UNIVERSITY OF TURKU
Turku School of Economics

Explaining trust and acceptance of AI in digital security

The influence of explainable AI on trust and technology acceptance in cybersecurity tools.

Cybersecurity/Turku School of Economics (TSE) &
Tilburg School of Economics and Management (TiSEM)
Master's thesis

Author(s):
T.G. Haesen

Supervisor(s):
I.F.Kanellopoulos

13.02.2025
Vught

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Master's thesis

Subject: Explainable AI, Cybersecurity, AI, Technology acceptance model, Trust
Author(s): Twan Haesen
Title: Explaining trust and acceptance of AI in digital security
Supervisor(s): Ioannis Filippos Kanellopoulos
Number of pages: 55 pages + appendices 25 pages
Date: 13.02.2025

Abstract

Explainable AI (XAI) has gained significant recognition in recent years, shaping the development of artificial intelligence (AI) systems by promoting transparency and interpretability. As AI-driven cybersecurity tools become increasingly more widespread, understanding the influences of XAI on user trust and technology acceptance is crucial. This study examines the relationship between XAI, trust, and the acceptance of cybersecurity tools, aiming to determine whether the presence of XAI enhances trust and facilitates greater adoption of AI-driven security measures.

To investigate these relationships, an online survey was conducted with 52 valid respondents, ranging in age from 18 to 66 (M = 35.08, SD = 15.25). The questionnaire used scales established in previous research for measuring trust and technology acceptance, with participants randomly assigned to scenarios involving a cybersecurity AI tool for phishing with or without XAI-based explanations. The results indicate a significant positive relationship between trust and technology acceptance, reinforcing previous findings that trust plays a critical role in user adoption of AI technologies. However, contrary to expectations, the presence of XAI did not strengthen this relationship. This unexpected finding suggests that explanations provided by XAI may not always be intuitive or beneficial to users, potentially due to information overload, cognitive complexity, or the lack of a clear and actionable explanation format.

These findings highlight the role of XAI in cybersecurity applications and challenge the assumption that increased explainability always leads to higher trust. Instead, they suggest that the effectiveness of XAI in increasing trust may depend on factors such as the complexity of the AI model, the clarity of the explanations provided, and the technical expertise of the end user. This study contributes to the growing body of research on human-AI interaction, emphasizing the need for further investigation into the design of explainability methods that balance transparency, usability, and trustworthiness in cybersecurity AI. Future research should explore alternative XAI approaches, assess the impact of contextual factors, and examine how different user demographics respond to AI explanations in real-world security settings.

Keywords: Explainable AI, Cybersecurity, Technology acceptance, Trust, Human-AI interaction, Artificial intelligence transparency

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

In this modern time, artificial intelligence (AI) is becoming a key technology that is connected to nearly everything that is happening in the information systems industry (Sarker et al., 2021). Every day, more users are subjected to the use of AI, for example people are using tools like ChatGPT to finish their work quicker or Grammarly to check their grammar. And not just these tools are seeing an increase in their use. Generative AI, a subset of AI mainly focused on generating new content like text or images from training data (Feuerriegel et al,. 2024), is seeing an increase in use among researchers with a 11% growth in use when researchers are facing barriers (Dorta-González et al., 2024). AI use by individuals is expected to keep on increasing as AI is rapidly being developed and keeps on becoming more essentials for users (Li et al., 2023).
From a business perspective, AI can have a different purpose, for example AI is mainly being used for analysing large amount of data like with big data (Kok et al., 2009). Another task for which AI is extensively employed is pattern recognition (Bharadiya et al., 2023). The AI systems used for these tasks are developed using a large amount of training data gathered from mostly historical data.

Furthermore, there are many different AI techniques that are being used by businesses, with some of the more widely used techniques being machine learning and deep learning. Machine learning is a subfield of AI originally defined by Arthur Samuel (1959) and focuses on giving computers a way to learn from data. Deep learning is an alternative to machine learning and uses even more data to get a more accurate result (Madakam et al. 2022). This data is being connected in a way that can be compared to how the human brain works. In contrast to machine learning, deep learning can learn from its own mistakes (Madakam et al., 2022).
While deep learning and machine learning are widely used techniques for AI, there are many more AI tools or techniques used in different industries. A key problem with many of these AI models is that they are hard to define or explain. AI systems that are hard to explain can be considered a black box (Arrieta et al., 2020).

With all this data, highly intelligent AI and black boxes, it can be difficult for users to understand how these AI processes work.
This poses an intriguing dilemma for businesses and individuals: when AI is not understood, there can be no guarantee that everything is being done according to the rules the developers set for the AI. Not knowing how AI works can pose problems in many different industries. One of these industries is cybersecurity, cybersecurity tools like anti-virus software, network monitoring systems and phishing tools could make use of understandable AI. However, most of

these tools also use a black-box AI model, which means that non-expert users may struggle to understand how these tools operate. The lack of transparency from these tools might create confusion and thus hesitation in following its recommendations. Hesitation or not knowing how AI provides results to tasks can create substantial security issues.
Recently it has become especially important for cybersecurity tools to be working like they should, given that according to Forbes, there were 2,365 reported cyberattacks in 2023, a 72% increase in cybersecurity attacks since 2021 (St John, 2024). These numbers show the importance of efficient and innovative cybersecurity strategies. In order to make the best possible decisions in the field of cybersecurity, an AI tool should be trusted and transparent (Vemuri et al., 2023).

Understandability of AI is a concept that has been around for quite some time and is continuously being developed and researched in the form of explainable AI (XAI). The method of using XAI is aimed at understanding the way AI works, how it gathers its data and how it provides solutions to problems (Arrieta et al., 2020). As an example, there are new deep learning models created called Deep Neural Networks (DNNs). DNNs operate in a black box which means DNNs are not understandable and thus not transparent (Arrieta et al., 2020). For these DNNs, XAI tries to provide an explanation on what data the AI used to provide an acceptable answer for the user. This explanation could be in the form of a graph or an explanation from the AI itself stating the reason this answer was given. With the use of XAI, models gain transparency, explainability, fairness and accountability (Shin, 2021d & Shin, 2022).

1.1 Problem statement

The problem of users or developers not trusting AI is not new, and has recently been getting more attention. Researching explaining AI paved a road for the concept of XAI and thus this concept got more attention from researchers. The method of using XAI is aimed at understanding the way AI works, how it gathers its data and how it provides solutions to problems (Arrieta et al., 2020). In general, humans are not eager to be using software which they cannot explain and this is no different with cybersecurity software (Ahmed et al., 2022).

In industries like cybersecurity, AI use and especially the understanding of AI is also receiving a lot of attention (Gade et al., 2019). Being able to understand AI integration in tools might be a possibility for expert users but non experts users will not be able to understand a cybersecurity tool that is making use of AI. XAI aims to change this and implement more transparency and

more trust for users. An increase of transparency and trust is favourable and can benefit cybersecurity by generating more users or making users more aware of the danger they could be in. Combining XAI with cybersecurity tools could be the difference that ensures non-expert users can understand the complex nature of AI models that are being used. XAI could be the solution of understandable cybersecurity tools for non-expert users giving users the ability to trust and act more effectively on cybersecurity threats.

There are only a few studies about cybersecurity tools using XAI, which exposes a significant research gap worthy of further research. There has, however, been a noticeable increase in the amount of research on XAI, AI for cybersecurity and XAI for cybersecurity, which provides a good theoretical foundation about the key subject of this study. Furthermore, the benefits of researching XAI in cybersecurity tools could prove valuable not only from the perspective of the scientific community but also from a business perspective. The businesses behind cybersecurity tools can benefit from increased explainability of their AI methods to identify errors and make sure the AI is working as it was intended (Arrieta et al., 2020). XAI might also attract more customers since the cybersecurity tools are easier to understand. Furthermore, this study will also show what makes users more receptive to using cybersecurity tools. When a user is more accepting and trusting of cybersecurity tools it might prove that users are quicker to react to, for example, a virus warning thus making a user safer (Zhang et al., 2022).
However, one of the key difficulties with XAI is to retain a high performance of AI systems while still increasing the explainability (Das, A., & Rad, P., 2020). For example, a more simplistic AI model is generally better explainable than a complex AI system. So when trying to implement XAI in a complex AI system the choice is either to compromise on the explanation or make the AI more simplistic. Cybersecurity tools need to be the best to ensure online security, meaning that effectiveness cannot be compromised to increased explainability (Dosilovic et al., 2018).

1.2 Research question

To investigate the trust in XAI the main research question (RQ) is: What is the moderating effect of explainable AI on the relationship between trust and the technology acceptance of cybersecurity tools?
Based on the research question the following sub questions are created:

1. What factors influence technology acceptance according to previous research?
2. Does XAI influence factors other than the relationship between trust and technology acceptance?

1.3 Research method

For the current study a literature review as well as a survey is conducted. The literature review shows previous research of key concepts like AI, XAI, technology acceptance and trust. This literature review includes information like the history of XAI, different AI and AI techniques, multiple technology acceptance models, trust and definitions. This literature review will be used to develop the research model and hypotheses.
As primary research method a survey was conducted to test the hypothesis developed using the literature review. For this study a survey was chosen to cost effectively collect data from a large population (Jack, B., & Clarke, A. M., 1998). The survey was adapted from existing questionnaires, and measures trust and technology acceptance with question making use of Likert scale style answers (Davis, 1989; Jian, et al. 2000). The survey is split into two groups one of these groups is presented with a scenario containing a cybersecurity tool with XAI and the other group is provided with the same scenario except the XAI integration. A pilot survey was conducted to ensure validity and reliability. In total, 52 useful responses were recorded from a random group of respondents. The data gathered with the survey was analysed using a regression analysis and t-tests.

1.4 Theoretical and social relevance

The concept of XAI has only recently been gaining attention in the scientific community, which means that there is less available research on specific subjects like XAI in cybersecurity (Adadi & Berrada, 2018). Further research of XAI can provide an insight in the future possibilities regarding regulations, design and usage of AI technology in many different industries. The combination of XAI and cybersecurity has been explored in previous research however, this research is mostly aimed at cybersecurity experts not the general user.

This study is theoretically relevant since there is yet to be, to the best of my knowledge, a study about the influence of XAI on the trust and acceptance in cybersecurity tools used by the general user. Insights gathered from this study can provide contributions to the design and use of XAI in AI oriented software tools. This research aims to fill the gap between literature aimed at cybersecurity experts and cybersecurity tool users as well as the gap of XAI usage in AI supported tools and standalone AI models. Furthermore, research gaps were identified in the difference between explaining and understanding the AI. This research adds to the knowledge of how users understand AI when explanations are provided.

Practical applications for this study range from policy making to understanding a customers' needs.

Recent interest from government institutions regarding the European general data protection regulations (GDPR) and the benefits XAI can provide these regulations shows further relevance of this research (Das, A., & Rad, P.,2020). Further practical applications of this research can be found in commercial perspectives. This study aims to find an influence in trust and acceptance, both of these concepts are highly sought after in the commercial context as these are seen as the catalyst of buyer-seller transactions (Pavlou, P. A., 2003). Any influences regarding this catalyst can be used to provide better marketing or develop better features in cybersecurity software products. Further details on social and practical relevance can be found in section 6.1 implications.

1.5 Structure

This study starts with a literature review which can be found in chapter 2, this literature shows relevant and previous studies available relevant to XAI, cybersecurity, trust and technology acceptance. This chapter is followed by chapter 3 where the details of the hypotheses and conceptual model are explained. Chapter 4 provides an overview of the research methodology, this includes the development of the survey, selection of respondents, procedures and measures. Next is the results of the study in chapter 5, this section describes the validity and reliability as well as the analysis of gathered data within the study. Chapter 6 provides a discussion about the findings, covering key findings, implications and future research. Finally chapter 7 is composed of a comprehensive conclusion finalizing the thesis.

2 Literature review

The key concepts of this study are cybersecurity, explainable artificial intelligence, trust, and technology acceptance which have been the focus of extensive research. Examining this body of work is essential to ensure that the current study not only builds upon existing knowledge but also makes a meaningful contribution to the scientific community.

2.1 Cybersecurity

Definition of Cybersecurity

Cybersecurity is a broad term that encompasses hardware, software, and information security. In the early days, the focus was mostly on restricting physical access to large fixed installations. However, there has been a shift with the increase of personal computers, and when IT system became increasingly more connected (Shah et al., 2023). The focus of cybersecurity in the later years was on information security rather than physical security. In the last few years, cybersecurity has become increasingly complex due to the interconnected devices, systems and networks. Together with the increase in digital economy infrastructure, this leads to a significant increase in cyberattacks with serious consequences (Kaur et al., 2023). Cybersecurity's main focus is on preventing cyber-attacks like malware attacks, ransomware, denial of service (DOS), phishing or social engineering, and many more (Sarker et al., 2021). These types of attacks can affect organizations and individuals, cause disruptions, as well as devastating financial loss.

There are multiple different defensive strategies that are typically used in the protection of associated data, hardware or software in cybersecurity. Some of these defensive strategies are as follows: Access control, firewall, anti-malware, sandbox, security information and event management, cryptography (Qi et al., 2018; Yin, 2016; Hunt et al., 2017; Irfan et al., 2016; Abood & Guirguis, 2018). Lately, the need for cybersecurity has been changing and the existing counteractivities like antivirus or firewalls may not be effective (Mohammadi et al., 2019). One of the main issues is that the traditional systems are usually operated by just a few security experts, data processing is carried out ad-hoc and can therefore not be run according to the needs (Foroughi & Luksch, 2018). This is where AI comes in, by focussing on AI driven cybersecurity the system becomes intelligent and can answer more specific needs.

2.2 AI

Definition of AI

AI is seeing an ever increasing collection of research being dedicated to it. AI has, especially the last few years, been labelled as the fourth industrial revolution which explains the increase in the amount of research (Zhai et al., 2021). Defining AI is an accomplishment many have tried but none have successfully accomplished as there is no widely accepted definition for AI. This is however not seen as a problem, as many researchers agree that a clear definition can only be accepted once enough research has been generated. Even for this study a definition is not strictly necessary as the AI model itself is not being researched, it is however better for this literature review to see some varying opinions on this matter. Thus some existing definitions of AI will be provided and evaluated.

Starting with one of the earlier definitions by Simmons and Steven (1988) who define AI as "The behaviour of a machine which, if a human behaves the same way, is considered intelligent.". With this definition the comparison between human and machine intelligence is made. This is a fairly common comparison which the following definition from Rada (1968) also encompasses: "Artificial Intelligence is the study of making computers do things which, at the moment, humans do better.". These definitions still hold value as it is the most basic way to describe AI however, more recent definitions might provide more depth. One of the more recent definitions is from Sheikh (2023) and states AI stands for the "imitation by computers of the intelligence inherent in humans.". All these different study's state that AI is hard to define since we ourselves do not fully understand it yet. Thus choosing a narrow definition would be unwise as this study aims to see AI as a whole and not a single technology or concept. In the current study, the following definition is chosen as it encompasses the most applications of AI, this definition is from the High-Level Expert Group on Artificial Intelligence of the European Commission and is as follows "Systems that display intelligent behaviour by analysing their environment and taking actions, with some degree of autonomy, to achieve specific goals"

Understanding AI

After defining AI, it is also essential to understand how it functions and how it is applied. AI is a term that is used for many different models and techniques. AI is build-up of 5 different branches: analytical AI, functional AI, interactive AI, textual AI and visual AI. Each of these different branches uses different models and techniques to accomplish their goal. Most of these techniques use a either machine learning (ML) or deep learning (DL) (figure 1) but there are more like advanced analytics, searching, reasoning or knowledge discovery (Sarker 2022).

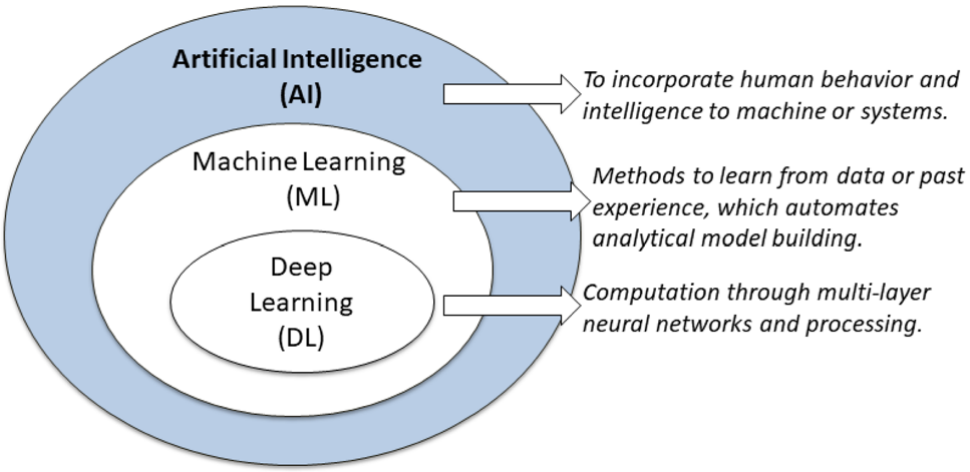


Figure 1 Position of Deep learning and Machine learning within the area of AI (Sarker 2022).

ML is known as one of the most promising AI technologies, it is usually comprised of a set of rules, procedures or sophisticated "transfer functions". These can be used to discover patterns or anticipate behaviour of users (Dua, 2016). ML makes predictions by studying data provided to it, like training data, doing this can solve many real life problems one of these are business risk predictions.

ML is comprised of different learning types, most importantly supervised learning and unsupervised learning. Supervised learning is a task driven strategy that uses labelled data to train models, an example of this is detecting spam emails. The two most common supervised learning tasks are classification and regression (Sarker 2022). Unsupervised learning is a more data driven approach and mostly focuses on pattern recognition from unlabelled data. There are many different applications such as clustering, visualization, dimensionality reduction, anomaly detections and finding associating rules (Sarker 2021). A complete summary of machine learning techniques can be found in figure 2.

Learning type	Model building	Tasks
Supervised	Algorithms or models learn from labeled data (Task-Driven Approach)	Classification, Regression
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, Associations, Dimensionality Reduction
Semi-supervised	Models are built using combined data (Labeled + Unlabeled)	Classification, Clustering
Reinforcement	Models are based on reward or penalty (Environment-Driven Approach)	Classification, Control

Figure 2 Various types of machine learning techniques (Sarker 2022).

Moreover, DL uses a structure similar to the human brain and is based on data representation and learning processes of ML. DL models are also known as DNN models, and combine low-level features to form high-level categories or features. The model has the ability to learn basic

characteristics from small data sets (Bengio, 2009; Goodfellow et al., 2016). DL has three major categories which are: deep networks for supervised or discriminative learning, deep networks for unsupervised or generative learning, deep networks for hybrid learning (Sarker, 2022).

AI techniques

Supervised or classification tasks are a type of deep learning that is employed to provide a discriminative function. These architectures are specifically designed to distinguish between patterns by modelling the posterior distributions of classes based on observable data (Deng, 2014). Examples of deep discriminative models include multi-layer perceptron's (MLPs) (Pedregosa et al., 2011), convolutional neural networks (CNNs or ConvNets) (LeCun et al., 1998), recurrent neural networks (RNNs) (Dupond, 2019; Mandic, 2001), and their various adaptations.

Unsupervised or generative deep learning approaches identify high-order correlations or features for pattern analysis and synthesis, without relying on supervisory information like class labels. These techniques are mainly used for unsupervised learning, feature learning, and data generation (Da'u & Salim, 2020; Deng, 2014). Generative models can also act as preprocessing steps for supervised tasks, enhancing the accuracy of discriminative models. Examples include Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Autoencoders (AEs) (Goodfellow et al., 2016), Restricted Boltzmann Machines (RBMs) (Marlin et al., 2010), Self-Organizing Maps (SOMs) (Kohonen, 1990), and Deep Belief Networks (DBNs) (Hinton, 2009), among others.

Hybrid networks combine both generative and discriminative models. Generative models can learn from both labelled and unlabelled data, while discriminative models excel in supervised tasks but require labelled data. These models may consist of two or more learning models, such as a generative model followed by a discriminative model or a generative/discriminative model integrated with a non-deep learning classifier, offering effective solutions to real-world problems (Sarker, 2022).

In conclusion, there are many different DL techniques. Some well-known techniques are shown in figure 3.

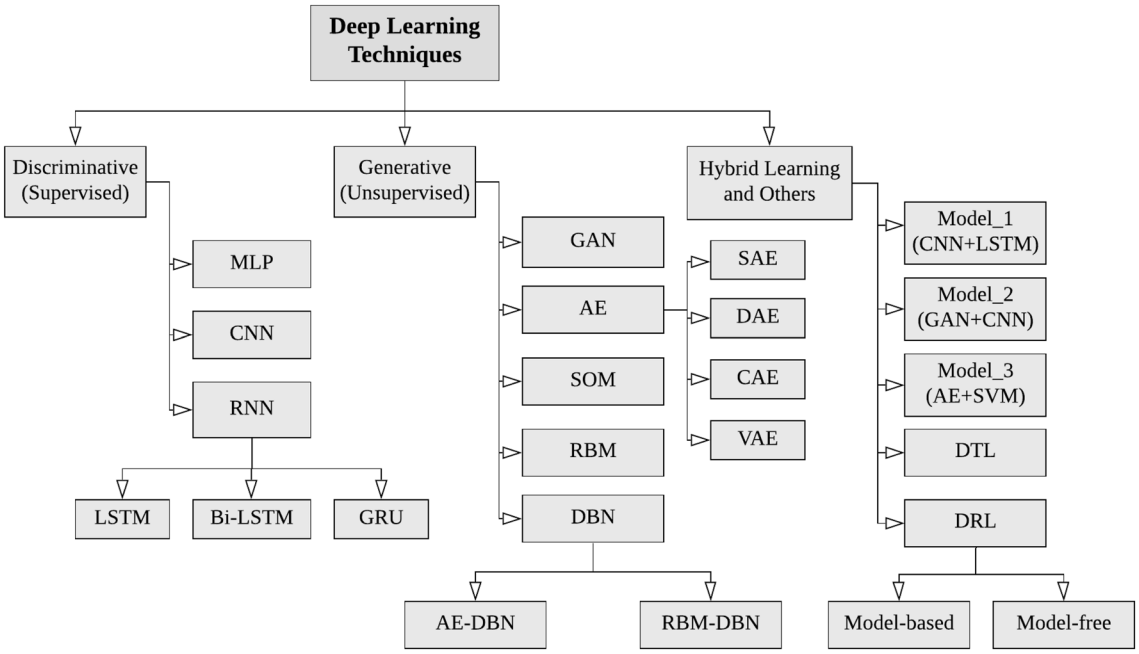


Figure 3 Deep learning techniques separated in three categories (Sarker, 2022).

While there are many more different AI techniques, ML and DL are techniques that are used the most frequently at this moment. A summarization of other popular AI techniques and their respective application areas and tasks can be found in figure 4 (Sarker 2022).

AI techniques	Application areas	Tasks
Machine learning	Healthcare	COVID-19 aid
	Cybersecurity	Anomaly and Attack Detection
	Smartcity	Smart parking pricing system
	Recommendation systems	Hotel recommendation
Neural network and deep learnig	Healthcare	Diagnosis of COVID-19
	Cybersecurity	Malware detection
	Smart cities	Smart parking system
	Smart Agriculture	Plant disease detection
	Business and Finance	Stock trend prediction
	Virtual Assistant	An intelligent chatbot
	Visual Recognition	Facial expression analysis
Data mining, knowledge discovery and advanced analytics	Education	Decision support systems
	Business	Maximising competitive advantage
	Cybersecurity	Human-centred data mining
	Diagnostic analytics	To mature gas fields
	Prescriptive analytics	Optimizing outpatient appointment
Rule-based modeling and decision-making	Intelligent systems	Mining contextual rules
	Healthcare	Identifying risk factors
	Recommendation system	Web page recommendation
	Smart systems	Risk prediction
Fuzzy logic-based approach	Healthcare	Heart disease diagnosis
	Agriculture	Smart irrigation
	Cybersecurity	Network anomaly detection system
	Business	Customer satisfaction
Knowledge representation, Uncertainty reasoning and Expert system modeling	Smart systems	Smart traffic monitoring
	cloud computing	Ontology data access control
	cybersecurity	Vulnerability management
	Mobile expert system	Personalized decision-making
Case-based reasoning	Healthcare	Breast cancer management
	Smart cities	Energy management
	Smart Industry	Fault detection system
	Recommendation Systems	Classification and regression tasks
Text mining and natural language processing	Sentiment analysis	Sentiment analysis of tweets
	Business	Product reviews sentiment
	Cybersecurity	Estimating security of events
	Healthcare	Effectiveness of social media
Visual analytics, computer vision and pattern recognition	Healthcare	Cervical cancer diagnostics
	Computer vision	Human fall detection
	Visual Analytics	Navigation mark classification
Hybrid approach, searching and optimization	Mobile application	Personalized decision-making
	Recommendation systems	Personalized hotel recommendation
	Sentiment analysis	Tweet sentiment accuracy analysis
	Business	Customer satisfaction
	Cybersecurity	Optimum feature selection

Figure 4 AI tasks and methods in several popular real-world applications areas (Sarker 2022).

Cybersecurity and AI

AI use in cybersecurity can be a key technology for solving cybersecurity issues that are too complex for the traditional methods, or could be done better with AI (Sarker et al., 2021). Some of the most used AI applications in cybersecurity are threat intelligence, phishing detection and incident response (Sarker et al., 2021). With threat intelligence the AI, usually a natural language processing (NLP), analyses an large amount of data to provide relevant threat indicators or contextual information. Phishing detection uses NLP techniques that are employed to examine the language and structure of emails to identify potential phishing attempts. And lastly, with incident response, NLP can assist cybersecurity teams with understanding and categorizing incident rapports (Balantrapu, 2024). Phishing is the one threat that is directly aimed at human interaction as it is a form of social engineering (Salloum et al., 2022). As this study aims to understand the influence of XAI on users' trust, human interaction is necessary.

Therefore, this study focusses on phishing detection. Over the years that AI has been applied to cybersecurity problems it has had some interesting impacts, both positive and negative.

Some of the positive impacts of AI on cybersecurity have been for businesses and companies. As shown by Perols and Murthy (2018), attacks on companies have been becoming more dangerous since attackers are increasing their knowledge and finding more weak spots in cybersecurity defences. The intelligence of AI makes sure that attackers can never use the same attack twice, and AI also ensures that (human) errors are avoided. Machine learning algorithms can process huge amounts of data flagging anomalies and enabling proactive defensive strategies before an attack can be executed (Rangaraju, 2023; Shah, 2021; Kasowaki & Emir 2023).
To ensure that the response to a cybersecurity threat is as efficient as possible, AI has taken on different roles within cybersecurity. These AI systems are continuing their own research to deal with every possible attack. The expectation is that the scale on which AI is incorporated in cybersecurity will continue to grow. The future for AI in cybersecurity is bright as companies are researching how AI systems can not only protect the companies systems but also themselves (Hamilton, 2020; Sharma et al., 2022).

However, applying AI in cybersecurity can also provide challenges. Ethical implications of AI systems in cybersecurity raise questions about accountability, transparency and unintended consequences. Automated decision making may lead to biases or errors that are blindly accepted as there is no insight in how AI generates these decisions. This leads to the importance of ethical AI development, government frameworks and transparent AI (Marda, 2018; Sontan & Samuel 2024). To start combatting AI challenges, building an understanding is step one. Without a clear sight of what AI is doing, government frameworks or accountability cannot be developed. Transparency is the key in this regard, as a transparent AI model can further be developed to be ethical and effective. In conclusion, transparency is critical to building trust and confidence in AI use in cybersecurity (Familoni, 2024).

2.3 Explainable AI

The concept of XAI has gained significant attention as AI models, particularly deep learning models, have become more opaque and are often considered "black box" systems (Castelvecchi, 2016). This growing focus is partly driven by the reluctance to adopt systems that are not interpretable, traceable, or trustworthy (Zhu et al., 2018). As a result, there is a rising demand

for ethical AI that includes these qualities (Goodman & Flaxman, 2017). Before looking into the theories and details of XAI, it is important to first define XAI.

Definition XAI

The question of how to define XAI is not easy, it is a collection of terms that all try to get to a similar result. However, there is a notable difference in the terminology of XAI as there is a common misuse of the concepts interpretability and explainability. Interpretability refers to the level at which a given model makes sense for an observer, similar to transparency (Guidotti et al., 2018). Explainability is an action of an AI model to clarify its internal function (Guidotti et al., 2018). In other words, a model is transparent if it is understandable by itself (Lipton, 2018). As stated by Arrieta et al. (2020), understandability is the most important concept for XAI. This concept is defined as "the characteristic of a model to make a human understand its function, how the model works, without any need for explaining its internal structure or the algorithmic means by which the model processes data internally." (Montavon et al., 2017).

Given these definitions of the key concepts in XAI being understandability, transparency and explainability, a definition of XAI could provide some further insight. One of these definitions is from Gunning et al. (2021), who defines XAI as "a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificial intelligence partners". This definition combines two concepts, being understanding and trust. However, this definition is far from complete as causality, transferability, informativeness, fairness and confidence are not mentioned (Lipton, 2018; Doran et al., 2017; Doshi-Velez & Kim, 2017; Vellido et al., 2012). Another definition of XAI comes from the Cambridge Dictionary's definition for explanation which is "the details or reasons that someone gives to make something clear or easy to understand" (Walter, 2005). This definition can be rewritten in the context of this study to "the details or reasons a model gives to make its functioning clear or easy to understand". The issue with this definition is the lack of a receiver of these details, there have to be users or an audience that finds it easy to understand. Lastly, there is the definition from Arrieta et al. (2020) that is more focused on XAI and explainability. This definition is as follows: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.". This definition seems to be encompassing the importance of the model explaining to an audience, and assumes XAI provide understandability and clarity which leads to better trustworthiness for the audience (Arrieta et al., 2020). For this study it is important to incorporate trustworthiness and XAI in the definition. When looking at the definition from Gunning (2021), all the necessary components as not only the user, but also trust are both mentioned. However, since the definition from Arrieta et al. (2020) provides the most clear and

easy to understand definition that incorporates users/audience and trust, this definition of XAI will be used for the study.

Explainability

There are many different goals that XAI aims to achieve such as trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity and privacy awareness. For this study, the focus will be on is trustworthiness, as trustworthiness is a goal that several authors use to describe XAI and its primary aim (Kim et al. 2015; Ribeiro et al., 2016). A model that is explainable does not equal a model that is trustworthy, trustworthiness is the confidence that a model will act as intended (Arrieta et al., 2020).

Providing explainability with XAI can be done in several different ways that can all provide the desired outcome of an understandable AI model. XAI explains its predictions that are derived from the training data collected by a company. There are two options for providing explanations, which are global interpretation or local interpretation (Mirhoseini et al., 2017). Global interpretation methods focus on analysing a model's overall behaviour. This involves defining variables, understanding their dependencies and interactions, and assigning importance to these components (Dwivedi et al., 2023).
Local interpretation involves analysing individual predictions and decisions made by the model to understand why it suggested a particular course of action. The focus is on the area surrounding the specific data point being analysed (Dwivedi et al., 2023).

These interpretation methods can be seen as categories for ways to explain a model. When using XAI, one of the most used ways to explain are visualisations. These visualisations are being generated using different techniques such as partial dependence plot (PDP), LIME and Shapley additive explanations (SHAP). All of these techniques are called feature-based techniques and they describe how input features contribute to the models output (Dwivedi et al., 2023).
PDP shows what features influence the prediction in a line graph -1 to 1 where 0 means there is no influence, -1 is a very negative influence and 1 is a very positive influence (Friedman, 2001; Brandon, 2017). LIME tries to understand the AI model by comprehending how the predictions change (Ribeiro et al., 2018), and SHAP is a game-theoretic approach to explaining output, interpretations measures the impact of having a certain value for a given feature in comparison to the prediction (Lundberg & Lee, 2017). Though there are many more techniques to implement XAI in an AI model, these three are some of the most used techniques.

In some available research the question is proposed if creating more transparency will impact the performance of AI models. This is mainly due to the fact that ‘complex’ models have more

layers, and therefore can be seen as a bigger model. Creating models that are less complex would mean decreasing the size and levels of the model which would also decrease the accuracy (Freitas, 2004). Jin and Sendhoff (2008) find a similar interaction and describe interpretability as mainly being influenced by complexity for neural networks.

XAI and trust

This study aims to investigate the relationship between XAI and trust. Existing literature on this relationship is minimal but there are some relevant previous studies. Lai and Tan (2019) have researched the impact of XAI in the form of heatmaps of relevant instances and example-based explanations, and asked users if a fictional hotel review was genuine or deceptive. This study showed that the XAI method that was used increased the level of measured trust. Trust was measured as the percentage of instances where participants relied on machine predictions. However, Cheng et al. (2019) presented a study that explains the inner workings of AI via the use of different UI interfaces. In this study, explaining AI did not affect the trust of the user which was measured with a survey scale. There are many different ways to measure and experiment with XAI and trust and these different ways show different results. Scharowski et al. (2022) describes that a possible reason for inconsistency in XAI research could be due to the differences in measurements.

2.4 Trust

Definition

Another key concept is trust which has been studied by multiple sources such as Binns et al. (2018) stating that “the provision of explanations can affect levels of trust and acceptance of algorithmic decisions” or Miller (2018) describing the relationship between explainability and the trust in AI as “to increase” or “to generate”. Trust is a multifaceted and thus complex concept, which makes trust hard to define (Chui, 2022). Philosophical analyses qualify trust as the decision to delegate a task, without any form of control or supervision over the way the task is executed (Taddeo, 2010). In the terms of cybersecurity, a user trusts that their system is kept free of any malware. Trust is considered a reputation of the tool that is being trusted (Henshel et al., 2015). In recent publications studying trust in the cyber domain, trust has been specified to “digital trust”. Digital trust encompasses the belief that systems, platforms, technologies or cybersecurity tools operate according to the intended use. Digital trust requires the assurance of privacy and safeguarding of personal data (Chui, 2022).

Trust and AI

When looking at trust in conjunction with AI, transparency is considered critical for developing trust as this refers to the inner logics and operating rules that are apparent for the user. Multiple other sources state that transparency and explanation increases trust for virtual AI (Fan et al., 2008; Wang & Benbasat, 2007; Wang et al., 2016). Researchers have identified different trust trajectories based on the several starting characteristics. An interesting theory for this study states that trust generally starts high with virtual AI and decreases following interaction between a user and the AI (Hoff & Bashir, 2015). Ben Mimoun et al. (2012) found a similar reaction when users are presented with a commercial website using virtual agents, over the years the actual use of these virtual agents significantly decreased. It is however possible to reverse this effect, in some cases it was found that the trajectory started out low and increase with time. Kopp et al. (2005) found this reversed trajectory when studying the use of a virtual agent in a museum. To explain this difference in trust trajectories, several features are mentioned where anthropomorphism can significantly increase user expectations and the AI machine intelligence moderated the direction of the trust trajectory. High machine intelligence provides a positive trust trajectory, and low machine intelligence underperforms when users have high expectations which creates a negative trust trajectory (Glikson & Woolley, 2020).

In recent literature the concepts that influence AI to be more trustworthy have been the main focus, an example of one of these studies is from Shin and Park (2019) where the roles of fairness, accountability and transparency (FAT) in predicting users' satisfaction was studied. In later studies, explainability was added creating the FATE framework and shows the determinants of trust in AI (Shin, 2021d, 2022). The overall findings show that FATE plays a pivotal role in the increase of user trust. For this study the findings show that explainability should provide a significant influence on trust.

In general, the literature shows that trust plays a critical role in the perception and acceptance of AI technologies in many different environments such as healthcare (Lee & Rich, 2021), hiring and work applications (Lee, 2018) and algorithmic journalism (Shin, 2020c).

Technology acceptance (TAM)

In this study, technology acceptance will be conceptualised in the form of the TAM model. Technology acceptance has always had a lot of interest in information system studies, and continues to be widely used by researchers as a concept in different studies. One of the most used models is the technology acceptance model (TAM) which emphasises an individual's perceptions of technology driving their acceptance and use behaviour (Carr et al., 2010; Marangunić & Granić, 2015). In the case of the TAM model, this is 'perceived ease of use' and 'perceived usefullnes', which are theorized to be the foundations of technology acceptance

(Davis, 1989). While many studies use the TAM model (Fallatah et al., 2024; Shrestha et al., 2021; Heierhoff & Choun, 2023), differentiating models like the unified theory of acceptance and use of technology are also being used (Tam et al., 2023; Robles-Gomez et al., 2021). Where some studies integrate trust into the TAM model (Dickson et al., 2021), others show that trust can increase the technology acceptance of a user (Lee & See, 2004). For this study, trust is seen as a separate concept so that the influence of trust on technology acceptance can be measured and the moderating influence of XAI on the relationship between trust and technology acceptance can be measured. This choice has been made, as it has been shown in previous study that trust has an positive influence on the key TAM constructs. Wu et al. (2011) and Mou et al. (2016) show that trust in AI predicts the perceived usefulness and ease of use of AI. This study uses the TAM model (Figure 5) to measure technology acceptance and aims to show the relationship trust has on its key concepts perceived usefulness and perceived ease of use.

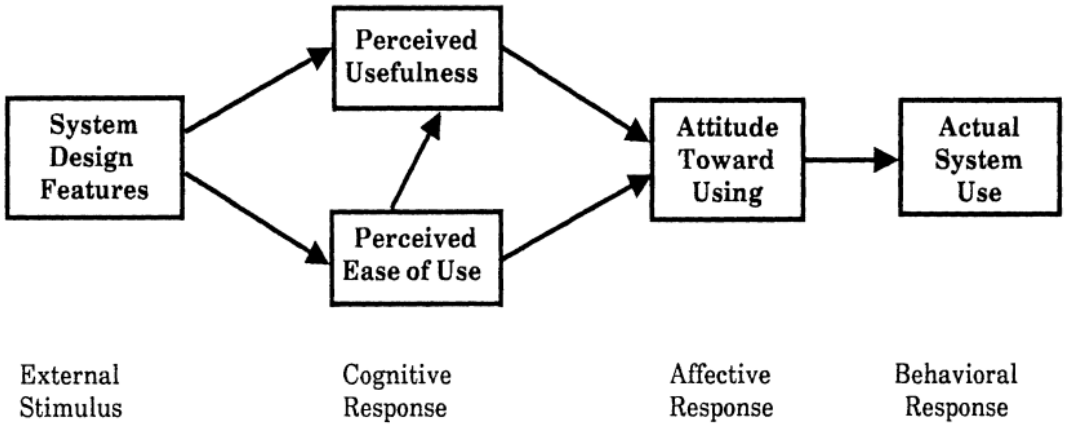


Figure 5 TAM model (Davis, 1989).

A study from Baroni et al. (2022) has integrated trust in AI within the TAM model and is composed of two more concepts than the original TAM model which are: Trust in AI and Quality of AI. This AI-TAM model shows a strong positive effect on behavioural intention, perceived usefulness and ease of use.

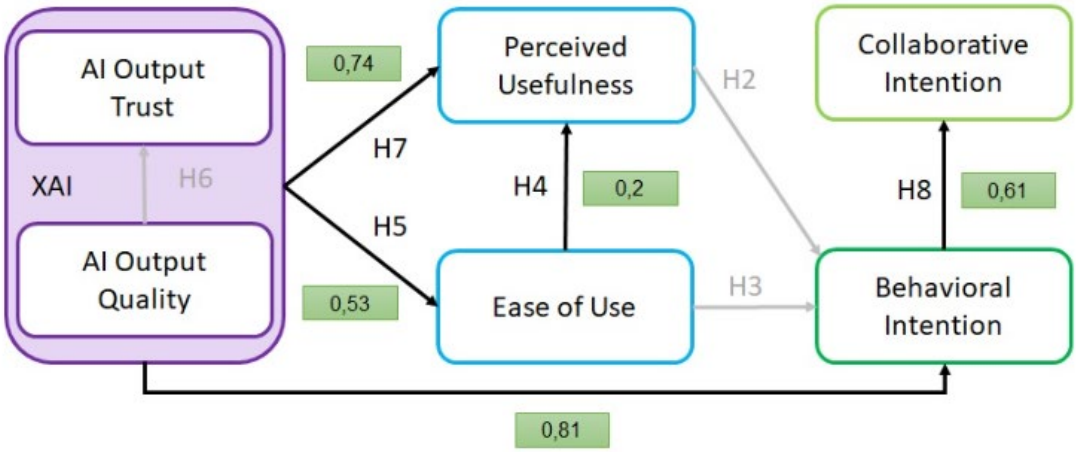


Figure 6 AI-TAM model integration (Baroni et al., 2022).

As was shown in the theories that have been discussed, trust is a concept that many have considered combining with the TAM model. How to combine these concepts into one model is a question that has a lot of different research directions. One important view comes from Gefen et al. (2003), who state that trust positively influences the intend to use (technology acceptance) and that trust positively influences the perceived usefulness. However, trust has been theorized be influenced positively by perceived ease of use (Ganesan, 1994).

Table 1 Summary of important literature

Author(s)	**Literature Category**	**Research Focus**	**Results**
Sarker et al. (2021)	Cybersecurity	Role of AI in addressing complex cybersecurity issues.	AI applications like threat intelligence, phishing detection, and incident response improve security.
Familoni (2024)	Cybersecurity	Navigation of cybersecurity challenges	By embracing a holistic view and using the latest technology advancements organizations can stay ahead of the threats.
Goodman and Flaxman (2017)	Explainable AI (XAI)	Introduction of explanation to the GDPR	By properly applying XAI methods algorithms can not only make more accurate predictions, but offer increased transparency and fairness.
Arrieta et al. (2020)	Explainable AI (XAI)	Challenges of black-box AI models and defining XAI.	Transparency and interpretability are critical for building trust; XAI provides clarity through methods like PDP, LIME, and SHAP.
Lai and Tan (2019)	XAI and Trust	Influence of XAI on user trust.	XAI explanations increased user reliance on AI predictions in specific contexts.
Cheng et al. (2019)	XAI and Trust	Explaining AI via different UI interfaces and its impact on trust.	No significant effect of explanations on trust.
Scharowski et al. (2022)	XAI and Trust	evaluating existing studies with the concepts trust and XAI.	Results of XAI studies fluctuates by context and methodology.

Author(s)	**Literature Category**	**Research Focus**	**Results**
Shin and Park (2019)	Trust in AI	FAT (Fairness, Accountability, Transparency) framework and its role in trust.	Transparency and explainability are pivotal for increasing trust in AI.
Jin & Sendhoff (2008)	XAI and Performance	Trade-off between explainability and AI performance.	Simplifying AI models to increase explainability often reduces effectiveness.
Baroni et al. (2022)	Technology Acceptance (AI-TAM)	Integrating AI and trust into the TAM model.	Trust and quality of AI positively influence behavioural intention and technology acceptance.

3 Hypotheses

Based on the research question: What is the moderating effect of explainable AI on the relationship between trust and the technology acceptance of cybersecurity tools? Hypotheses are developed below.

Trust in AI-driven decision-making has been the focus of several studies, with research highlighting the role of interpretability in fostering trust. Ashoori and Weisz (2019) argue that trust in AI is closely linked to how well users can interpret its decisions, whereas black-box models tend to reduce trust due to their lack of transparency. This distinction makes the direct relationship between XAI and trust particularly relevant to this study, separate from XAI's potential moderating effect on technology acceptance.
Baroni et al. (2022) identified XAI as a distinct construct, though their research combined it with trust rather than examining its standalone impact. This study builds on that work by investigating XAI as an independent factor influencing trust. By analysing the relationship between XAI and trust separately from technology acceptance, this research aims to clarify whether XAI serves primarily as a moderator or whether it has a direct effect on trust. Based on this, the following hypothesis is proposed:
H1: XAI has a positive influence on trust in AI.

The relationship between trust and technology acceptance has been widely explored in previous research, yet its complexity makes it difficult to generalize across different settings. While trust is often seen as a key factor in the adoption of AI-driven systems, its precise influence can vary depending on context. Dahlberg et al. (2003) states that a person's attitude, such as trust, affects both perceived ease of use and perceived usefulness, ultimately influencing technology acceptance. Similarly, Lee and See (2004) found that greater trust leads to higher technology acceptance, reinforcing the idea that trust plays a crucial role in user intention to use.
Previous studies also highlight the direct relationship between trust and perceived usefulness. Gefen et al. (2003) argue that users are more likely to perceive a system as useful when they trust its reliability and decision-making processes. Additionally, Ganesan (1994) suggests that perceived ease of use can positively influence trust, as users tend to trust systems that are intuitive and effortless to navigate.
Building on these insights, this study applies the TAM framework (Davis, 1989) alongside trust as an additional factor, as demonstrated in the AI-TAM model by Baroni et al. (2022). Based on this theoretical foundation, the following hypotheses are proposed:

H2: Trust in AI has a positive influence on technology acceptance of AI.

H3: Trust in AI has a positive influence on perceived usefulness of AI.
H4: Perceived ease of use of AI has a positive influence on trust in AI.

XAI aims to enhance transparency, thereby increasing trust in AI models (Kim et al., 2015; Ribeiro et al., 2016). However, research by Scharowski et al. (2022) indicates that the effect of XAI on trust varies across studies, likely due to differences in measurement approaches. While trust in XAI has been explored in prior research, this study seeks to expand the existing knowledge by examining the impact of XAI on trust using alternative methods and different experimental settings (Cheng et al., 2019; Lai & Tan, 2019).
As demonstrated in the AI-TAM model by Baroni et al. (2022), AI trust directly influences perceived usefulness and perceived ease of use, which are key factors in technology acceptance. However, this study investigates whether XAI plays a moderating role in the relationship between trust and technology acceptance rather than simply acting as a direct influence. Additionally, previous research suggests that XAI may positively impact both trust and technology acceptance (Weitz et al., 2019). Based on these considerations, the following hypothesis is proposed:
H5: The relationship between trust in AI and technology acceptance of AI is positively moderated by XAI.

3.1 Conceptual model

A conceptual model was developed to summarize the hypotheses and provide a clear overview of the key concepts and their proposed relationships. As shown in the model, the TAM framework has been largely incorporated, with the exception of system design features, which are not relevant to this study (Davis, 1989). The model begins by examining the influence of trust on both the intention to use (H2) and perceived usefulness (H3). It then explores the impact of XAI on trust (H1) and its moderating effect on the relationship between trust and the intention to use (H5). Additionally, the model includes the influence of perceived ease of use on trust (H4). While perceived usefulness and perceived ease of use have multiple other connections within the model that where not mentioned, these relationships have already been extensively studied by Davis (1989) and will not be the primary focus of this research. Finally the model considers the demographic variables measured during this study which are Experience with AI, Education level and Age.

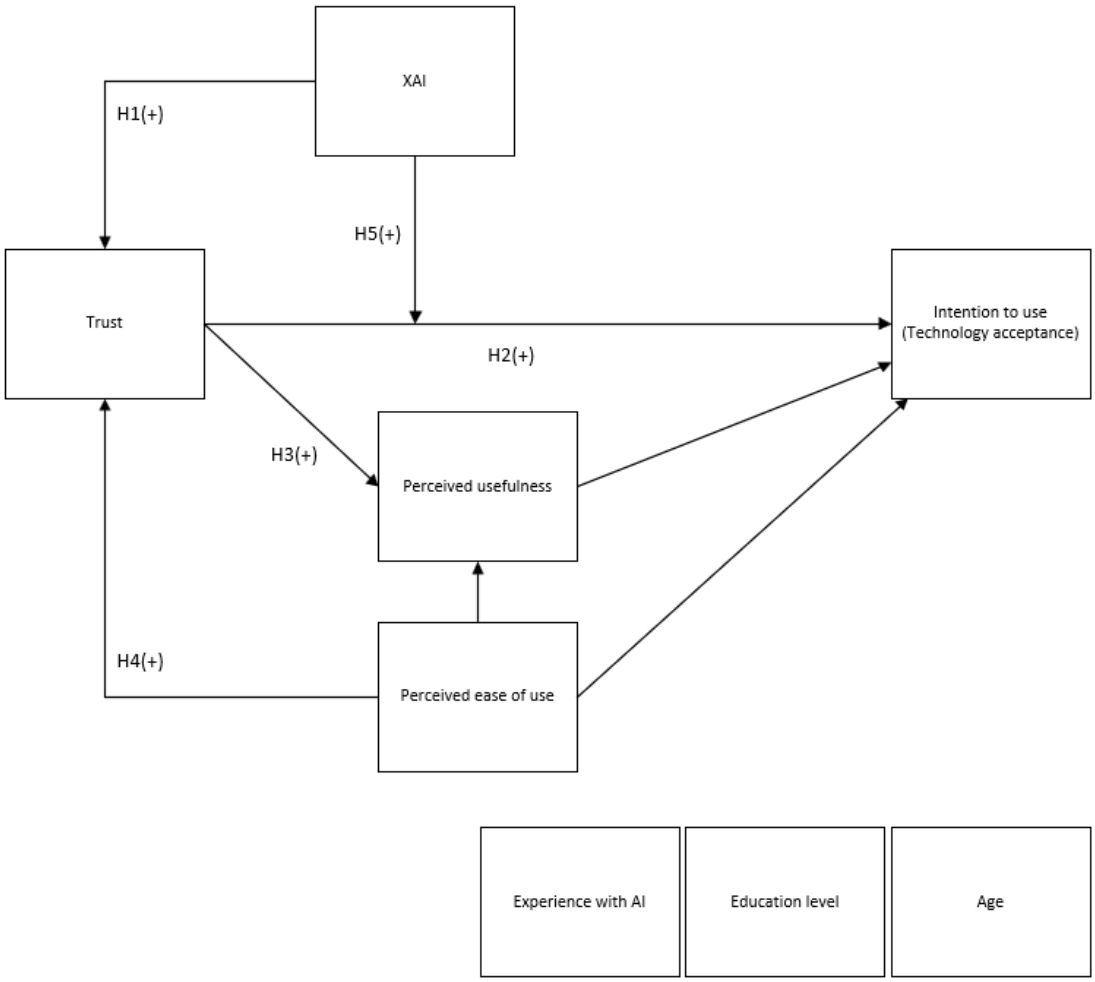


Figure 7 Conceptual model.

4 Method

his study aims to examine how Explainable AI (XAI) influences the relationship between trust and technology acceptance among cybersecurity tool users. Specifically, it explores whether XAI positively impacts trust, enhances the acceptance of cybersecurity tools, and strengthens the link between trust and technology acceptance. Additionally, the study investigates potential relationships between perceived ease of use and trust, as well as between trust and perceived usefulness. These connections are analysed in greater depth to understand their implications.

4.1 Participants

This study benefits from a randomly selected group of participants as this gives a broad view across demographics. To achieve this there was no selection criteria for participants and through different social media platforms such as LinkedIn, Instagram, WhatsApp and Facebook everyone willing was invited to participate in the study. Furthermore, networking was used to further find participants, as every respondent was asked if they were willing to share the questionnaire with their network. At the end of the questionnaire period, there were 131 responses, 2 of these responses failed an attention check question and 77 responses where left unfinished. In total the survey was completed with 52 useful responses for the study. Of these 52 responses 24 respondents completed the questionnaire with XAI and 28 respondents completed the questionnaire without the XAI. The average response time of the survey was 319,48 seconds (5,32 minutes).
The age of the participants was measured to be between 18 and 66 (M=35.08, SD=15.251). 36,5% of participants finished a bachelor's degree, 28,8% a Master’s degree, 17.3% finished secondary vocational education and 13,5% finished high school. Just 11.5% of the respondents had more than 3 years of experience with AI the remaining 88,5% had less than 3 years experience.

4.2 Procedure

For the data collection of this study an online questionnaire was used. An anonymous link to the questionnaire was send out via text message or could be found in an online post on one of the earlier mentioned social media platforms. A participant could access the questionnaire via this anonymous link. The questionnaire would first confirm consent by stating how the data gathered in the questionnaire would be used and stored and if the respondent is above the age of 16. After confirming consent the questionnaire is explained and one of two scenarios are presented to the participant in one of these scenarios a phishing mail is discovered by an AI tool on a work

laptop. When a phishing mail is detected in this scenario a cybersecurity tool that uses AI provides the user of the email with a popup that the email they are currently reading might be a phishing email. The second scenario provides the same situation as the first however the AI tool explains the reasoning it might suspect the mail is a phishing mail. The explanation is based on XAI principles and uses a Partial Dependence Plot (PDP) graph to visualize the explanation. PDP is regarded as one of the better ways to visualize the workings of AI (Aechtner et al., 2022). After the scenario 28 items are presented to the participants. The first twelve items measure trust followed by an attention check then twelve questions measure the technology acceptance through perceived ease of use and perceived usefulness, after these twelve questions three demographic questions are given and thus concluding the questionnaire.

As a pre-test the survey was send to a small group of 7 people to check the survey before publishing. In this pre-test the wording of some demographic question was altered to increase clarity. Furthermore, some grammar mistakes where corrected and the demographic questions where placed at the end of the survey instead of the beginning. And finally, an attention check question was added to increase reliability.

4.3 Measures

The questionnaire that was used for the study is based on different existing scales that give insight in the concepts trust and technology acceptance. For the trust measurement the scale of Jian et al. (2000) was used. For the measurement of technology acceptance a scale form Davis (1989) was used.

Trust

The scale of Jian et al. (2000) has been used in many different studies regarding trust in automation. The scale was developed by using three different experiments, a word elicitation study, a questionnaire study, and a paired comparison study. The scale was developed using a factor and cluster analysis of the data gathered from these experiments. The scale uses 12 items for measuring the trust between people and automation. The statements are scored on a 7-point Likert scale from ‘Not at all’ to, ‘Extremely’. The scale can be found in appendix 1. The only change made to this scale for this study is the replacement of the word ‘System’ with ‘Cybersecurity tool’.

Technology acceptance

Technology acceptance was measured based on the Davis (1989) scale. This scale is developed with two additional constructs perceived ease of use and perceived usefulness. Using previous research and pretest interviews 10 items were developed for each construct. Using a field test with 112 users the scale was further refined to 6 items per construct and finally a lab study with 40 participants was conducted to further ensure validity. The Davis (1989) scale uses a 7 point Likert scale from extremely likely (7) to extremely unlikely (1).

4.4 Data analysis

For the statistical data analysis of this study IBM SPSS (Version 30) was used. To examine the relationship between XAI and trust (H1) an independent T-test was performed to compare the trust in the XAI and NoXAI groups. For H2, H3, H4 and H5 a linear regression was used. The variable for the trust scale was created by computing the means of the results from the 12 items of the Jian, et al. (2000) scale. For the technology acceptance variable the means where computed of both subscales (6 items) of the Davies (1989) scale. Finally, the interaction variable was computed by multiplying the trust scale with group scale of what scenario a respondent was presented with, either a 1 for the scenario with XAI or a 0 for the scenario without XAI.

5 Results

5.1 Descriptives

Out of all the respondents none where younger than 18, 30.8% is between the ages of 18 and 24, 38.5% is 25-35, 1.9% is between 36-45, 7.7% is between 46-55 and 21.2% is 55+. As stated in the participants section the oldest respondent was 66 and the average age of a respondent is 35.08.

Table 2 Descriptives Age & Education

Age	N	Percent (%)	Education	N	Percent (%)
18-24	16	30,8	None	1	1,9
25-35	20	38,5	Secondary/high school education	7	13,5
36-45	1	1,9	Secondary vocational education (MBO)	9	17,3
46-55	4	7,7	Associate's degree	1	1,9
55+	11	21,2	Bachelor's degree (HBO/WO Bachelor)	19	36,5
Total	52	100	Master's degree (WO Master)	15	28,8
			Total	52	100

Of these respondents one did not finish an education, 13.5% finished secondary/high school, 17.3% finished secondary vocational school (MBO), 1.9% finished an associate degree, 36.5% finished an bachelor's degree (HBO/WO bachelor) and 28.8% finished their master's degree (WO master).

Table 3 Descriptives Experience with AI

Experience with AI	N	Percent (%)
None	8	15,4
A little (<1 year)	20	38,5
Some (1 - 2 years)	18	34,6
Quite a bit (3-4 years)	5	9,6
A lot (>4 years)	1	1,9
Total	52	100

The respondents were asked about their experience with AI. 15.4% of the respondents stated that they have no experience with AI or machine learning, 38.5% has a little experience with AI around 1 year, 34.6% has between 1 or 2 years of experience, only 9.6% has 3 to 4 years of experience and just one person has more than 4 years of experience with AI.

Table 4 Descriptives min, max, mean and std deviation

		N	Minimum	Maximum	Mean	Std. Deviation
NoXAI	Trust	28	3,42	6,25	4,91	0,82
	PU	28	1,33	6,5	4,43	1,44
	PEOU	28	2,67	7	5,06	1,07
	TA	28	2,25	6,33	4,75	1,02
XAI	Trust	24	3,33	5,67	4,41	0,78
	PU	24	2,67	6,5	4,53	1,07
	PEOU	24	2,83	6	4,57	0,95
	TA	24	3,33	6,25	4,55	0,91

In table 3 the descriptive statistics provide insights into the differences between the XAI and non-XAI groups in terms of trust, perceived usefulness (PU), perceived ease of use (PEOU), and technology acceptance (TA). Overall, the means suggest that respondents in both conditions reported moderate to high levels of trust and TA. However, differences emerge when comparing the two groups. Trust was higher in the non-XAI group (M = 4.91, SD = 0.82) compared to the XAI group (M = 4.41, SD = 0.78), suggesting that the presence of explainability did not necessarily enhance trust and may have even reduced it. Similarly, PEOU was higher in the non-XAI group (M = 5.06, SD = 1.07) than in the XAI group (M = 4.57, SD = 0.95), indicating that participants who received AI explanations found the tool slightly less easy to use. PU scores were relatively similar across groups, with the non-XAI group scoring slightly lower (M = 4.43, SD = 1.44) than the XAI group (M = 4.53, SD = 1.07), though the standard deviation was higher in the non-XAI condition, reflecting more variation in responses. TA followed a similar pattern, with the non-XAI group scoring slightly higher (M = 4.75, SD = 1.02) than the XAI group (M = 4.55, SD = 0.91). These results suggest that, contrary to expectations, the presence of XAI did not enhance trust or TA and may have introduced complexity that affected PEOU and overall perceptions.

5.2 Validity and reliability

Before moving on to the regression the construct validity of the study was determined. To achieve this a confirmatory factor analysis was conducted for two constructs: trust and technology acceptance which includes perceived ease of use and perceived usefulness. The complete results can be found in appendix 3. All 24 items used in the survey were added to the factor analysis. The Kaiser-Meyer-Olkin (KMO) was above the .5 (.702) and the Bartlett's test was found to be significant ($p<.001$). The factor analysis resulted in 6 factors with an eigenvalue greater than 1. However since there where 3 factors expected (trust, perceived ease of use,

perceived usefulness) the factor analysis was limited to 3 factors. With the 3 factors together 60% of the variance of this study can be explained resulting in discriminant validity for the 2 constructs.

Convergent validity was also measured for this study using a principal component analysis (PCA). The complete results of this analysis can be found in appendix 3.
For this validity a PCA was run on the factors independently of each other. Starting off with the 12 items used to measure trust. The KMO score is .760 which is acceptable as it exceeds .5 and the Bartlett's test was also found to be significant ($p<.001$). The PCA resulted in 2 factors that explain 62% of the variance. While only 1 factor was expected the trust scale is dived in positive and negatively worded questions which could lead to 2 factors. This means convergent validity can be guaranteed as 2 factors explain more than 50% of the variance.
For technology acceptance the same PCA analysis was conducted as for trust. This resulted in a KMO of .844 and the Bartlett's test was found to be significant ($p<.001$). For this variable there are excepted to be 2 factors as technology acceptance is split in two subscales perceived ease of use and perceived usefulness. The result of the PCA analysis supports this expectation there are two factors that explain 68% of the variance. For this reason, convergent validity can be guaranteed for technology acceptance.

Lastly the reliability of the scales were analysed using Cronbach's alpha. The internal consistency of the trust scale is good ($\alpha= .807$). The internal consistency of the technology acceptance is also good ($\alpha= .891$) and when technology acceptance is dived in its subscales the internal consistency of 'Perceived Usefulness' is very good ($\alpha= .907$), the second construct scale 'Perceived Ease of Use' is ($\alpha= .877$).

5.3 Hypotheses results

The first hypothesis examines the influence of XAI on trust (H1). To test this, an independent samples T-test was conducted to compare trust levels between the XAI and NoXAI groups. Levene's test for equality of variances shows a non-significant result ($p = 0.936$), indicating that equal variances could be assumed. The T-test results revealed a statistically significant difference in trust between the two conditions ($t(50) = -2.250$, $p = 0.029$). The mean difference of -0.50 suggests that trust levels were significantly lower in the NoXAI group compared to the XAI group. These findings do not reject the hypothesis, demonstrating that XAI positively influences trust, as participants exposed to explainable AI reported higher trust levels than those who were not.

To provide a complete overview of the concepts from this study the effect of XAI on TA was also tested with an independent samples T-test. Once again Levene's test for equality of variances shows a non-significant result (p= 0.844), meaning that equal variances could be assumed. This T-test results without a significant statistical difference in TA between the two groups (t(50) = -.735, p = 0.466). The mean difference is -0.20 meaning that the TA was slightly lower in the NoXAI group in comparison to the XAI group.

The next hypothesis: trust has a positive effect on technology acceptance (H2) was tested separately for the NoXAI and XAI groups, as respondents in each condition were presented with different scenarios.

Table 5 Regression model Trust on TA

	Model 1	B	Std. Error	t	Sig.
NoXAI	Trust	0,730	0,198	3,686	0,001
XAI	Trust	0,716	0,196	3,652	0,001
a. Dependent Variable: TA					

The regression results show that trust has a significant positive effect on technology acceptance in both groups (p = 0.001). In the NoXAI group, trust strongly predicts acceptance (B = 0.73), indicating that higher trust levels are associated with greater acceptance when no explainable AI is provided. Similarly, in the XAI group, trust remains a significant predictor of acceptance (B = 0.716, p = 0.001). While the effect of trust is slightly lower in the XAI group, the difference is minimal, suggesting that trust plays a crucial role in technology acceptance regardless of the presence of explainable AI. These findings do not reject the hypothesis in both conditions, reinforcing that trust positively influences technology acceptance, even when respondents are exposed to different scenarios.

Table 6 Model 3 regression Trust on PU

	Model 2	B	Std. Error	t	Sig.
NoXAI	Trust	0,689	0,316	2,182	0,038
XAI	Trust	0,853	0,231	3,689	0,001
a. Dependent Variable: PU					

The hypothesis that trust has a positive effect on PU (H3) was also tested separately for the NoXAI and XAI groups. The regression results indicate that trust significantly influences PU in both conditions, though the strength of the effect varies. In the NoXAI condition, trust has a moderate positive effect on PU (B = 0.689, SE = 0.316, t = 2.182, p = 0.038), suggesting that higher trust levels lead to greater PU when no explainable AI is provided. However, the effect is stronger in the XAI condition (B = 0.853, SE = 0.231, t = 3.689, p = 0.001), indicating that trust

plays an more critical role in shaping PU when XAI is present. The larger coefficient and stronger significance in the XAI group suggest that users who trust the system perceive it as more useful when explanations are provided, highlighting the importance of explainability in reinforcing the relationship of trust on PU. Following these results it can be concluded that this hypothesis is not rejected.

Table 7 Regression of PEOU on trust

	Model 3	B	Std. Error	t	Sig.
NoXAI	PEOU	0.452	0.121	3.729	<0.001
XAI	PEOU	0.391	0.154	2.535	0.019
a. Dependent Variable: Trust					

Table 6 shows the regression of PEOU on trust (H4) once again separated in two groups, one with XAI and one without XAI. The results in this table indicate that PEOU has a significant positive effect on trust in both groups. In the NoXAI group this effect shows that respondents who perceive technology as easy to use have a higher trust in the AI tool ($B = 0.452$, $t = 3.729$, $p < 0.001$). Interestingly, the group with XAI shows the same result although this interaction is slightly lower than the group without XAI indicating a weaker relationship between PEOU and trust when XAI is present ($B = 0.391$, $t = 2.535$, $p = 0.019$). These findings support theories that trust increases when the ease of use increases. Furthermore, these results show that H4 will not be rejected based on the regression results.

Finally, hypothesis (H5) regarding the moderating effect of XAI on the relationship between trust and TA was tested. In this regression model, trust, XAI, and their interaction term (trust*XAI) were included as independent variables to examine their influence on TA, the dependent variable. Table 5 below presents the results of this analysis.

Table 8 Moderation regression

Model 4	B	Std. Error	t	Sig.
Trust	0.730	0.186	3.920	<0.001
XAI	0.239	1.329	0.172	0.864
Trust*XAI	-0.014	0.283	-0.049	0.961
a. Dependent Variable: TA				

The result show that trust has a significant influence on TA ($B = 0.730$, $t = 3.920$, $p < 0.001$) this interaction has already been found in H2. The direct effect of XAI on TA show no significant effect ($B = 0.239$, $t = 0.172$, $p = 0.864$), meaning that the addition of XAI has no significant positive influence on TA. The interaction variable shows if the XAI moderates the

relationship between trust an TA and shows no significant results (B = -0.014, t = -0.049, p = 0.961). These results suggest that while trust has a significant influence on TA, XAI does not appear to further strengthen or weaken this relationship, and thus hypothesis 5 is rejected based on the results.

Table 9 Pearson Correlations model

	Trust	TA	PU	PEOU	XAI
Trust	1.00	.598**	.441**	.574**	-.303*
TA	.598**	1.00	.871**	.797**	-0.10
PU	.441**	.871**	1.00	.397**	0.04
PEOU	.574**	.797**	.397**	1.00	-0.24
XAI	-.303*	-0.10	0.04	-0.240	1.00
**. Correlation is significant at the 0.01 level (2-tailed).					
*. Correlation is significant at the 0.05 level (2-tailed).					

To provide more insight in the relationship between variables of trust, TA, PU, PEOU and XAI a correlation analysis was also performed. The results show a strong correlation between trust and TA as is expected following the linear regression also showing a positive significant relationship between trust and TA ($r = 0.598$, $p < 0.01$). Furthermore there is are positive relationship between trust and PU ($r = 0.441$, $p = 0.01$) and between trust and PEOU ($r = 0.574$, $p = < 0.01$). This reinforces the idea that trust is positively related with the ease of use and usefulness of AI tools.

Interestingly the correlations shows a significant negative correlation between XAI and trust ($r= -0.303$, $p= 0.029$), XAI and TA also show a negative correlation however this is not significant. This indicates that XAI had a negative effect on trust and shows signs of negative correlation with technology acceptance.

Overall, this analysis shows that trust plays a big part in technology acceptance, ease of use and usefulness while also raising questions about the unexpected opposite results of XAI.

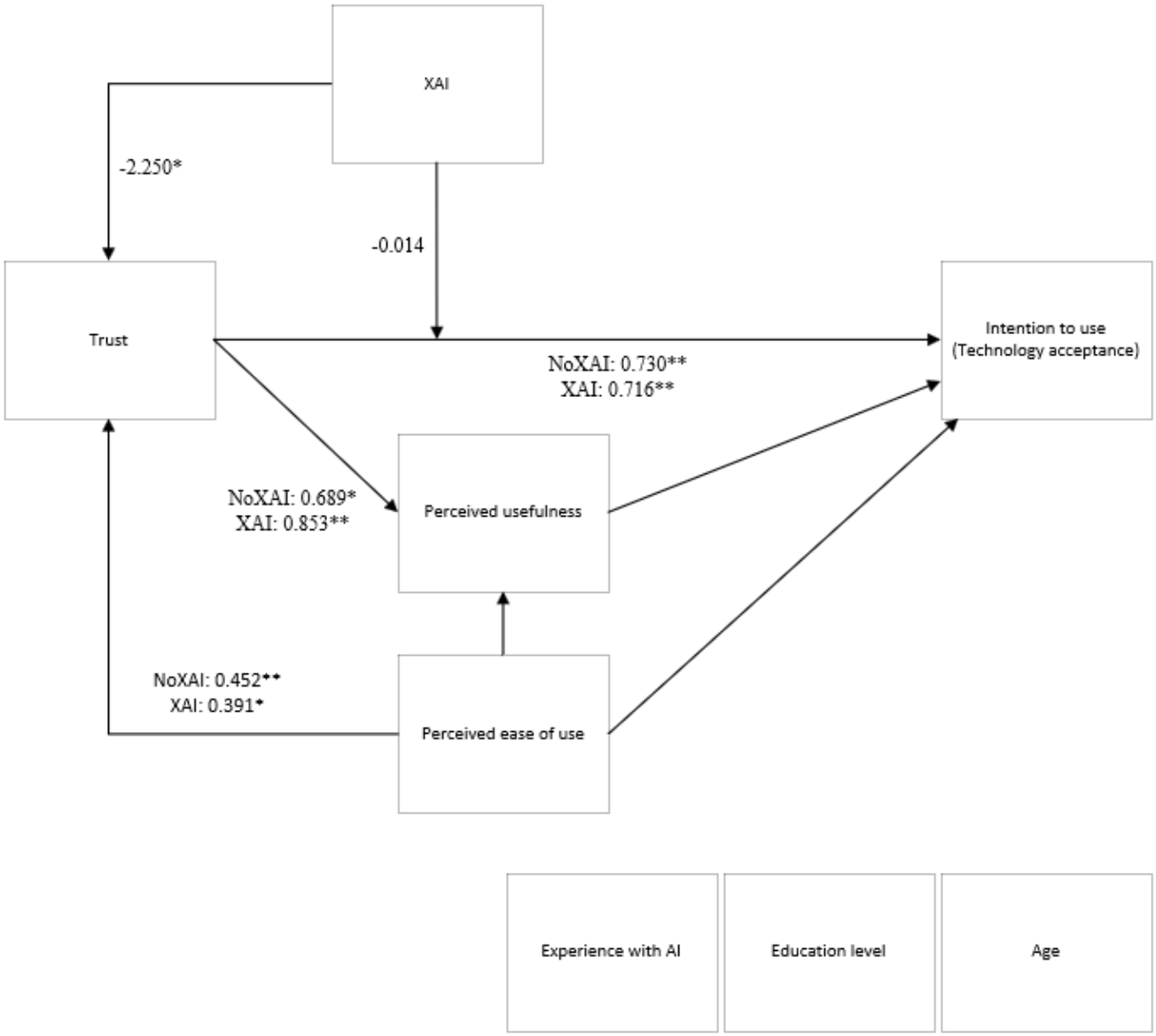


Figure 8 Summary of results in conceptual model.

Lastly some environmental variables were considered in relation to trust and TA. First, the impact of AI experience was examined, revealing that greater experience with AI is associated with higher trust in AI. Similarly, as users' experience with AI increases, their technology acceptance also rises, as illustrated in table 8.

Table 10 Experience with AI on trust and TA

Experience with artificial intelligence		
None	TA	3.85
	Trust	4.31
A little (<1 year)	TA	4.77
	Trust	4.85

Some (1 - 2 years)	TA	4.69
	Trust	4.58
Quite a bit (3-4 years)	TA	5.20
	Trust	4.70
A lot (>4 years)	TA	5.50
	Trust	5.83

The influences of education level on trust and TA was also analysed and shows the same trend as with experience with AI. As education level increases, TA consistently rises, suggesting that individuals with higher education levels are more likely to accept and adopt AI technologies. Trust also shows increases but this is only a slight increase. Respondents with no formal education reported the lowest levels of TA (3.75) and trust (3.33), while those with a master's degree exhibited the highest TA (5.01) and relatively high trust (4.78). Trust does not increase as consistently as TA. Overall these findings suggest that education level does influence trust in AI and TA.

Table 11 Mean trust grouped on education level

Education level		
None	TA	3.75
	Trust	3.33
Secondary/high school education	TA	4.38
	Trust	4.32
Secondary vocational education (MBO)	TA	4.19
	Trust	4.74
Associate's degree	TA	4.92
	Trust	4.08
Bachelor's degree (HBO/WO Bachelor)	TA	4.73
	Trust	4.80
Master's degree (WO Master)	TA	5.01
	Trust	4.78

This analysis was also done with the variable age however, this showed no interesting influence in trust or TA.

6 Discussion

This study aims to investigate the moderating effect of explainable AI on the relationship between trust and technology acceptance, in the context of cybersecurity tools. To answer this question, a survey was constructed using existing scales developed for trust and technology acceptance. Two different versions of the survey were created: one with the inclusion of XAI and one that did not include XAI. Each respondent was randomly allocated to one of these surveys this allocation was evenly distributed to make sure both survey have enough respondents. The survey itself was also distributed through the social media platforms LinkedIn, WhatsApp, Instagram and Facebook.
The key findings in this study are the positive significant relation between trust and technology acceptance, and the significant negative correlation between XAI and trust. A significant positive relationship was found between trust and perceived usefulness and between perceived ease of use and trust. However, no significant relation was found between XAI as a moderator of trust and technology acceptance. Given these results, there is still a lot of potential for studying XAI in relation to trust or technology acceptance. When looking at previous studies on this subject, explanations might be found for the results of this study which will be discussed below.

The first hypothesis (H1) stated the following: *XAI has a positive influence on trust*. Evidence for this hypothesis has not been found, on the contrary: a negative influence of XAI on trust has been found in the current study. A possible explanation for these results, is that the AI of this survey is already easy to understand and thus already more trustworthy. As stated by Lukyanenko (2022), AI can benefit from explanations. However, when AI gets either too complex or is already easy to understand, the AI will not benefit from explanation. For example, when AI is very complex, not just the users but also the developers of these AI systems will not understand the inner workings of their technology, and therefore the explainability of AI is not a realistic goal. This shows that there could be a fine line between AI that is too simple and will not need explanations to AI that is too complex and cannot be explained. This was also made apparent in the literature review by Jin and Sendhoff (2008) and Freitas (2004) the positive effects of the AI can be compromised when explainability is increased since the model becomes simpler. Possibly, in the current research, an XAI was used that did not contribute to the explainability of the AI, because the AI was already clear enough and therefore did not need further explanation.
Additionally, unintuitive explanations may have also played a part in the negative relation between XAI and trust (Schmidt et al., 2020). In other words, the understandability of an explanation could be the problem. Further support for this explanation was the feedback

received by respondents, most respondents mentioned that the survey was hard to understand, used difficult wording and was not available in their native language. These are factors that could influence the way respondents understood the scenarios and could influence the results.

Finally, a possible explanation for no significant increase in trust with XAI, might be that XAI alone is not enough to increase trustworthiness, especially for the less technically experienced users (von Eschenbach, 2021). As was stated in the introduction, trust is hard to conceptualize because there are a lot of variables that should be considered and not just transparency.

Furthermore, H2 stated: *trust has a positive influence on technology acceptance of AI.* After looking at the results, this hypothesis is accepted as a significant positive relationship was found between trust and technology acceptance. This means that more trust results in a higher rate of technology acceptance. This result supports different previous studies that were done on this concept (e.g. Wu et al., 2011; Dickson et al., 2021). Some studies state that trust is generally important in the adoption of new technologies which is supported by the findings from this study (Fukuyama, 1996). The importance of trust on the technology acceptance is dependent on the situation that was measured. For example, a different level of trust can be found between commercial AI tools, and non-commercial AI tools. When the research object is commercial, the level of trust is explicitly lower than when the research object is non-commercial, like measuring technology acceptance (Tan and Thoen, 2001). In the case of this study there was no commercial research object but rather a security issue. In other words, it is easier for users to trust an AI system when it's not trying to sell a product.

H3 stated *Trust in AI has a positive influence on perceived usefulness of AI.* Gefen et al. (2003) found that an increase of trust makes users more likely to find a tool useful. Dahlberg et al. (2003) also found similar results, the persons attitude can affect the perceived usefulness and ease of use of the tool. This study finds comparable results as a significant positive effect was found between trust and PU, thus this hypothesis can be accepted.

H4 stated that *perceived ease of use has a positive influence on trust in AI*. As was theorized by Ganesan (1994), PEOU does show an significant positive relationship with trust. Other studies show similar results when investigating the relationship between PEOU and trust (Nuseir et al., 2022). This shows that when users find tools easy to use they also find the tools more trustworthy. The results show that PEOU has a significant positive effect on trust and thus this hypothesis is accepted.

H5 stated the following: *The relationship between trust and technology acceptance is positively*

moderated by XAI. In the current study no evidence was found for the moderating role of XAI on the relationship between trust and technology acceptance. Meaning that H5 cannot be accepted.
There could be many explanations for these results apart from any limitations in this study (Section 6.2), one of these possible explanation could be a lack of trust from unintuitive explanations, as theorized by Schmidt et al. (2020). This study states that a lack of trust can emerge from unintuitive explanations, which means that explanations might be difficult to understand. When looking at the current study, it might be a reason for the results if respondents did not understand the questions in the survey or find them intuitive, which could mean the score on trust would not increase. This could even be a reason for the negative relation found between XAI and trust.

Moreover, Alufaisan et al. (2021) have presented mixed results regarding advantages of employing XAI. In their study, no evidence was found that users follow explained AI choices more often than non-explained AI choices. This aligns with the findings in the current study, and is not the only study where (AI) explanation adds no benefits (Green and Chen 2019; Poursabzi-Sangdeh et al., 2018; Zhang et al., 2020). As stated in the literature survey Scharowski et al. (2022) also found varying studies with different results on the impact of XAI on trust, in this case the difference in measurements where speculated to be the cause of these varying results. Another theory for finding no added benefits for explanations is that providing more information does not necessarily mean that there is more trust in a system or that the explanation would be more accurate (Gigerenzer and Brighton 2009; Goldstein and Gigerenzer 2002; Nadav-Greenberg and Joslyn 2009). This scenario can be attributed to cognitive limitations, and coincides with Poursabzi-Sangdeh et al. (2018) stating that explanation can cause an information overload. This could be the case for this study, as it was a lot of information was provided in the XAI case, to get a good understanding of the AI. All those factors considered, it could be possible that the scenario has led to an information overload.

Given these results it can be concluded that even though trust would fit within a TAM model, it cannot be said that XAI has a significant positive influence on trust or technology acceptance and is not the solutions to all the transparency problems of AI.

6.1 Implications

Theoretical

This study adds to the growing body of research on XAI in cybersecurity by addressing the key gap of how explainability influences trust among general users rather than IT experts. While

much of the existing research focuses on applying XAI to cybersecurity datasets, this study looks at how XAI affects real-world cybersecurity tools and their adoption. XAI has been shown to be a good fit for use in applications (Srinivasu et al., 2022). However, this study shows that not all areas of applications react the same to XAI being introduced. The findings suggest that XAI does not always lead to increased trust, challenging the assumption that explainability is universally beneficial. In some cases, cybersecurity tools may already be perceived as trustworthy without the need for additional explanations, meaning XAI might not always be necessary.

The application of XAI in a realistic setting while focussing on cybersecurity tools is a contribution because literature reviews of the concepts XAI for cybersecurity shows that most research focusses on the application of XAI on cybersecurity datasets and not on cybersecurity tools (Charmet et al., 2022).
Apart from XAI, the concept and knowledge of technology acceptance has also been expanded upon. The influence of trust on technology acceptance was already found in previous studies and has been found once again in a cybersecurity environment (Gefen et al., 2003; Baroni et al., 2022; von Eschenbach, 2021).
These results also raise questions about current trust models, such as the FATE framework, which assume that explainability and transparency directly contribute to trust. Instead, this study suggests that trust in AI is more complex, factors like perceived accuracy and reliability may play a bigger role than transparency alone. However, it is possible that explainability, when combined with fairness and accountability, could still have a more significant impact on trust, something future research should explore.

By showing that XAI does not always enhance trust and could even reduce it in some cases, this study provides a foundation for further research into when and how XAI should be applied in cybersecurity. These insights help refine our understanding of XAI's role in security tools and highlight the need for a more defined approach to its implementation.

Practical implications

Although cybersecurity firms have yet to widely implement XAI in their tools or workflows, this research does still provides valuable insights for businesses considering its adoption.
A key takeaway of this research is that XAI cannot be considered a one-size-fits-all solution for businesses looking to enhance their customers trust. The findings suggest that AI-driven cybersecurity tools do not benefit from the adoption of XAI and the added explainability this provides. Especially when a tool is already intuitive XAI can lead to information overload or add confusion. Businesses should critically asses their customers' needs on a case-by-case basis

to ensure that any changes like the addition of XAI aligns with their user's needs. Factors such as the complexity of the AI system, the technical proficiency of users, regulatory requirements, and the level of risk associated with the AI's decisions should be considered before implementing XAI.
Furthermore, businesses can use these findings to refine how they present AI explanations to maximize trust. Research suggests that different XAI techniques, such as Partial Dependence Plots (PDP), SHAP, or LIME, vary in effectiveness depending on user familiarity and cognitive load (Dwivedi et al., 2023). Rather than assuming that all XAI increases trust companies should explore what explanation techniques work best for the given situation. This can range from basic explanations to detailed graphs and technical explanations. It might be better to choose for a more flexible, adaptive system to help prevent information overload.

Other implications are the reinforcement of the critical role of trust in technology acceptance, a factor businesses should consider when using AI tools in their security products. As was stated in the theoretical implications the link between trust and technology acceptance has already been shown and was further confirmed by this study, firms should recognize this relationship when trying to increase the acceptance of their technology. Not just trust but other concepts like reliability accuracy and usability should be recognized as possible influences of technology acceptance and trust.

6.2 Limitations and future research

A limitation of this study is that when creating the survey, the choice was made not to translate the survey, even though the majority of the respondents were going to be Dutch. This choice was made as translation can make slight changes to the meaning of the items in the scales and impact the interpretability (Baumgartner & Weijters, 2017). However this means that the survey was only available in English and this could have an influence on how respondents with a different native languages would interpret the items in the survey. As it was expected that most of the respondents were going to be Dutch this could have impacted how each respondents filled out the survey. So for this study it could be the case that the questions or the scenario of the survey were not clear to the respondent leading to unreliable responses.

Furthermore, a limitation is the size of the study as a lot of responses had to be deleted due to unfinished responses. The final population was only 52, Baruch (1999) states that any response rate below 60 +- 20 could impact the generalizability of the study. For this study the effective response rate is 24-26 as the survey is halved due to the addition of two different scenarios. This

could have been because the survey was only public for about a month or due to the large amount of incomplete responses.
Finally, this study used two different scenarios for this survey, to work in this context the respondents need to be as random as possible. As the survey was mainly completed by networking and social media there would be no guarantee that the respondents were completely random as the network of the researcher mainly contains students. Thus the sample for this study was not large or representative enough.

Another limitation that could have impacted the study is that the variable intention to use, which is part of the TAM model, was never measured during the survey period. Due to not knowing the real intention to use the technology acceptance has been based on perceived usefulness and perceived ease of use. If the intention to use the technology had been measured the study would have been more reliable.

Future research is needed to explore the importance of context when defining the research object. A possible research subject can be the difference in XAI results between a commercial and a non-commercial setting. The results might be significantly different when the context is changed. Trust or technology acceptance can change significantly if the context provided is, for example, stressful. Not only context is important, but also a the use of a baseline measurement of trust. In future research, the focus should be on measure trust at different times to accomplish this. Instead of the two scenarios, a combination can be made that presents both scenarios with and without XAI to the same respondent and trust can be measured after each scenario. Measuring trust twice will show the level of trust before and after each scenario, to gain insight in the amount of trust a respondent already has, and moreover making sure that there is not already a lot of trust in a scenario or AI tool. Measuring at different times already provides some insight into the understandability of the scenario, however adding understandability as a concept in the questionnaire can provide interesting results in future research.
Future research could also explore how demographic factors, like age, influence AI acceptance, as well as how varying education levels impact perceptions of AI. Additionally, examining user experience with AI could provide valuable insights, particularly in distinguishing between general users and expert users.

7 Conclusion

This study explores the insights into the effect of trust on technology acceptance with XAI as a moderator. This study used data from a questionnaire with a population of 52. The questionnaire used two scenarios: one where XAI is used in conjunction with a cybersecurity tool and one without the XAI. With the use of regression analysis and T-tests on the data, five hypotheses where answered. The results show that XAI did not have a significant positive influence on trust. The relationship between trust and technology acceptance was found to be significantly positive. The relationship between trust and perceived usefulness was found to be significantly positive. The relationship between perceived ease of use and trust was also found to be significantly positive. Lastly, the there is no positive moderating effect of XAI on the relation between trust and technology acceptance. On the contrary, XAI was found to have a significant negative impact on trust. Two of the five hypotheses were rejected and three were accepted, these findings may be influenced by the technological or user context in which the study was conducted. Possible explanations for these results like unintuitive explanations, information overload or a simplistic AI model are described in the discussion section.

During the process of this study, some important limitations of the study were discovered. Firstly, there was a low amount of usable responses for this study. Even though there where over 100 responses to the questionnaire, only 52 responses were usable. Due to a limited amount of time to keep the questionnaire online, it was not possible to increase the number of usable responses. Secondly, some respondents gave feedback about the questionnaire, which mentioned the difficult wording of the questionnaire and the language barrier as most of the respondents were native Dutch speakers. The language barrier and difficult wording were already a known issue when developing the questionnaire, however the choice was made not to translate the questionnaire, as this could influence the interpretation and therefore outcome of the questionnaire. Lastly, although insights were gained about the interaction of XAI and trust, some concepts like context and interpretability were not accounted for in the current study.

The addition of the concepts context and interpretability could be interesting for future studies, and could build upon the research presented in this study. Creating more understanding of XAI and its interactions with concepts like trust and technology acceptance can provide answers to business questions. For example, in what contexts XAI would feel the most trustworthy, or whether or not company’s should use XAI in an effort to increase trust. With the addition of future research on different concepts, more questions can be answered like what the impact of XAI could be on AI use, when is XAI the most efficient. The results might even provide a foundation for governmental decisions, like regulating the use of AI without XAI. The use of

XAI could be a big development for the implementation of AI, which makes the current study an important contribution to this field of research.

References

Abood, O. G., & Guirguis, S. K. (2018). A Survey on Cryptography Algorithms. *International Journal of Scientific and Research Publications (IJSRP)*, *8*(7). https://doi.org/10.29322/ijsrp.8.7.2018.p7978

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/access.2018.2870052

Aechtner, J., Cabrera, L., Katwal, D., Onghena, P., Valenzuela, D. P., & Wilbik, A. (2022). Comparing User Perception of Explanations Developed with XAI Methods. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–7. https://doi.org/10.1109/fuzz-ieee55066.2022.9882743

Ahmed, I., Jeon, G., & Piccialli, F. (2022). From Artificial Intelligence to Explainable Artificial intelligence in Industry 4.0: A survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, *18*(8), 5031–5042. https://doi.org/10.1109/tii.2022.3146552

Arrieta, A. B., 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

Ashoori, M., & Weisz, J. D. (2019, December 5). *In AI we trust? Factors that influence trustworthiness of AI-infused Decision-Making processes*. https://doi.org/10.48550/arXiv.1912.02675

Balantrapu, S. S. (2024). A Comprehensive Review of AI Applications in Cybersecurity. *International Machine learning journal and Computer Engineering*, *7*(7).

Baroni, I., Calegari, G. R., Scandolari, D., & Celino, I. (2022). AI-TAM: a model to investigate user acceptance and collaborative intention inhuman-in-the-loop AI applications. *Human Computation*, *9*(1), 1–21. https://doi.org/10.15346/hc.v9i1.134

Baruch, Y. (1999). Response Rate in Academic Studies-A Comparative Analysis. Human Relations, 52(4), 421–438. https://doi.org/10.1177/001872679905200401

Baumgartner, H., & Weijters, B. (2017). Methodological Issues in Cross-Cultural Research. In Springer eBooks (pp. 169–190). https://doi.org/10.1007/978-3-319-65091-3_10

Bengio, Y. (2009). Learning deep architectures for AI. Foundations and Trends® in Machine Learning, 2(1), 1–127. https://doi.org/10.1561/2200000006

Bharadiya, J.P., Thomas, R.K., & Ahmed, F. (2023). Rise of artificial intelligence in business and industry. *Journal of Engineering Research and Reports*, 25(3), 85–103. https://doi.org/10.9734/jerr/2023/v25i3893

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's Reducing a Human Being to a Percentage." *N CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–14. H*. https://doi.org/10.1145/3173574.3173951

Brandon, M. G. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1), 421. https://doi.org/10.32614/rj-2017-016

Carr, A. S., Zhang, M., Klopping, I., & Min, H. (2010). RFID tech-nology: Implications for healthcare organizations. American Journal of Business, 25(2), 25– 40. https://doi.org/10.1108/19355181201000008

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Charmet, F., Tanuwidjaja, H.C., Ayoubi, S. (2022) Explainable artificial intelligence for cybersecurity: a literature survey. *Ann. Telecommun*. 77, 789–812 . https://doi.org/10.1007/s12243-022-00926-7

Cheng, H., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300789

Chui, K. T. (2022). Building digital Trust: Challenges and strategies in Cybersecurity. *Cyber Security Insights Magazine*, *05*, 15. Retrieved from: https://insights2techinfo.com/wp-content/uploads/2023/08/Building-Digital-Trust-Challenges-and-Strategies-in-Cybersecurity.pdf

Dahlberg, T., Mallat, N., & Öörni, A. (2003). Trust enhanced technology acceptance modelconsumer acceptance of mobile payment solutions: Tentative evidence. Stockholm Mobility Roundtable, 22(1), 145.

Das, A., & Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): a survey. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2006.11371

Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of Artificial Intelligence in Machine Learning: Review and Prospect. *International Journal Of Computer Applications*, *115*(9), 31–41. https://doi.org/10.5120/20182-2402

Da'u, A., & Salim, N. (2020). Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review*, 53(4), 2709–2748. https://doi.org/10.1007/s10462-019-09744-1

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

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing, 3,* e20. https://doi.org/10.1017/atsip.2013.9

Dhagarra, D., Goswami, M., & Kumar, G. (2020). Impact of Trust and Privacy Concerns on Technology Acceptance in Healthcare: An Indian Perspective. *International Journal of Medical Informatics*, 141, 1386-5056. https://doi.org/10.1016/j.ijmedinf.2020.104164

Dickson B.U., Oby B.O., Samuel N.N., Udoka S.O. (2021). Integrating Trust into Technology Acceptance Model (TAM), the Conceptual Framework for E-Payment Platform Acceptance. *British Journal of Management and Marketing Studies. 4*(4), 34-56. https://doi.org/10.52589/BJMMSTB3XTKPI

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1710.00794

Dorta-González, P., López-Puig, A. J., Dorta-González, M. I., & González-Betancor, S. M. (2024). Generative artificial intelligence usage by researchers at work: Effects of gender, career stage, type of workplace, and perceived barriers. *Telematics and Informatics*, *94*, 102187. https://doi.org/10.1016/j.tele.2024.102187

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1702.08608

Dosilovic, F. K., Brcic, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *MIPRO 2018*. https://doi.org/10.23919/mipro.2018.8400040

Dupond, S. (2019). A thorough review on the current advance of neural network structures. *Annual Reviews in Control, 14,* 200–230. https://doi.org/10.1016/j.arcontrol.2019.03.001

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 1-33. https://doi.org/10.1145/3561048

Fallatah, W., Kävrestad, J., & Furnell, S. (2024). Establishing a model for the user acceptance of cybersecurity training. *Future Internet*, *16*(8), 294. https://doi.org/10.3390/fi16080294

Familoni, B. T. (2024). Cybersecurity challenges in the age of ai: theoretical approaches and practical solutions. *Computer Science & IT Research Journal*, *5*(3), 703–724. https://doi.org/10.51594/csitrj.v5i3.930

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in human-agent collaboration. *ECCE '08:*

Proceedings of The5th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction. https://doi.org/10.1145/1473018.1473028

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Foroughi, F., & Luksch, P. (2018). Data science methodology for cybersecurity projects. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1803.04219

Freitas, A. A. (2004). A critical review of multi-objective optimization in data mining. *ACM SIGKDD Explorations Newsletter*, *6*(2), 77–86. https://doi.org/10.1145/1046456.1046467

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5). https://doi.org/10.1214/aos/1013203451

Fukuyama, F. (1996). Trust: The social virtues and the creation of prosperity. New York: Free Press.

Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019). Explainable AI in Industry. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'19)*. (Vols. 3203–3204). ACM. https://doi.org/10.1145/3292500.3332281

Ganesan, S. (1994). Determinants of Long-Term orientation in Buyer-Seller Relationships. *Journal of Marketing*, 58(2), 1. https://doi.org/10.2307/1252265

Gefen, N., Karahanna, N., & Straub, N. (2003). Trust and TAM in online shopping: an integrated model. *MIS Quarterly*, 27(1), 51. https://doi.org/10.2307/30036519

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical research. *Academy of Management Annals*, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goldstein, Daniel & Gigerenzer, Gerd. (2002). Models of Ecological Rationality: The Recognition Heuristic. *Psychological review, 109,* 75-90. https://doi.org/10.1037//0033-295X.109.1.75

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). Cambridge: MIT press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1406.2661

Goodman, B., & Flaxman, S. (2017). European Union regulations on Algorithmic Decision making and a "Right to Explanation." *AI Magazine*, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Green, B., & Chen, Y. (2019). The principles and Limits of Algorithm-in-the-Loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*, 1–24. https://doi.org/10.1145/3359152

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainableAI(XAI) program: A retrospective. *Applied AI Letters*, 2(4). https://doi.org/10.1002/ail2.61

Hamilton, W. L. (2020). Graph Representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *14*(3), 1–159. https://doi.org/10.2200/s01045ed1v01y202009aim046

Heierhoff, S. & Choun, I. H. (2023). The Impact of Cybersecurity and Innovation on Mobility Technology Acceptance. *PACIS 2023 Proceedings*. 53. Retrieved from: https://aisel.aisnet.org/pacis2023/53

Henshel, D., Cains, M., Hoffman, B., & Kelley, T. (2015). Trust as a human factor in holistic cyber Security risk assessment. *Procedia Manufacturing*, *3*, 1117–1124. https://doi.org/10.1016/j.promfg.2015.07.186

Hinton, G. (2009). Deep belief networks. Scholarpedia, 4(5), 5947. https://doi.org/10.4249/scholarpedia.5947

Hoff, K. A., & Bashir, M. (2015). Trust in automation. *Human Factors the Journal of the Human Factors and Ergonomics Society*, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Hunt, T., Zhu, Z., Xu, Y., Peter, S., & Witchel, E. (2017). Ryoan: A Distributed Sandbox for Untrusted Computation on Secret Data. *ACM Transactions on Computer Systems*, 35(4), 1–32. https://doi.org/10.1145/3231594

IBM Threat Detection and Response Services. (n.d.). https://www.ibm.com/services/threat-detection-response

Irfan, M., Abbas, H., Sun, Y., Sajid, A., & Pasha, M. (2016). A framework for cloud forensics evidence collection and analysis using security information and event management. *Security and Communication Networks*, 9(16), 3790-3807. https://doi.org/10.1002/sec.1538

Jack, B., & Clarke, A. M. (1998). The purpose and use of questionnaires in research. *Professional nurse* (London, England), 14(3), 176–179.

Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/s15327566ijce0401_04

Kasowaki, L., & Emir, K. (2023). AI and Machine Learning in Cybersecurity: Leveraging Technology to Combat Threats (No. 11610). EasyChair.

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Kim, B., Glassman, E.L., Johnson, B., & Shah, J.A. (2015). iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction.

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464–1480. https://doi.org/10.1109/5.58325

Kok, J. N., N., Boers, E. J. W., W., Kosters, W. A., A., Van Der Putten, P., & Mannes Poel (2009). Artificial intelligence: definition, trends, techniques, and cases. In: Encyclopedia of Life Support Systems (EOLSS), *Encyclopedia of Life Support Systems (EOLSS)*. https://www.eolss.net/Sample-Chapters/C15/E6-44.pdf

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A Conversational Agent as Museum Guide – Design and evaluation of a Real-World Application. *Lecture notes in computer science* (pp. 329–343). https://doi.org/10.1007/11550617_28

Kraus, L., Wechsung, I., & Möller, S. (2017). Psychological needs as motivators for security and privacy actions on smartphones. *Journal of Information Security and Applications*, *34*, 34–45. https://doi.org/10.1016/j.jisa.2016.10.002

Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. 2019. Explainable AI in industry. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'19). ACM*, New York, NY, 3203–3204. https://doi.org/10.1145/3292500.3332281

Lai, V., & Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38. https://doi.org/10.1145/3287560.3287590

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

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

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors the Journal of the Human Factors and Ergonomics Society*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1). https://doi.org/10.1177/2053951718756684

Lee, M. K., & Rich, K. (2021). Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445570

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9), 1-46. https://doi.org/10.1145/3555803

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

Lukyanenko, R., Maass, W. & Storey, V.C. Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electron Markets. 32*, 1993–2020 (2022). https://doi.org/10.1007/s12525-022-00605-4

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.1705.07874

Madakam, S., Uchiya, T., Mark, S., & Lurie, Y. (2022). Artificial Intelligence, Machine Learning and Deep Learning (Literature: Review and Metrics). *Asia-Pacific Journal Of Management Research And Innovation*, *18*(1–2), 7–23. https://doi.org/10.1177/2319510x221136682

Mandic, D., & Chambers, J. (2001). Recurrent neural networks for prediction: Learning algorithms, architectures, and stability. *Wiley*. https://doi.org/10.1002/047084535X

Marangunić, N., & Granić, A. (2015). Technology acceptance model: A literature review from 1986 to 2013. *Universal Access in the Information Society, 14*(1), 81–95. https://doi.org/10.1007/s1020 9- 014- 0348- 1

Marda, V. (2018). Artificial intelligence policy in India: a framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, 376(2133), 20180087. https://doi.org/10.1098/rsta.2018.0087

Marlin, B., Swersky, K., Chen, B., & Freitas, N. (2010). Inductive principles for restricted Boltzmann machine learning. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 509-516). JMLR Workshop and Conference Proceedings.

Mimoun, M. S. B., Poncin, I., & Garnier, M. (2012). Case study—Embodied virtual agents: An analysis on reasons for failure. *Journal of Retailing and Consumer Services*, 19(6), 605–612. https://doi.org/10.1016/j.jretconser.2012.07.006

Mirhoseini, A., Pham, H., Le, Q. V., Steiner, B., Larsen, R., Zhou, Y., Kumar, N., Norouzi, M., Bengio, S., & Dean, J. (2017). Device Placement Optimization with Reinforcement Learning. *ArXiv*. https://arxiv.org/abs/1706.04972

Mohammadi, S., Mirvaziri, H., Ghazizadeh-Ahsaee, M., & Karimipour, H. (2019). Cyber intrusion detection by combined feature selection algorithm. *Journal of Information Security and Applications*, 44, 80-88. https://doi.org/10.1016/j.jisa.2018.11.007

Montavon, G., Samek, W., & Müller, K. (2017). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Mou, J., Shin, D., & Cohen, J. (2016). Understanding trust and perceived usefulness in the consumer acceptance of an e-service: a longitudinal investigation. *Behaviour and Information Technology*, 36(2), 125–139. https://doi.org/10.1080/0144929x.2016.1203024

Nadav-Greenberg, L., & Joslyn, S. L. (2009). Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making, 3*(3), 209–227. https://doi.org/10.1518/155534309X474460

Nuseir, M. T., Aljumah, A. I., & Refae, G. a. E. (2022). Trust in adoption of Internet of Things: role of perceived ease of use and security. *2022 International Arab Conference on Information Technology (ACIT)*, 16, 1–7. https://doi.org/10.1109/acit57182.2022.9994207

Pavlou, P. A. (2003). Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model. *International Journal of Electronic Commerce*, 7(3), 101–134. https://doi.org/10.1080/10864415.2003.11044275

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://doi.org/10.5555/1953048.2078195

Perols, R., & Murthy, U. S. (2018). The impact of cybersecurity risk management examinations and cybersecurity incidents on investor perceptions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3112872

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). *Manipulating and measuring model interpretability*. *arXiv*. https://arxiv.org/abs/1802.07810

Qi, H., Di, X., & Li, J. (2018). Formal definition and analysis of access control model based on role and attribute. *Journal of Information Security and Applications*, 43, 53–60. https://doi.org/10.1016/j.jisa.2018.09.001

Rada, R. (1986). Artificial intelligence. *Artificial Intelligence*, 28(1), 119–121. https://doi.org/10.1016/0004-3702(86)90034-2

Rangaraju, S. (2023). Secure by intelligence: enhancing products with ai-driven security measures. *eph - International Journal of Science and Engineering*, 9(3), 36–41. https://doi.org/10.53555/ephijse.v9i3.212

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). https://doi.org/10.1609/aaai.v32i1.11491

Robles-Gomez, A., Tobarra, L., Pastor-Vargas, R., Hernandez, R., & Haut, J. M. (2021). Analyzing the users' acceptance of an IoT cloud platform using the UTAUT/TAM model. *IEEE Access*, *9*, 150004–150020. https://doi.org/10.1109/access.2021.3125497

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2022). A Systematic Literature Review on Phishing email detection using natural language processing techniques. *IEEE Access*, 10, 65703–65727. https://doi.org/10.1109/access.2022.3183083

Sarker, I. H. (2021). Machine learning: algorithms, Real-World applications and research directions. *SN Computer Science*, 2(3). https://doi.org/10.1007/s42979-021-00592-x

Sarker, I. H. (2022). AI-Based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, 3(2). https://doi.org/10.1007/s42979-022-01043-x

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, security intelligence modeling and research directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00557-0

Scharowski, N., Perrig, S. a. C., Nick, V. F., & Brühlmann, F. (2022). Trust and reliance in XAI -- Distinguishing between attitudinal and behavioral measures. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.12318

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Shah, C., Nachand, D., Wald, C., & Chen, P. (2023). Keeping Patient Data Secure in the Age of Radiology Artificial Intelligence: Cybersecurity Considerations and Future Directions. *Journal Of The American College Of Radiology*, *20*(9), 828–835. https://doi.org/10.1016/j.jacr.2023.06.023

Shah, V. (2021). Machine Learning Algorithms for Cybersecurity: Detecting and Preventing Threats. *Revista Espanola de Documentacion Cientifica*, 15(4), 42-66.

Sharma, P., Dash, B., & Ansari, M. F. (2022). Anti-Phishing Techniques – A review of cyber defense mechanisms. *IJARCCE*, 11(7). https://doi.org/10.17148/ijarcce.2022.11728

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial intelligence: definition and background. *Research for policy* (pp. 15–41). https://doi.org/10.1007/978-3-031-21448-6_2

Shin, D. (2020c). Expanding the role of trust in the experience of Algorithmic journalism: User sensemaking of Algorithmic heuristics in Korean users. Journalism Practice, 16(6), 1168–1191. https://doi.org/10.1080/17512786.2020.1841018

Shin, D. (2021b). How do people judge the credibility of algorithmic sources? *AI & Society*, 37(1), 81–96. https://doi.org/10.1007/s00146-021-01158-4

Shin, D. (2022). How do people judge the credibility of algorithmic sources? *AI & Society*, 37(1), 81–96. https://doi.org/10.1007/s00146-021-01158-4

Shin, D. (2023). Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. *Journal of Information Science*, 49(1), 18–31. https://doi.org/10.1177/0165551520985495

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Shrestha, A. K., Vassileva, J., Joshi, S., & Just, J. (2021). Augmenting the technology acceptance model with trust model for the initial adoption of a blockchain-based system. *PeerJ Computer Science*, *7*, 502. https://doi.org/10.7717/peerj-cs.502

Sontan, N. a. D., & Samuel, N. S. V. (2024). The intersection of Artificial Intelligence and cybersecurity: Challenges and opportunities. *World Journal of Advanced Research and Reviews*, 21(2), 1720–1736. https://doi.org/10.30574/wjarr.2024.21.2.0607

Srinivasu, Parvathaneni Naga, Sandhya, N., Jhaveri, Rutvij H., Raut, Roshani (2022) Blackbox to Explainable AI *Healthcare: Existing Tools and Case Studies, Mobile Information Systems*, 8167821, 20 https://doi.org/10.1155/2022/8167821

St John, M. (2024, August 28). *Cybersecurity stats: facts and figures you should know*. Forbes Advisor. https://www.forbes.com/advisor/education/it-and-tech/cybersecurity-statistics/

Szychter, A., Ameur, H., Antonio, C., Kung, H., Daussin, & CoESSI (2018). The Impact of Artificial Intelligence on Security: a Dual Perspective.

Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-Trust. *Minds and Machines*, *20*(2), 243–257. https://doi.org/10.1007/s11023-010-9201-3

Tam, C., Balau, M., & Oliveira, T. (2023). What Influences People's Adoption of Cognitive Cybersecurity?. *International Journal of Human–Computer Interaction*, 1-18. https://doi.org/10.1080/10447318.2023.2279411

Tan, Y., & Thoen, W. (2000). Toward a generic model of trust for electronic commerce. *International Journal of Electronic Commerce*, *5*(2), 61–74. https://doi.org/10.1080/10864415.2000.11044201

Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2012, April). Making machine learning models interpretable. *ESANN* (Vol. 12, pp. 163-172).

Vemuri, N., Thaneeru, N., & Tatikonda, V. M. (2023). Securing Trust: Ethical Considerations in AI for Cybersecurity. *Journal of Knowledge Learning and Science Technology ISSN 2959-6386 (Online)*, *2*(2), 167–175. https://doi.org/10.60087/jklst.vol2.n2.p175

Vilone, G., & Longo, L. (2023). Development of a Human-Centred Psychometric Test for the evaluation of explanations produced by XAI methods. In *Communications in computer and information science*. 205–232. https://doi.org/10.1007/978-3-031-44070-0_11

von Eschenbach, W.J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophical Technology. 34*, 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Walter, E. (2005). Cambridge advanced learner's dictionary.

Wang, W., & Benbasat, I. (2007). Recommendation Agents for Electronic Commerce: Effects of explanation facilities on trusting Beliefs. *Journal of Management Information Systems*, 23(4), 217–246. https://doi.org/10.2753/mis0742-1222230410

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems*, 86, 48–60. https://doi.org/10.1016/j.dss.2016.03.007

Wu, K., Zhao, Y., Zhu, Q., Tan, X., & Zheng, H. (2011). A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management*, *31*(6), 572–581. https://doi.org/10.1016/j.ijinfomgt.2011.03.004

Yin, J., (2016). Firewall policy management. US Patent, 9,338,134.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021, 1-18. https://doi.org/10.1155/2021/8812542

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. https://doi.org/10.1145/3351095.3372852

Zhang, Z., Hamadi, H. A., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access*, *10*, 93104-93139. https://doi.org/10.1109/ACCESS.2022.3204051

Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (2018). Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. *2018 IEEE Conference on Computational Intelligence and Games* (CIG), 1–8. https://doi.org/10.1109/cig.2018.8490433

Appendices

Appendix 1 questionnaire items

Item	Question	Response options
1	The cybersecurity tool is deceptive.	not at all=1 extremely=7
2	The cybersecurity tool behaves in an underhanded matter.	not at all=1 extremely=7
3	I am suspicious of the cybersecurity tool's intent, action, or outputs.	not at all=1 extremely=7
4	I am wary of the cybersecurity tool.	not at all=1 extremely=7
5	The cybersecurity tool's actions will have a harmful or injurious outcome.	not at all=1 extremely=7
6	I am confident in the cybersecurity tool.	not at all=1 extremely=7
7	The cybersecurity tool provides security.	not at all=1 extremely=7
8	The cybersecurity tool has integrity.	not at all=1 extremely=7
9	The cybersecurity tool is dependable.	not at all=1 extremely=7
10	The cybersecurity tool is reliable.	not at all=1 extremely=7
11	I can trust the cybersecurity tool.	not at all=1 extremely=7
12	I am familiar with the cybersecurity tool.	not at all=1 extremely=7

Item	Question	Response options
13	Using this cybersecurity tool would enable me to make decisions more quickly.	not at all=1 extremely=7
14	Using the cybersecurity tool would help me make better decisions.	not at all=1 extremely=7
15	Using the cybersecurity tool would increase my productivity.	not at all=1 extremely=7
16	Using the cybersecurity tool would enhance the effectiveness of my decisions.	not at all=1 extremely=7
17	Using the cybersecurity tool would make it easier to make decisions.	not at all=1 extremely=7

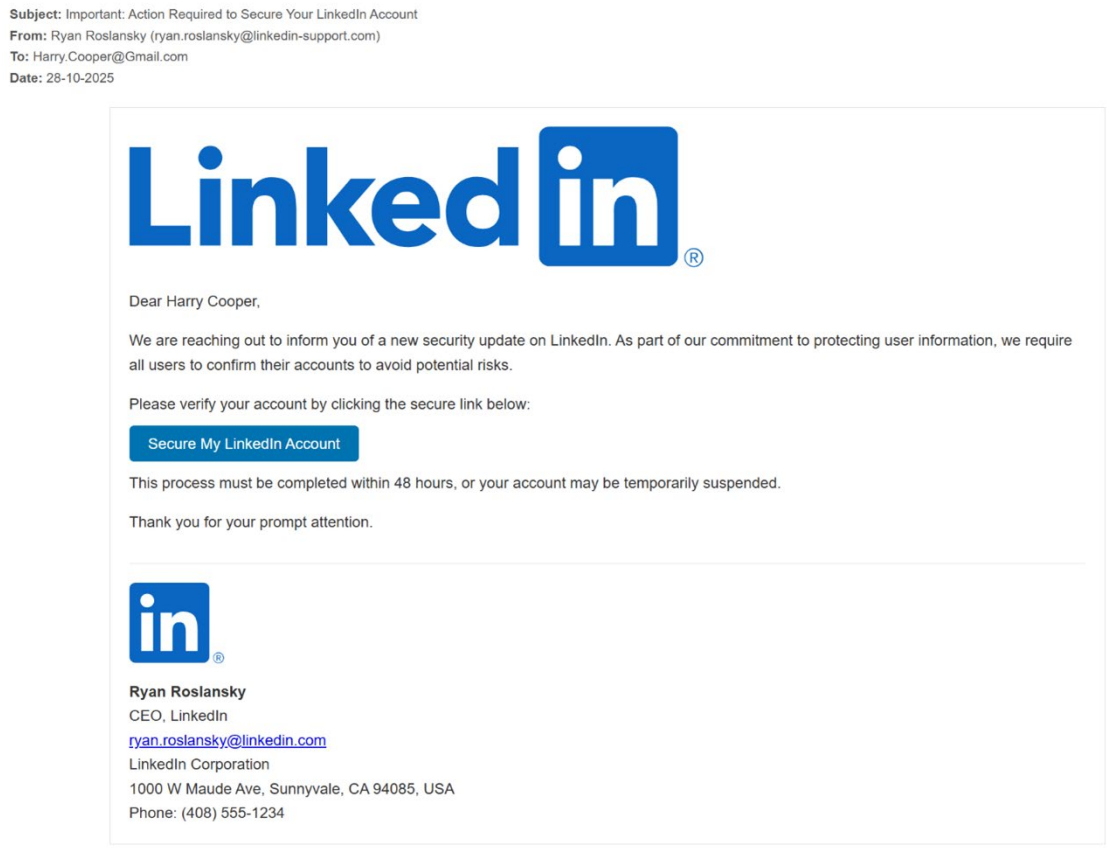
18	I would find the cybersecurity tool useful.	not at all=1 extremely=7
19	Learning to operate the cybersecurity tool would be easy for me.	not at all=1 extremely=7
20	I would find it easy to get the cybersecurity tool to do what I want it to do.	not at all=1 extremely=7
21	My interaction with the cybersecurity tool would be clear and understandable.	not at all=1 extremely=7
22	I would find the cybersecurity tool to be flexible to interact with.	not at all=1 extremely=7
23	It would be easy for me to become skillful at using the cybersecurity tool.	not at all=1 extremely=7
24	I would find the cybersecurity tool easy to use.	not at all=1 extremely=7

Item	Question	Response option
25	What is your age?	-
26	What is the highest level of education you have completed?	None Primary school Secondary/high school education Secondary vocational education (MBO) Associate's degree, Bachelor's degree (HBO/WO Bachelor) Master's degree (WO Master)
27	How would you quantify your experience with artificial intelligence technologies/ machine learning techniques?	None A little Some (1 - 2 years) Quite a bit (3-4 years) A lot (>4 years)

Appendix 2 Questionnaire scenarios

Scenario with XAI:

Harry is working at the office and receives the following email:



Subject: Important: Action Required to Secure Your LinkedIn Account
From: Ryan Roslansky (ryan.roslansky@linkedin-support.com)
To: Harry.Cooper@Gmail.com
Date: 28-10-2025

Dear Harry Cooper,

We are reaching out to inform you of a new security update on LinkedIn. As part of our commitment to protecting user information, we require all users to confirm their accounts to avoid potential risks.

Please verify your account by clicking the secure link below:

Secure My LinkedIn Account

This process must be completed within 48 hours, or your account may be temporarily suspended.

Thank you for your prompt attention.

Ryan Roslansky
CEO, LinkedIn
ryan.roslansky@linkedin.com
LinkedIn Corporation
1000 W Maude Ave, Sunnyvale, CA 94085, USA
Phone: (408) 555-1234

Harry uses an anti-phishing tool that uses XAI. While reading the email, Harry's cybersecurity anti-phishing tool gives the following warning:

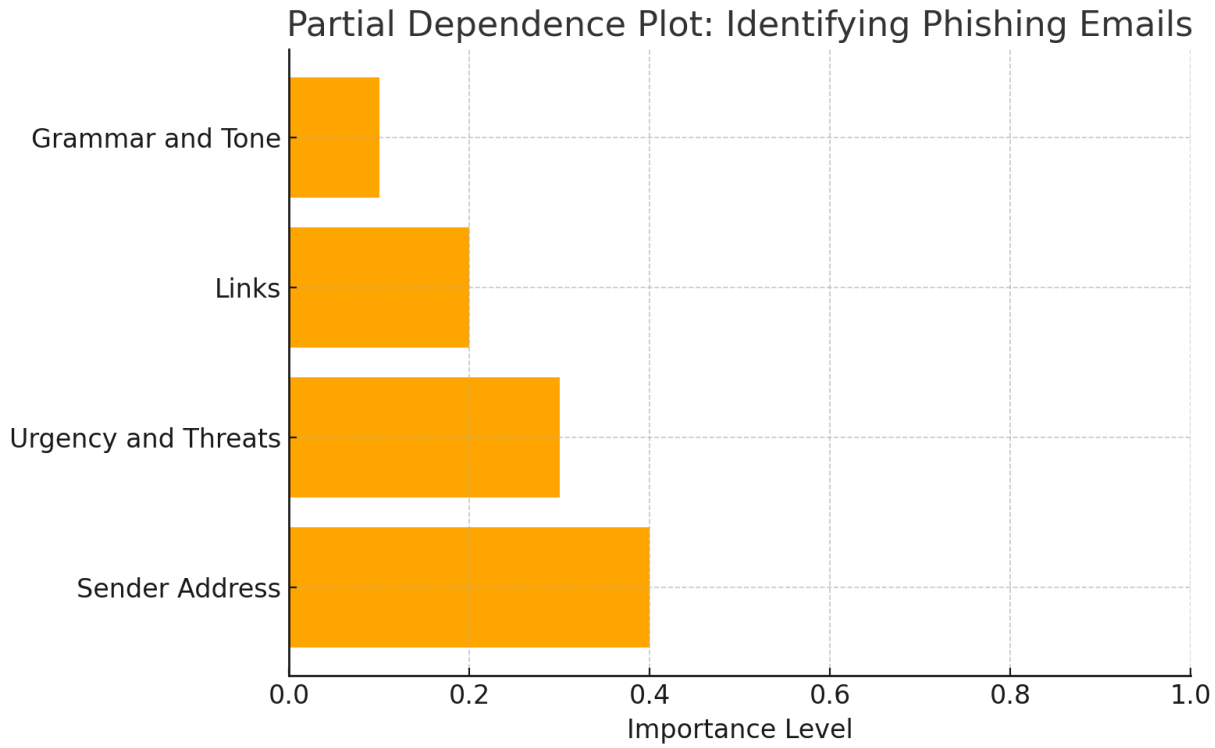
Warning: Phishing Email Detected!

Detection Summary:

- ***Sender: ryan.roslansky@linkedin-support.com***
- ***Subject: Important: Action Required to Secure Your LinkedIn Account***
- ***Date Received: October 28, 2024***

Risk Assessment: Our AI system has identified this email as a potential phishing attempt

based on several key indicators, including:



Recommended Actions:

• ***Do Not Interact:*** *Avoid clicking any links or downloading attachments from this email.*

• ***Report as Spam:*** *Mark the email as spam in your email client.*

• ***Delete the Email:*** *Remove it from your inbox to prevent accidental engagement.*

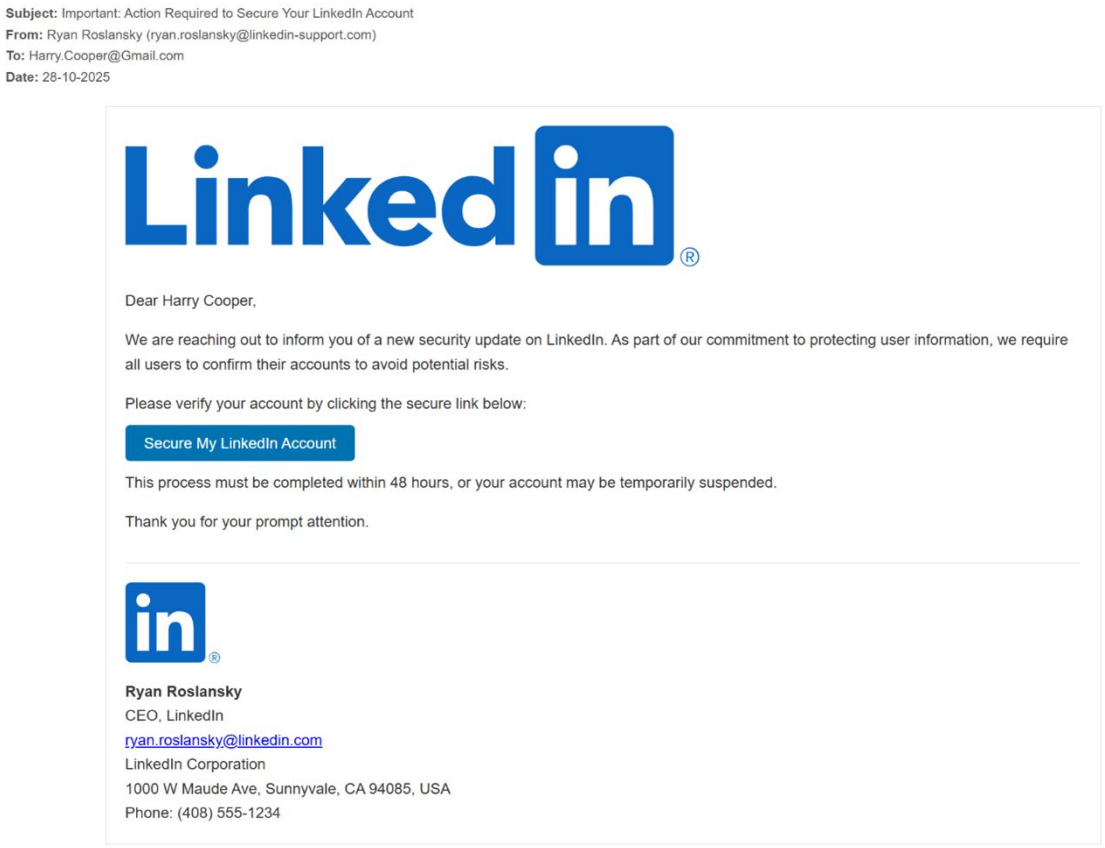
• ***Notify Your Bank:*** *If you have interacted with the email, contact your bank immediately for further guidance.*

Stay Safe Online! For more tips on recognizing phishing emails, please visit our Phishing Awareness Center.

The following statements are to provide insight into the trust of this anti-phishing tool. Please answer them based on the scenario above, and to the best of your ability.

Scenario Without XAI:

Harry is working at the office and receives the following email:



Subject: Important: Action Required to Secure Your LinkedIn Account
From: Ryan Roslansky (ryan.roslansky@linkedin-support.com)
To: Harry.Cooper@Gmail.com
Date: 28-10-2025

Dear Harry Cooper,

We are reaching out to inform you of a new security update on LinkedIn. As part of our commitment to protecting user information, we require all users to confirm their accounts to avoid potential risks.

Please verify your account by clicking the secure link below:

Secure My LinkedIn Account

This process must be completed within 48 hours, or your account may be temporarily suspended.

Thank you for your prompt attention.

Ryan Roslansky
CEO, LinkedIn
ryan.roslansky@linkedin.com
LinkedIn Corporation
1000 W Maude Ave, Sunnyvale, CA 94085, USA
Phone: (408) 555-1234

Harry uses an anti-phishing tool that uses AI. While reading the email, Harry's cybersecurity anti-phishing tool gives the following warning:

Warning: Phishing Email Detected!

Recommended Actions:

• ***Do Not Interact:*** *Avoid clicking any links or downloading attachments from this email.*

• ***Report as Spam:*** *Mark the email as spam in your email client.*

• ***Delete the Email:*** *Remove it from your inbox to prevent accidental engagement.*

• ***Notify Your Bank:*** *If you have interacted with the email, contact your bank immediately for further guidance.*

Stay Safe Online! For more tips on recognizing phishing emails, please visit our Phishing Awareness Center.

The following statements are to provide insight into the trust of this anti-phishing tool. Please answer them based on the scenario above, and to the best of your ability.

Appendix 3 Construct validity SPSS results

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,702
Bartlett's Test of Sphericity	Approx. Chi-Square	932,776
	df	276
	Sig.	<,001

Communalities

	Initial	Extraction
The Cybersecurity tool is deceptive not at all=1 extremely=7	1,000	,692
The cybersecurity tool behaves in an underhanded matter. not at all=1 extremely=7	1,000	,636
I am suspicious of the cybersecurity tool's intent, action, or outputs. not at all=1 extremely=7	1,000	,705
I am wary of the cybersecurity tool. not at all=1 extremely=7	1,000	,413
The cybersecurity tool's actions will have a harmful or injurious outcome. not at all=1 extremely=7	1,000	,532
I am confident in the cybersecurity tool. not at all=1 extremely=7	1,000	,660

The cybersecurity tool provides security. not at all=1 extremely=7	1,000	,507
The cybersecurity tool has integrity. not at all=1 extremely=7	1,000	,442
The cybersecurity tool is dependable. not at all=1 extremely=7	1,000	,394
The cybersecurity tool is reliable. not at all=1 extremely=7	1,000	,507
I can trust the cybersecurity tool. not at all=1 extremely=7	1,000	,661
Using this cybersecurity tool would enable me to make decisions more quickly. not at all=1 extremely=7	1,000	,739
Using the cybersecurity tool would help me make better decisions. not at all=1 extremely=7	1,000	,840
Using the cybersecurity tool would enhance the effectiveness of my decisions. not at all=1 extremely=7	1,000	,817

Using the cybersecurity tool would make it easier to make decisions. not at all=1 extremely=7	1,000	,779
I would find the cybersecurity tool useful. not at all=1 extremely=7	1,000	,734
Learning to operate the cybersecurity tool would be easy for me. not at all=1 extremely=7	1,000	,439
I would find it easy to get the cybersecurity tool to do what I want it to do. not at all=1 extremely=7	1,000	,602
My interaction with the cybersecurity tool would be clear and understandable. not at all=1 extremely=7	1,000	,511
I would find the cybersecurity tool to be flexible to interact with. not at all=1 extremely=7	1,000	,710
It would be easy for me to become skillful at using the cybersecurity tool. not at all=1 extremely=7	1,000	,555
I would find the cybersecurity tool easy to use. not at all=1 extremely=7	1,000	,616

I am familiar with the cybersecurity tool. not at all=1 extremely=7	1,000	,473
Using the cybersecurity tool would increase my productivity. not at all=1 extremely=7	1,000	,604

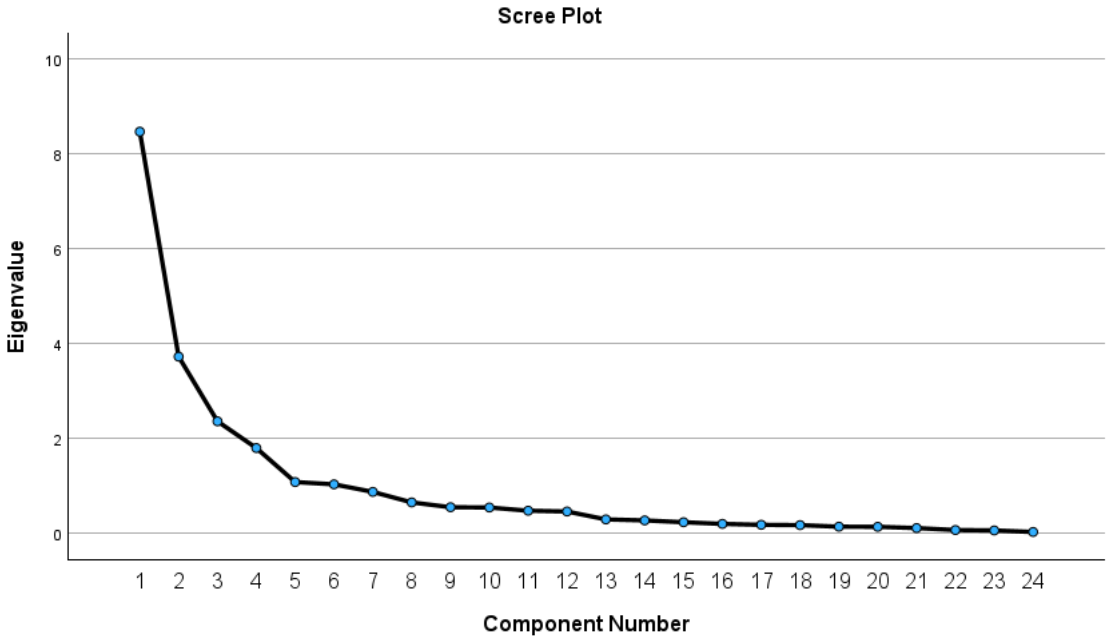
Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8,473	35,303	35,303	8,473	35,303	35,303	6,560	27,333	27,333
2	3,729	15,536	50,840	3,729	15,536	50,840	4,290	17,873	45,206
3	2,365	9,854	60,694	2,365	9,854	60,694	3,717	15,488	60,694
4	1,801	7,506	68,201						
5	1,084	4,518	72,719						
6	1,036	4,317	77,036						
7	,876	3,652	80,688						
8	,655	2,730	83,418						
9	,554	2,307	85,725						
10	,548	2,281	88,006						
11	,480	1,999	90,006						
12	,463	1,931	91,936						
13	,297	1,236	93,172						
14	,275	1,148	94,320						
15	,237	,988	95,308						
16	,202	,842	96,149						

17	,182	,760	96,910						
18	,176	,733	97,642						
19	,143	,597	98,239						
20	,140	,585	98,824						
21	,114	,475	99,299						
22	,073	,303	99,602						
23	,064	,267	99,869						
24	,032	,131	100,000						

Extraction Method: Principal Component Analysis.



Rotated Component Matrix[a]

	Component		
	1	2	3
I would find the cybersecurity tool to be flexible to interact with. not at all=1 extremely=7	,786		
I would find it easy to get the cybersecurity tool to do what I want it to do. not at all=1 extremely=7	,773		
I can trust the cybersecurity tool. not at all=1 extremely=7	,770		
I would find the cybersecurity tool easy to use. not at all=1 extremely=7	,756		

I am confident in the cybersecurity tool. not at all=1 extremely=7	,733		
It would be easy for me to become skillful at using the cybersecurity tool. not at all=1 extremely=7	,730		
My interaction with the cybersecurity tool would be clear and understandable. not at all=1 extremely=7	,689		
I would find the cybersecurity tool useful. not at all=1 extremely=7	,661	,520	
The cybersecurity tool provides security. not at all=1 extremely=7	,641		
The cybersecurity tool has integrity. not at all=1 extremely=7	,637		
The cybersecurity tool is reliable. not at all=1 extremely=7	,633		
The cybersecurity tool is dependable. not at all=1 extremely=7	,548		

Learning to operate the cybersecurity tool would be easy for me. not at all=1 extremely=7	,521		,404
Using the cybersecurity tool would enhance the effectiveness of my decisions. not at all=1 extremely=7		,861	
Using the cybersecurity tool would make it easier to make decisions. not at all=1 extremely=7		,853	
Using this cybersecurity tool would enable me to make decisions more quickly. not at all=1 extremely=7		,846	
Using the cybersecurity tool would help me make better decisions. not at all=1 extremely=7	,363	,841	
Using the cybersecurity tool would increase my productivity. not at all=1 extremely=7		,694	-,347
The Cybersecurity tool is deceptive not at all=1 extremely=7			,825
I am suspicious of the cybersecurity tool's intent, action, or outputs. not at all=1 extremely=7			,819

The cybersecurity tool behaves in an underhanded matter. not at all=1 extremely=7			,778
The cybersecurity tool's actions will have a harmful or injurious outcome. not at all=1 extremely=7			,728
I am wary of the cybersecurity tool. not at all=1 extremely=7			,585
I am familiar with the cybersecurity tool. not at all=1 extremely=7		,380	-,543

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Construct validity Trust

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,760
Bartlett's Test of Sphericity	Approx. Chi-Square	347,120
	df	66
	Sig.	<,001

Communalities

	Initial	Extraction
The Cybersecurity tool is deceptive not at all=1 extremely=7	1,000	,728
The cybersecurity tool behaves in an underhanded matter. not at all=1 extremely=7	1,000	,658
I am suspicious of the cybersecurity tool's intent, action, or outputs. not at all=1 extremely=7	1,000	,722
I am wary of the cybersecurity tool. not at all=1 extremely=7	1,000	,413
The cybersecurity tool's actions will have a harmful or injurious outcome. not at all=1 extremely=7	1,000	,539
The cybersecurity tool provides security. not at all=1 extremely=7	1,000	,579

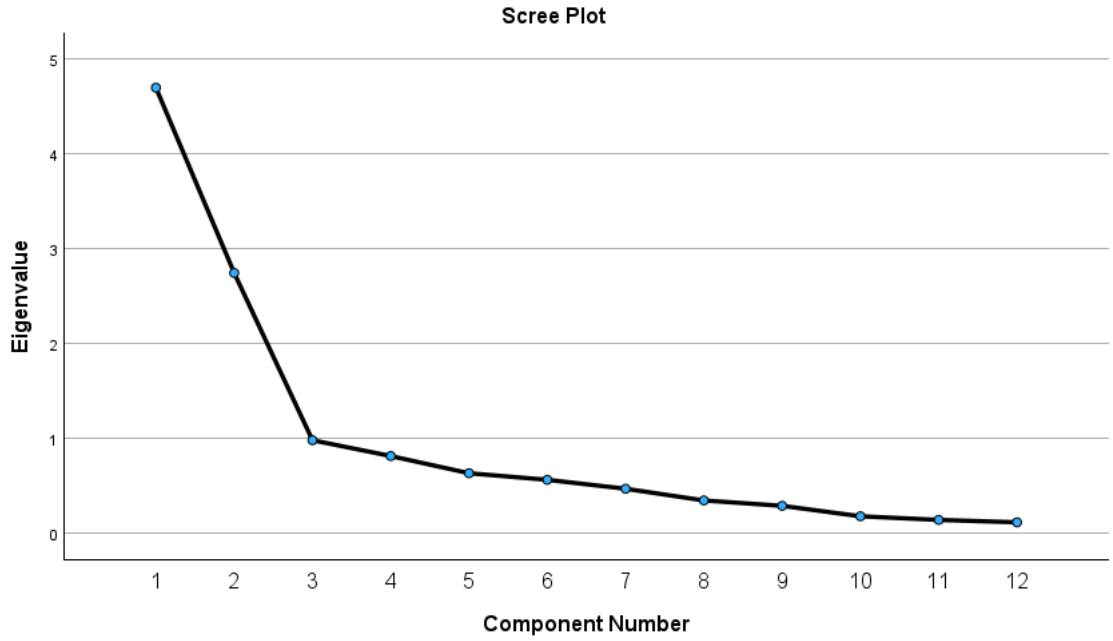
The cybersecurity tool has integrity. not at all=1 extremely=7	1,000	,585
The cybersecurity tool is dependable. not at all=1 extremely=7	1,000	,504
The cybersecurity tool is reliable. not at all=1 extremely=7	1,000	,749
I can trust the cybersecurity tool. not at all=1 extremely=7	1,000	,850
I am familiar with the cybersecurity tool. not at all=1 extremely=7	1,000	,346
I am confident in the cybersecurity tool. not at all=1 extremely=7	1,000	,774

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,700	39,169	39,169	4,700	39,169	39,169	4,186	34,882	34,882
2	2,747	22,889	62,058	2,747	22,889	62,058	3,261	27,177	62,058
3	,983	8,193	70,251						
4	,816	6,802	77,053						
5	,635	5,292	82,345						
6	,566	4,714	87,060						
7	,471	3,929	90,989						
8	,348	2,901	93,890						
9	,291	2,428	96,318						
10	,181	1,510	97,828						
11	,143	1,191	99,019						
12	,118	,981	100,000						

Extraction Method: Principal Component Analysis.



Rotated Component Matrix[a]

	Component	
	1	2
I can trust the cybersecurity tool. not at all=1 extremely=7	,912	
I am confident in the cybersecurity tool. not at all=1 extremely=7	,868	
The cybersecurity tool is reliable. not at all=1 extremely=7	,852	
The cybersecurity tool has integrity. not at all=1 extremely=7	,763	

The cybersecurity tool provides security. not at all=1 extremely=7	,754	
The cybersecurity tool is dependable. not at all=1 extremely=7	,709	
The Cybersecurity tool is deceptive not at all=1 extremely=7		,842
I am suspicious of the cybersecurity tool's intent, action, or outputs. not at all=1 extremely=7		,828
The cybersecurity tool behaves in an underhanded matter. not at all=1 extremely=7		,768
The cybersecurity tool's actions will have a harmful or injurious outcome. not at all=1 extremely=7		,732
I am wary of the cybersecurity tool. not at all=1 extremely=7		,612
I am familiar with the cybersecurity tool. not at all=1 extremely=7		-,540

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Construct validity technology acceptance

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,714
Bartlett's Test of Sphericity	Approx. Chi-Square	281,896
	df	55
	Sig.	<,001

Communalities

	Initial	Extraction
The Cybersecurity tool is deceptive not at all=1 extremely=7	1,000	,728
The cybersecurity tool behaves in an underhanded matter. not at all=1 extremely=7	1,000	,655
I am suspicious of the cybersecurity tool's intent, action, or outputs. not at all=1 extremely=7	1,000	,723
I am wary of the cybersecurity tool. not at all=1 extremely=7	1,000	,420
The cybersecurity tool's actions will have a harmful or injurious outcome. not at all=1 extremely=7	1,000	,534

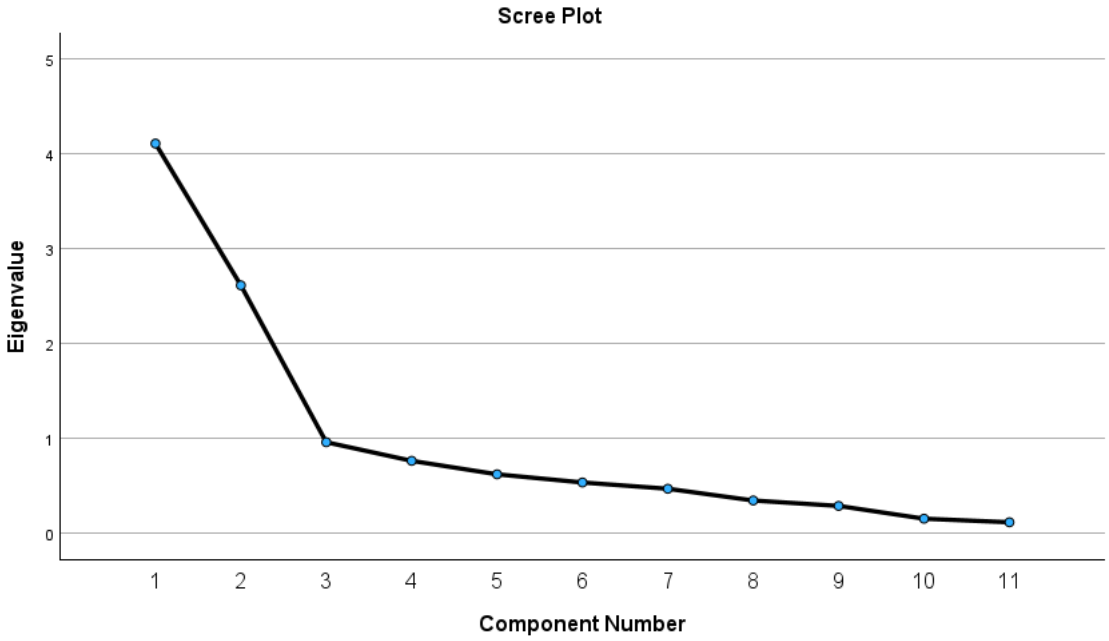
The cybersecurity tool provides security. not at all=1 extremely=7	1,000	,518
The cybersecurity tool has integrity. not at all=1 extremely=7	1,000	,633
The cybersecurity tool is dependable. not at all=1 extremely=7	1,000	,554
The cybersecurity tool is reliable. not at all=1 extremely=7	1,000	,755
I can trust the cybersecurity tool. not at all=1 extremely=7	1,000	,847
I am familiar with the cybersecurity tool. not at all=1 extremely=7	1,000	,360

Extraction Method: Principal Component Analysis.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,111	37,373	37,373	4,111	37,373	37,373	3,468	31,526	31,526
2	2,616	23,782	61,155	2,616	23,782	61,155	3,259	29,629	61,155
3	,962	8,747	69,902						
4	,766	6,962	76,864						
5	,624	5,671	82,535						
6	,538	4,888	87,422						
7	,471	4,286	91,708						
8	,348	3,161	94,869						
9	,290	2,641	97,510						
10	,156	1,416	98,926						
11	,118	1,074	100,000						

Extraction Method: Principal Component Analysis.



Rotated Component Matrix[a]

	Component	
	1	2
I can trust the cybersecurity tool. not at all=1 extremely=7	,909	
The cybersecurity tool is reliable. not at all=1 extremely=7	,854	
The cybersecurity tool has integrity. not at all=1 extremely=7	,794	
The cybersecurity tool is dependable. not at all=1 extremely=7	,743	

The cybersecurity tool provides security. not at all=1 extremely=7	,711	
The Cybersecurity tool is deceptive not at all=1 extremely=7		,844
I am suspicious of the cybersecurity tool's intent, action, or outputs. not at all=1 extremely=7		,830
The cybersecurity tool behaves in an underhanded matter. not at all=1 extremely=7		,773
The cybersecurity tool's actions will have a harmful or injurious outcome. not at all=1 extremely=7		,728
I am wary of the cybersecurity tool. not at all=1 extremely=7		,612
I am familiar with the cybersecurity tool. not at all=1 extremely=7		-,539

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.