRUE
) +guides(fill=guide_legend(nrow=2,byrow=TRUE))
p

ggsave("likert_plot_tmp.svg", plot = p, width = 16, height = 16, units = "cm")
```

The following code was used to plot second and third round together.

```
# Load required libraries
library(ggstats)
library(dplyr)
library(readr)   # for read_csv() if needed
library(tidyr)

# Step 1: Read data (use read.csv() as you initially wanted)
data <- read.csv("R2R3.csv", stringsAsFactors = FALSE)
names(data) <- gsub("X", "S", names(data))

# Step 2: Set Likert scale levels (adjust if needed)
likert_levels <- c("Fully disagree", "Disagree", "Slightly disagree", "No comment", "Slightly
agree", "Agree", "Fully agree")

# Step 3: Convert all Likert response columns to ordered factors
# Assume all columns except the last one are Likert items
question_cols <- names(data)[-ncol(data)]
data[question_cols] <- lapply(data[question_cols], function(x) {
  factor(x, levels = likert_levels, ordered = TRUE)
})

# Step 4: Convert group column to factor (optional but preferred)
group_col <- names(data)[ncol(data)]
data[[group_col]] <- as.factor(data[[group_col]])

# Step 5: Plot using gglikert with grouping by 'group' column
p <-gglikert(
  data,
  S1:S8,  # questions to plot
#  S9:S16,  # questions to plot
  y = group_col,                      # group column
  facet_rows = vars(.question),       # facet per question
```

```
  totals_include_center = FALSE,
  totals_size = 3,
  totals_fontface = "plain", # not bold
  symmetric=TRUE
) + guides(fill=guide_legend(nrow=2,byrow=TRUE))
# +  labs(title = "Likert Responses by Round per Statement")
p
ggsave("likert_plot_tmp.svg", plot = p, width = 16, height = 16, units = "cm")
```

H.4 Wilcoxon matched pairs signed rank test

The following shows the code used to do a Wilcoxon matched pairs signed rank test, using an exact distribution and the Pratt zero method.

```
> # Round 2 versus Round 3
library(coin)

# Example: List of 16 paired test cases (replace with your actual vectors)
# Each item is a list with two numeric vectors: x and y
test_cases <- list(
list(x=c(6,6,6,5,4,5,5,4),y=c(6,5,6,5,4,5,5,5)),
list(x=c(4,2,6,5,5,3,3),y=c(5,2,5,4,5,3,4)),
list(x=c(3,5,5,6,5,5),y=c(3,3,5,6,6,5)),
list(x=c(4,4,3,3,3,3,2),y=c(5,5,3,3,3,3,3)),
list(x=c(3,6,5,5,5,5,4),y=c(3,6,6,5,5,5,4)),
list(x=c(4,5,4,6,6,6),y=c(5,4,4,6,6,5)),
list(x=c(5,6,6,5,4,5,4,5),y=c(5,6,1,5,5,5,5,5)),
list(x=c(5,6,3,4,4,3,4,3),y=c(5,5,1,4,5,3,4,4)),
list(x=c(5,6,1,4,4,2,2,2),y=c(4,5,1,4,5,2,2,3)),
list(x=c(2,4,6,6,4,3,3),y=c(2,3,6,5,4,3,3)),
list(x=c(5,4,4,4,6,5,5),y=c(5,4,5,5,6,5,5)),
list(x=c(6,5,6,5,2,2),y=c(6,5,5,5,3,4)),
list(x=c(2,3,4,3,3,2,3),y=c(2,3,3,3,3,2,3)),
list(x=c(6,6,5,5,6,6,4),y=c(6,6,6,6,6,6,5)),
list(x=c(2,6,6,3,6,6,5),y=c(6,6,6,4,6,6,5))
)

# Initialize empty results list
results <- data.frame(question = integer(),
Z = numeric(),
p = numeric())

# Loop through each test case
for (i in seq_along(test_cases)) {
case <- test_cases[[i]]
w <- wilcoxsign_test(case$x ~ case$y, distribution = "exact", zero.method = "Pratt")
z <- w@statistic@teststatistic
p <- pvalue(w)
results <- rbind(results, data.frame(question = i, Z = z, p = p))
}

# Print the results table
print(results)
```

H.5 Compare Round 2 to Round 3

The following code generates plots that compare Round 2 and Round 3 per statement.

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(tibble)

# Statements
statements <- list(
list(x=c(6,6,6,5,4,5,5,4),y=c(6,5,6,5,4,5,5,5),pa=c("P1","P2","P3","P4","P5","P6","P7","P8"))
,
list(x=c(4,2,4,2,3),y=c(2,3,4,2,4),pa=c("P2","P4","P6","P7","P8")),
list(x=c(4,2,6,5,5,3,3),y=c(5,2,5,4,5,3,4),pa=c("P1","P2","P3","P4","P6","P7","P8")),
list(x=c(3,5,5,6,5,5),y=c(3,3,5,6,6,5),pa=c("P1","P2","P4","P6","P7","P8")),
list(x=c(4,4,3,3,3,3,2),y=c(5,5,3,3,3,3,3),pa=c("P1","P2","P3","P4","P6","P7","P8")),
list(x=c(3,6,5,5,5,5,4),y=c(3,6,6,5,5,5,4),pa=c("P1","P2","P3","P4","P6","P7","P8")),
list(x=c(4,5,4,6,6,6),y=c(5,4,4,6,6,5),pa=c("P2","P3","P4","P6","P7","P8")),
list(x=c(5,6,6,5,4,5,4,5),y=c(5,6,1,5,5,5,5,5),pa=c("P1","P2","P3","P4","P5","P6","P7","P8"))
,
list(x=c(5,6,3,4,4,3,4,3),y=c(5,5,1,4,5,3,4,4),pa=c("P1","P2","P3","P4","P5","P6","P7","P8"))
,
list(x=c(5,6,1,4,4,2,2,2),y=c(4,5,1,4,5,2,2,3),pa=c("P1","P2","P3","P4","P5","P6","P7","P8"))
,
list(x=c(2,4,6,6,4,3,3),y=c(2,3,6,5,4,3,3),pa=c("P1","P2","P3","P4","P6","P7","P8")),
list(x=c(5,4,4,4,6,5,5),y=c(5,4,5,5,6,5,5),pa=c("P1","P2","P3","P4","P6","P7","P8")),
list(x=c(6,5,6,5,2,2),y=c(6,5,5,5,3,4),pa=c("P1","P2","P3","P6","P7","P8")),
list(x=c(2,3,4,3,3,2,3),y=c(2,3,3,3,3,2,3),pa=c("P1","P2","P3","P4","P6","P7","P8")),
list(x=c(6,6,5,5,6,6,4),y=c(6,6,6,6,6,6,5),pa=c("P1","P2","P3","P4","P6","P7","P8")),
list(x=c(2,6,6,3,6,6,5),y=c(6,6,6,4,6,6,5),pa=c("P2","P3","P4","P5","P6","P7","P8"))
)
```

```
# Create consistent shape mapping for all PA labels
all_pa <- unique(unlist(lapply(statements, function(tc) tc$pa)))
shape_map <- setNames(0:(length(all_pa) - 1), all_pa)

# Jitter function
get_jittered_df <- function(x, y, pa) {
  df <- tibble(x = x, y = y, pa = pa) %>%
    mutate(point = factor(row_number()))

  df <- df %>%
    group_by(x, y) %>%
    mutate(
      dup_count = n(),
      dup_index = row_number()
    ) %>%
    ungroup()

  get_offsets <- function(n) {
    if (n == 1) return(tibble(dx = 0, dy = 0))
    angle <- seq(0, 2 * pi, length.out = n + 1)[- (n + 1)]
    radius <- 0.15
    tibble(dx = radius * cos(angle), dy = radius * sin(angle))
  }

  df <- df %>%
    group_by(x, y) %>%
    mutate(
      dx = get_offsets(dup_count[1])$dx[dup_index],
      dy = get_offsets(dup_count[1])$dy[dup_index],
      x_jitter = x + dx,
      y_jitter = y + dy
    ) %>%
    ungroup()

  df
}

# Plot each case
for (i in seq_along(statements)) {
  case <- statements[[i]]
  df_jitter <- get_jittered_df(case$x, case$y, case$pa)

  # Use only shape values for the current case's PAs
  current_pa <- unique(df_jitter$pa)
  shape_values <- shape_map[current_pa]

  diag_df <- data.frame(x = 1:6, y = 1:6)

  p <- ggplot(df_jitter, aes(x = x_jitter, y = y_jitter)) +
    geom_line(data = diag_df, aes(x = x, y = y),
              inherit.aes = FALSE, color = "gray70", linetype = "dashed", size = 0.2) +
    geom_point(aes(shape = pa), size = 3, color = "black") +
    scale_shape_manual(
      values = shape_values,
      labels = names(shape_values),
      name = NULL
    ) +
       # limits more than 6.0 to allow for jitter
    scale_x_continuous(breaks = 1:6, limits = c(1, 6.2)) +
    scale_y_continuous(breaks = 1:6, limits = c(1, 6.2)) +
    coord_fixed() +
    labs(
      title = paste("Statement", i),
      x = "Round 2",
      y = "Round 3"
    ) +
    theme_minimal(base_size = 12) +
    theme(
      legend.position = "bottom",
      legend.direction = "horizontal",
      legend.title = element_blank(),
      legend.text = element_text(size = 11),
      legend.key.size = unit(0.8, "cm"),
      plot.title = element_text(face = "bold", size = 14),
      legend.box = "horizontal",
      panel.grid.minor = element_blank()
    ) +
    guides(shape = guide_legend(
      nrow = 2,
      byrow = TRUE,
      override.aes = list(size = 4, color = "black")
    ))

  filename <- paste0("likert_plot", i, ".svg")
  ggsave(filename, plot = p, width = 16, height = 16, units = "cm")
}
```

H.6 Result

The R code displays the tabulated data as shown in Table H1 for Round 2 and Table H2 for Round 3.

Table H1

Second round IQR, quartiles and median

Seq	IQR_R6	Q1_R6	Median_R6	Q3_R6	IQR_R8	Q1_R8	Median_R8	Q3_R8
1	1.75	4.25	5.00	6.00	1.58	4.42	5.00	6.00
2	2.00	2.00	3.50	4.00	2.00	2.00	3.50	4.00
3	2.00	3.00	3.50	5.00	2.00	3.00	3.50	5.00
4	0.75	5.00	5.00	5.75	0.58	5.00	5.00	5.58
5	1.00	3.00	3.00	4.00	0.83	3.00	3.00	3.83
6	1.50	4.25	5.00	5.75	1.17	4.42	5.00	5.58
7	2.00	4.00	5.00	6.00	2.00	4.00	5.00	6.00
8	1.50	4.25	5.00	5.75	1.17	4.42	5.00	5.58
9	1.75	3.00	4.00	4.75	1.58	3.00	4.00	4.58
10	2.75	2.00	3.00	4.75	2.58	2.00	3.00	4.58
11	2.50	3.00	4.00	5.50	2.17	3.00	4.00	5.17
12	1.00	4.00	5.00	5.00	1.00	4.00	5.00	5.00
13	4.00	2.00	5.00	6.00	3.50	2.33	5.00	5.83
14	0.75	2.25	3.00	3.00	0.58	2.42	3.00	3.00
15	1.00	5.00	6.00	6.00	1.00	5.00	6.00	6.00
16	3.00	3.00	6.00	6.00	2.67	3.33	6.00	6.00

Table H2

Third round IQR, quartiles and median

Seq	IQR_T6	Q1_T6	Median_T6	Q3_T6	IQR_R8	Q1_R8	Median_R8	Q3_R8
1	1.00	5.00	5.00	6.00	0.83	5.00	5.00	5.83
2	2.00	2.00	3.50	4.00	2.00	2.00	3.50	4.00
3	2.00	3.00	4.00	5.00	1.83	3.17	4.00	5.00
4	3.00	3.00	5.00	6.00	3.00	3.00	5.00	6.00
5	2.00	3.00	3.00	5.00	1.67	3.00	3.00	4.67
6	2.00	4.00	5.00	6.00	1.67	4.17	5.00	5.83
7	2.00	4.00	5.00	6.00	1.83	4.00	5.00	5.83
8	0.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00
9	2.00	3.00	4.00	5.00	1.67	3.17	4.00	4.83
10	2.00	2.00	3.00	4.00	2.00	2.00	3.00	4.00
11	2.00	3.00	3.00	5.00	1.83	3.00	3.00	4.83
12	0.00	5.00	5.00	5.00	0.00	5.00	5.00	5.00
13	1.00	4.00	5.00	5.00	1.00	4.00	5.00	5.00
14	1.00	2.00	3.00	3.00	0.83	2.17	3.00	3.00
15	0.00	6.00	6.00	6.00	0.00	6.00	6.00	6.00
16	1.00	5.00	6.00	6.00	0.83	5.17	6.00	6.00