Analyzing Trust in AI: Factors in Safety-critical and non-safety-critical Situations

AI systems are complex and difficult to understand, which makes it essential to establish a basis of trust between the user and the system to enable comfortable interactions. AI systems present both opportunities and risks, hence developers should endeavor to minimize risks as much as possible. Users, particularly in the realm of safety-critical systems, require sufficient trust in any new technology to consider its usage. Uncertainty and missing understanding about the field of AI and algorithms in general hinders and stops its usage. Therefore, transparency and comprehensibility should be goals in the development of AI systems. These factors can be divided into three groups: Technical factors, social factors, and evaluation factors. Each of these groups consists of five factors that can either enhance or diminish trust in an AI system. Based on these factors, a reference model is created to classify these three groups of factors within the overall context of trust formation. A study was conducted with 258 participants using a questionnaire to examine and subsequently analyze and evaluate the previously formulated reference model. An attempt is made to hierarchize the individual factors within the three groups and to identify any differences related to safety-critical and non-safety-critical systems.

Keywords: a*rtificial intelligence, trustworthy ai, responsible ai*

Introduction

Machine learning, particularly deep neural networks, are often challenging to comprehend. A known input value is transformed into a potential output value predetermined by the network developer's design decisions which control the complex computations within the network. The structure of the network, with its various layers, is specified by the developer. However, the precise calculations and intermediate values that an input value undergoes on its path to becoming an output value are unknown. Even developers of highly complex neural networks typically only know what type of operations are performed in each respective layer. These networks are referred to as black boxes, as the exact behaviors of the neural network cannot be determined or described. Consequently, developers of such neural networks must focus particularly on input data and its structure. Input data requires special attention, as the exact processing within the network is unknown. For instance, it is challenging to identify sample biases within the input data that may render the results of the neural network unusable. Such biases could occur due to unnoticed characteristics within the data.

Users of AI systems thus desire transparency and a certain degree of comprehensibility, especially when employing machine learning algorithms in safety-critical application domains, such as medical technology or autonomous vehicles. Physicians receiving a diagnosis from an algorithm can better assess the accuracy of the diagnosis if they have background information about the algorithm's decision-making process. Furthermore, a more thorough understanding of a black

box can aid in error corrections or the debugging process. People typically fear the unknown, and similar sentiments apply to trusting something opaque (Carleton, 2016). Explainable artificial intelligence approaches aim to counter this issue, as more and more systems are equipped with AI functionalities, and trust in AI systems should be strengthened. Trust plays a significant role in the utilization of mechanisms and devices whose functionalities cannot be fully understood by its users. This trust can be gained through test results or personal experiences. Whether it concerns the use of a technical device or trust in another individual, it is more about an inner conviction, typically based on positive experiences. Thus, the concept of trust can always be associated with a dynamic relationship, which can be established with both humans and objects.

According to Luhmann, trust can be used to reduce complexity in modern societies. Uncertainties and dangers can be weakened. Applied to AI systems, this means that a certain level of trust is required to interact meaningfully with a complex system. The lack of information about the system must be compensated for and bridged with the help of trust (Luhmann, 1968). In interpersonal relationships, trust plays a crucial role. This trust must be built over an extended period and requires constant attention. Trust in another person can quickly be compromised. When using algorithms, it is expected that their use should occur directly and without a prolonged period of trial and adjustments. Experiences are typically relayed second-hand, through consultations, or from specialized personnel. Trust can be challenging to establish under such conditions. Personal experiences can only be gained through active use, which is not an effective way to build trust, particularly with novel and safety-critical systems. Even with non-critical systems, such as using a chatbot, users often lack the necessary information to understand how the system generates a response.

Literature Review

In terms of trust in AI systems, there are already numerous research papers and scientific articles. The European Commission sets an important benchmark for the development of AI systems in Europe with their "Ethics Guidelines for Trustworthy AI" (European Commission, 2019). Accordingly, trustworthy AI is characterized by seven components. The expert group identifies human agency, robustness, privacy, transparency, ethical behavior, sustainability, and accountability as the most important components in the development of trustworthy AI systems. The expert group of the European Commission focuses on the conception of AI systems and describes factors that form a basis for trustworthy AI systems. Without these fundamental factors, trust in AI systems is not possible. Ideally, all three factors should apply to an AI system for it to be considered trustworthy. But other factors such as explainability, as well as security, controllability, reliability, and experience, should not be underestimated (Jacovi, 2021). These factors are equally important when it comes to creating genuine trust in an AI system. The more positive factors associated with a system, the easier it is to build trust in that system.

In view of this multitude of factors, this work divides these factors into two groups. On the one hand, there are technical factors, and on the other hand, there are social factors. Technical factors are not directly transferable to a trust relationship between two individuals, whereas this is the case for social factors. Hoff et al. write that relationships between humans and machines as well as between two humans are different but fundamentally based on the same principles (Hoff, 2015). Taking a closer look at the factors of an interpersonal relationship, factors such as reliability or traceability emerge, which can also be transferred to trust in an AI system. The foundation of trust outlined by the expert group of the European Commission can thus be compared to a first impression of another person. Trust itself can only be built and strengthened as the relationship progresses, offering much scope for considering the additional factors for a more detailed analysis. Factors such as transparency and explainability play a prominent role here, while factors such as reliability or security are also important but may be weighed differently by various users.

Parasuraman et al. mention an interesting observation here. It was found that people continue to use a system after an error has occurred, even without understanding the error (Parasuraman, 1997). Such behavior would be atypical in an interpersonal relationship and is likely more common in the use of computer systems. Without an explanation of why something did not work, few people would probably trust another person again. Parasuraman et al. suggest a dependency between humans and machines here, as, for example, an activity may not be possible without the assistance of the machine. If such a dependency exists, people are also willing to continue working with a system even in the case of serious errors. Consequently, trust can be replaced by dependencies. However, according to Ferronato et al., a lack of trust and forced interaction with a system, such as in a work context, lead to a reduction in effectiveness and safety (Ferronato, 2020). In the context of interpersonal relationships, such dependence could be compared to coercion and thus does not reflect a desirable state. Trust should therefore always be an important goal in the design of AI systems, while efforts should be made to avoid usage due to dependency.

Ferronato et al. also address another interesting aspect. Trust itself can be further subdivided into "Primary Trust" and "Secondary Trust" (Ferronato, 2020). "Primary Trust" refers to the initial contact with something or someone, the first impression. "Secondary Trust", on the other hand, entails the trust that develops over time. It has been observed that the first impression has a very strong influence on "Secondary Trust". Furthermore, personal predisposition, which varies from person to person, plays a crucial role in the formation of "Primary Trust". Factors such as prior experiences, experiences of third parties, or the reputation of the system are important here (Ferronato, 2020). Trust in an AI system is therefore significantly based on the first impression, which is also strongly influenced by the fundamental attitude towards AI systems in general. Thus, one could speak of a "three-step process" (fundamental attitude, first impression, usage) of trust, as usage is highly unlikely to occur in safety-critical domains with a negative fundamental attitude or a negative first impression.

Additional characteristics of trust were discovered by Alarcon et al. They report on the dangers of "over-trusted" and "under-trusted" systems (Alarcon, 2020). If a system is trusted too much, the capabilities and functionalities of a system are overestimated. This leads to the danger that users interact with a system too casually and, for example, do not perform quality checks on the results of an AI system. It is assumed here that the system will provide correct results anyway. Conversely, if a system is trusted too little, functionalities of a system are not utilized because it is assumed from the outset that the system will not provide useful results (Alarcon, 2020). Both scenarios are suboptimal as they result in the underutilization of a system's full potential. Establishing trust in a system is crucial, but it's equally important not to overestimate its capabilities. Consequently, developers of AI systems should incorporate trust considerations into the development process. Trust is elusive and demands a considerable degree of attention and deliberation.

But how does trust arise and how can a trusting relationship be described? Hancock et al. write that the principle of trust is always based on three elements. These include two agents and a communication channel. One of the agents is the sender, and the other is the receiver (Hancock, 2011). The sender is the agent who transmits trust, that is, who trusts the other agent. The receiver is trusted by the sender; thus, the receiver receives trust through the communication channel. In the context of AI systems, the communication channel can be referred to as the operational environment. This could be one's own home computer for a software application or one's own car for a hardware application. Furthermore, Hancock et al. write that trust is a dynamic variable (dimension) that becomes increasingly important with increasing autonomy (Hancock, 2011). The concept of a humanoid robot, as depicted in science fiction films, would represent one possible stage of this dimension, as trust in such a robot would be nearly equivalent to that in another human. However, all intermediate stages and developments of current AI systems need to be considered in a more nuanced manner. They are gradually approaching a classical trust relationship between two humans as AI systems become more advanced and autonomous in decision-making.

In reference to the previously mentioned image of two agents, Devitt writes that the sender always makes themselves vulnerable, as they trust that the receiver will not exploit them (Devit, 2018). This fundamental fear applies only partially to AI systems, as it involves the "user - application" situation. The user tends to trust that the system will perform a task correctly rather than trusting that the system will perform any task at all. Therefore, in the context of AI systems, special attention must be paid to the level of functional trust. Functional trust describes the trust in a system's ability to successfully execute a specific task. Consequently, the focus in further examination is placed more on the topic of functional trust, while interpersonal factors are considered only marginally. Hancock et al. address an important point here. They highlight the role of "false trust" in the development of trustworthy AI systems. False trust involves the concept of deception. Through deception, a system may appear trustworthy or successful in its actions when it is not (Hancock, 2011). Such a system could, for example, secretly collect data, and the expected task is merely a secondary activity to appear trustworthy. If this fact were to be revealed, no one would knowingly interact with the system anymore

because without "false trust", it would not enjoy genuine trust. However, there is also a danger here if system functions offer significant advantages or substantial workflow simplification. For example, it might be possible for AI systems to be used, especially in non-critical areas, without being trusted to a high degree. Devitt mentions companies as an example of the relevance of this consideration, which do not prioritize data protection and share personal data to increase their own revenue (Devit, 2018). In this regard, there are interesting questions for further investigation. How significant must the functional advantage be for a person to interact with a non-trustworthy AI system? Functional trust plays a crucial role in AI systems and can be enhanced through transparency and explainability of algorithms. This is important, because the potential underuse of AI technology could be one logical consequence of the fear of misuse (Floridi, 2018).

To better understand the concept of deception in relation to trust, it is important to consider the opposite of trust. After a brief consideration, one may think of the term mistrust. But is mistrust truly the opposite of trust? Smithson writes that a more nuanced examination is required to answer this question (Smithson, 2018). According to Smithson, mistrust and the absence of trust are not synonymous. In the case of mistrust, clear doubts about trustworthiness can be identified based on information. In the case of the absence of trust, there is not enough information available to make a definitive decision about whether the system is trustworthy or not. However, the absence of trust is clearly the scenario that should primarily be considered when developing new AI systems. Mistrust is too strong a term when it is a matter of the absence of trust. Deception and mistrust should therefore be identified as extreme values.

In considering functional trust, the focus is on establishing trust in the functionalities of an AI system. Individuals lacking an understanding of a system's functions must be able to place their trust in the system's functionalities to utilize it efficiently. The process of building trust is therefore essential for the utilization of a system. As this process varies from system to system, preconceptions should be made during the development process.

The Human Black Box

The preceding examination of trust between humans and AI systems raises some questions regarding the human decision-making process. When one person trusts another person, they are also trusting a black box. When a person interacts with an AI system, they desire transparency and an understanding of its behavior. This paradox has also been addressed and analyzed by Bonezzi et al. They discuss the illusion that the decision-making process of a human is more transparent than that of an AI system (Bonezzi, 2022). A person is more likely to trust another person because they see themselves as a template and can more easily transfer their own decision-making process onto other people. An AI algorithm is foreign, and one's own decision-making process cannot be readily transferred onto an algorithm.

Many regulations are being introduced to demand transparency of algorithms, such as those by the European Union, to provide people with a better understanding of AI systems (Goodman, 2016). The intentions behind these mechanisms are clear.

However, daily, humans interact with a multitude of black boxes that are not required to openly disclose why they behave as they do. Examples of this include doctors (Bonezzi, 2022). Diagnoses are made without further explanation to a patient. It is trusted that a physician makes a correct diagnosis, but misdiagnoses occur. However, there are no initiatives to make human decisions explainable and transparent. The decisions of a taxi driver are as much a black box as those of an autonomous vehicle. Trust is established through laws, rules, and the transferability of one's own decision-making process to the taxi driver, even though no trust relationship has been built with an individual. This concept of trustworthiness is an important dimension when dealing with an uncertain future (Kusche, 2023). Kusche describes it as bridging the distinction between decision-makers and those affected. An AI system does not make conscious decisions and is therefore a foreign entity to many people when considering trusting something rather than someone.

Bonezzi et al. have found that people tend to empathize with other people, while this is not the case with AI systems. When people are closer to each other, this behavior occurs more frequently, and trust can be built more easily (Bonezzi, 2022). People belonging to a cohesive group are more likely to trust each other than those without close contact. However, in both cases, trust is placed in a black box, whether it involves friends or strangers. The human being uses themselves as a template to trust another human. The question that arises is: If an algorithm, which is a white box, could be used as a template to illustrate the decision-making process of another algorithm, which is a black box, would a person trust this algorithm?

To answer this question, three types of factors are distinguished: technical, social, and evaluation factors. Technical factors and social factors are no new concept, other works have already established similar kinds of factors. Ikkatai et al. have created an octagon metric with eight important factors for formulating trust (Ikkatai, 2022). Kaplan et al. and Bach et al. also worked on the identification process for factors in which trust can be measured (Kaplan, 2021; Bach, 2022). The evaluation factors are an additional set of factors that this work introduces to better understand outside influences that can determine the process of forming trust.

Methodology

The following section will focus on the factors influencing trust. These factors were decided upon reviewing other works in this field. Additionally, a set of nine test runs were conducted before the actual study took place to validate the approach. This led to the sub factors in the next part for generating a better understanding of main factors for participants of the study.

Regarding the technical factors, the AI system itself is considered. This includes stability, applicability, behavior, comprehensibility, and accuracy. These factors consist of several sub-factors. Technical factors are distinguished from social and evaluation factors in that they cannot be easily transferred to a trust relationship between two people. For instance, a person's behavior might be attributed to their political or religious beliefs. Since technical factors are intended to form the basis

for social factors, it was decided to include data protection and ethical correctness in the category of technical factors.

- Stability (Robustness/Responsiveness)
- Applicability (Usability/Difficulty)
- Behavior (Privacy/Ethics)
- Comprehensibility (Transparency/Explainability)
- Accuracy (Results/Reliability)

The second group of factors comprises social factors. These are characterized by their resemblance to elements found in human trust relationships. They are easier to transfer than technical factors and consider the AI system as an entity worthy of trust. The technical implementation is consequently functional and instills basic trust in the form of the factors already presented. Social factors include traceability, reliability, accountability, overall competence, and transferability. Transferability refers to the ability to transfer one's own values onto the AI system.

- Traceability
- Reliability
- Accountability
- Competence
- Transferability

The third group of factors comprises evaluation factors. These factors encompass the overarching theme of AI systems and societal perception. They examine reputation as well as experiential values. This specific group of factors aims to illustrate the influence of these values on trust in AI systems. It considers the media and popular culture reputation as well as the reputation of the developer or manufacturer. Furthermore, it investigates whether reviews, experiences with other AI systems, or the experiences of other individuals impact trust in AI systems.

- Reputation (Media/Popular culture)
- Reputation (Developer/Manufacturer)
- Experiences (Other AI systems)
- Experiences (Other people)
- Testing reports (Magazines/Consultations)

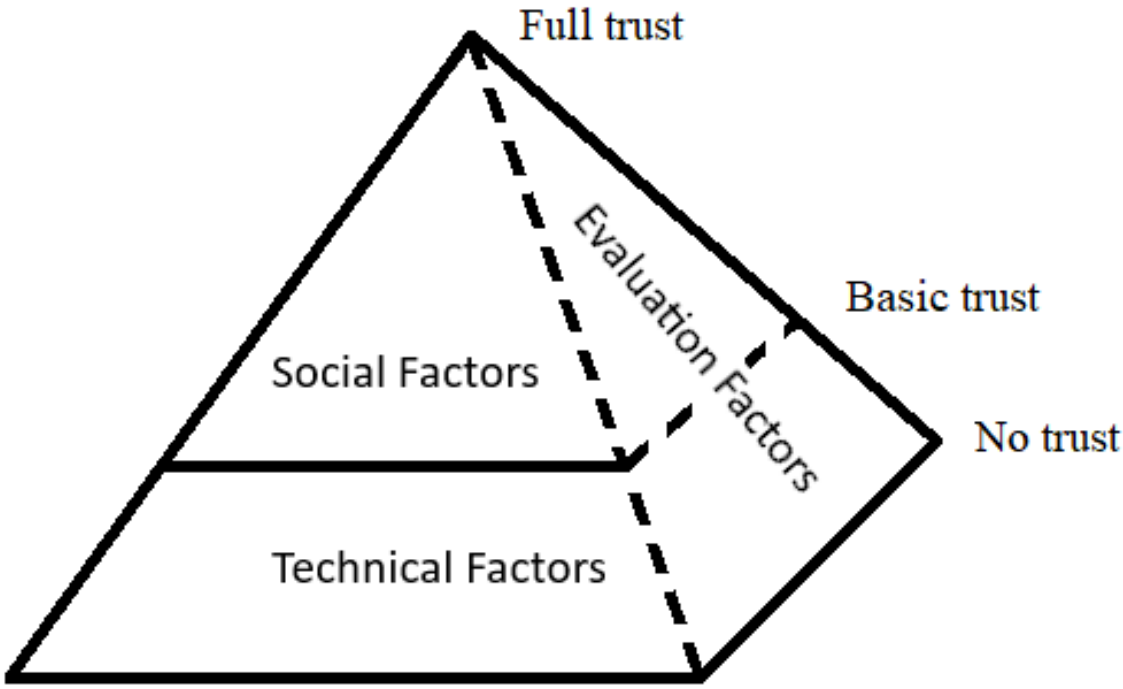
Introducing the Trust Pyramid as a Hierarchy

A study was conducted to evaluate the aforementioned factors and find answers to the following two questions:

1. What factors hold the highest priority when it comes to fostering trust in an AI system?
2. Does the context in which an AI system makes its decision play a crucial role?

The first question concerns the factors or groups of factors that have already been identified, while the second question relates to the context of use, with a focus on security aspects. Therefore, when both questions are combined, two scenarios are emphasized: one involving an AI system in a non-safety-critical context and the other involving an AI system in a safety-critical context. Figure 1 illustrates the trust pyramid that has been developed, which relates the trust factors to each other and attempts to classify the factors in terms of overall trust in any given system.

Figure 1. *Trust Pyramid - Based on own representation*



The trust pyramid always refers to a single separately considered AI system from the perspective of a singular individual. The basis on which trust is forming is different for everyone. Because of this the process of trusting a system or a person is not generalized by this approach. This work assumes a personal fundamental attitude of each human being. This varies from person to person and is based on experiences, one's own knowledge base, and personal worldview. If this fundamental attitude is not yet developed or if someone has no idea what an AI system is, it's a different process then for someone who has broad experiences with different AI systems. At the base of the pyramid, someone has no trust in a specific AI system. This is generally the case with any AI system, as there has been no engagement with it yet.

The bottom layer of the actual pyramid represents the technical factors. Here, a foundational trust is established based on stability, behavior, comprehensibility, accuracy of results, and usability. Without these factors, social factors don't play any role in generating trust, as trust in an AI system is not possible without trust in the technology. A broken system that does not work correctly or does not produce

results will have a hard time generating trust. The same is the case for any given system that can't be understood or behaves in a way that is not making sense to its users. Therefore, this work assumes a dependency between these two groups of factors. The dashed line on the right side of the pyramid represents an overlap with the domain of the evaluation factors. It is assumed that even if a system is technically trustworthy, a significant portion of the foundational trust depends on trust in the developer or manufacturer. The reputation of a company or the political stance of a CEO can also have an impact. These factors are typically independent of the software or hardware components.

The social factors and evaluation factors fundamentally build upon the technical factors. Social factors such as transparency, reliability, accountability, competence, and transferability should be considered after establishing the technical framework. If, in addition to the technical factors, the social factors can also be fulfilled, strong trust in an AI system can be inferred. The left side of the pyramid composed of technical and social factors represents the maximum trust that can be achieved from an AI system. The right side of the pyramid is only attainable through additional trust in the evaluation factors. However, these factors may deviate even in a perfect system. For example, if negative news about AI systems is currently prevalent in the media, this perception may be transferred to other similar systems. Thus, complete trust is anchored at the absolute pinnacle of the trust pyramid and requires, in addition to a trustworthy AI system, a trustworthy developer/ manufacturer and a generally positive perception of AI systems in the media or in the everyday environment.

Study Design

For the measurement within this study, a questionnaire is utilized. This questionnaire was developed based on the "OECD Guidelines on Measuring Trust" (OECD, 2017). As the conducted study is an online study, it mainly adhered to structural guidelines. Furthermore, no interviews were conducted. The study was held during a 46-day period with 258 active participants. The participants had to answer a questionnaire with questions related to different types of AI systems. Each question of the given scenarios was related to one of the factors.

For the completion of the questionnaire, a time span of 10 to 15 minutes was allocated, which was confirmed to be suitable through feedback from the respondents. The time frame was determined through a pre-test of the questionnaire. This pre-test also ensured that no technical difficulties would arise during the completion of the questionnaire. It was established beforehand what constitutes as "Valid cases" and what can be disregarded during data analysis. Additionally, this ensured that no necessary questions could be skipped and that dependencies within the questionnaire functioned correctly. Similarly, before the actual survey commenced, it was ensured that the questions were posed in a clear and comprehensible manner.

During the survey period, the questionnaire was completed by 294 individuals. However, 19 of these responses had to be excluded due to incomplete data. Consequently, 275 questionnaires were fully completed. Out of the 275 completed questionnaires, a further 17 were subsequently excluded due to erroneous responses

in the control questions. Thus, 258 datasets are relevant for the subsequent analysis of the results. 138 male, 118 female, 1 diverse and 1 participant without gender information were recorded. One objective of the study was to identify the factors that play the most significant role in fostering trust in an AI system.

Results

A preliminary conclusion can be drawn regarding the initial question of this study. The objective of the question was to identify the factors that play the most significant role in fostering trust in an AI system. Concerning the technical factors, the factors of stability, applicability, and accuracy must be mentioned. These were identified in the study as the most important technical factors. For the social factors, also three factors can be identified: reliability, accountability, and competence. The factor of traceability can be somewhat related to the technical factor of comprehensibility. It has been highlighted that a complete understanding is not always necessary when using a system. The ability to review or question the decisions of an algorithm is sufficient for many participants. Furthermore, trust gained through other factors can be used to fill this "knowledge gap". The factor of transferability was challenging for many individuals to imagine. This was easier for questions regarding a chatbot compared to an autonomous vehicle. Specifically, this factor needs to be examined to determine when people desire human-likeness and when they do not.

In terms of the evaluation factors, the most significant factors that were identified were the reputation towards developers/manufacturers, the personal experience regarding AI systems, and professional test reports. While media and films/games can have an influence, objective, industry-specific articles, and opinions based on test reports have a higher significance when it comes to building trust. Similarly, the experiences of other individuals have a lesser impact compared to one's own experiences. The trust pyramid is intended to serve as the initial means of such a classification and to relate the three groups of factors to one another. In already mentioned works, all factors were considered equally. The study conducted as part of this work demonstrated that different systems require a varied weighting of individual factors within the factor groups. While stability and accuracy were more critical for autonomous vehicles, participants emphasized the behavior regarding privacy and ethics of the system concerning chatbots. Individuals are willing to disclose more data for safety-critical systems than for non-safety-critical systems. Conversely, safety-critical systems are expected to have a higher degree of accuracy and stability. These differences in expectations influence trust in a system and must therefore be given higher weight. A clear hierarchy of factors cannot therefore not be established without the context of the system in question.

The introduced groups of factors aim to provide a better description of AI systems and allow for the classification of an individual’s trust in specific aspects of a system. The introduction of hierarchized factors adds to the approaches of Ikkatai et al., Kaplan et al. and Bach et al. by incorporating a factor of external impact, which can have a significant influence on trust formation (Ikkatai, 2022; Kaplan,

2021; Bach, 2022). The technical and social factors can also be found in those works but are complemented by the evaluation factors to categorize multiple factors. This helps facilitating the classification of interconnected aspects.

Discussion

The second question of this study will now be addressed: "Does the context in which an AI system makes its decision play a crucial role?" An attempt will be made to find an answer to the issue just posed. If the context of an AI system does play a crucial role, then a clear sequence cannot be established across multiple systems. In the following, two different AI systems were compared. This involved distinguishing between a chatbot as a non-safety-critical system and an autonomous vehicle as a safety-critical system. To compare these two different scenarios, 17 questions from the previously presented factors were selected and posed again in relation to these two types of systems. Each question always refers to one of the three factor groups.

As mentioned in the previous section, the type of AI system likely plays a role in many factors. For example, differences in the complexity of tasks of an AI system or safety concerns can affect the creation of trust. An autonomous vehicle, for instance, affects the physical well-being of a user, while chatbots are currently deployed in less risky environments. Autonomous vehicles operate in an uncertain and unpredictable environment and therefore must solve very complex tasks accordingly. Furthermore, chatbots have been a more familiar technology for some time, especially after the introduction of ChatGPT 3 to the public. Autonomous vehicles, on the other hand, are a newer development and are not yet very prevalent in people's everyday lives. Media coverage and personal experiences are also possible criteria for differentiating the evaluation of these two systems.

To obtain a more precise picture of the situation, the same 258 participants were also asked about their trust in chatbots and autonomous vehicles. All questions were asked with a given scenario about the AI system. These questions were created with the OECD guidelines as a basis for measuring trust. In addition, the questions were also adjusted based on the test runs before the actual study took place.

1. For me to use the AI system, the system must be responsive.
2. For me to use the AI system, the system must be easy to use.
3. For me to use the AI system, the system must be limited to relevant data.
4. For me to use the AI system, the system must behave respectfully towards me.
5. For me to use the AI system, I must understand the system's procedures.
6. For me to use the AI system, the results must be precise and the predictions accurate.
7. For me to use the AI system, I must be able to inspect its decision-making process.
8. For me to use the AI system, the system must deliver reliable results.
9. For me to use the AI system, I must be able to assess the responsibilities.

10. For me to use the AI system, the system must competently complete assigned tasks.
11. For me to use the AI system, the system must behave similarly to a human.
12. For me to use the AI system, it must be recommended to me by other people.
13. For me to use the AI system, the system must be positively reported in the media.
14. For me to use the AI system, I must trust the developer/manufacturer.
15. For me to use the AI system, I must already trust other AI systems from the same developer/manufacturer.
16. For me to use the AI system, many people must use the system.
17. For me to use the AI system, test reports must be positive.

Several T-tests were used to determine whether there are significant differences between the two systems. For a more detailed representation of the differences, a significant level of 0.001 was chosen. Table 1 provides an overview of the pairs and whether the t-value is <0.001. Table 1 shows 17 pairs, each regarding one of the 15 factors listed under methodology. Two factors were split to get a better picture of the two systems in comparison. These factors are behavior and experience. The factor's behavior was split into privacy and ethical behavior, while the factor experience was split into own experiences and experiences of other people.

Table 1. *T-test for comparing safety-critical and non-safety-critical AI systems*

Pairs	T	One-Sided p	Two-Sided p	Factor
Pair 1	-10.340	<.001	<.001	Stability
Pair 2	-4.059	<.001	<.001	Applicability
Pair 3	1.540	.062	.125	Behavior (Privacy)
Pair 4	1.895	.030	.059	Behavior (Ethics)
Pair 5	-8.766	<.001	<.001	Comprehensebility
Pair 6	-7.625	<.001	<.001	Accuracy
Pair 7	-9.579	<.001	<.001	Traceability
Pair 8	-6.387	<.001	<.001	Reliability
Pair 9	-10.092	<.001	<.001	Accountability
Pair 10	-7.433	<.001	<.001	Competence
Pair 11	-1.840	.033	.067	Transferability
Pair 12	-8.560	<.001	<.001	Reputation (Media)
Pair 13	-10.259	<.001	<.001	Reputation (Developer)
Pair 14	-9.864	<.001	<.001	Experience (AI systems)
Pair 15	-7.611	<.001	<.001	Experience (Other)
Pair 16	-9.679	<.001	<.001	Experience (Mine)
Pair 17	-10.900	<.001	<.001	Testing reports

From this table, it becomes evident that there are significant differences between chatbots and autonomous vehicles in almost all factors. No differences were observed in the factors "Behavior (Privacy)", "Behavior (Ethics)", and "Transferability". For these three questions, it was important for the participating individuals that both systems operate on the same principles. Significant differences were detected for all other factors. Thus, it can be concluded that there is a varying

perception and evaluation of the aforementioned factors for most aspects. Additionally, the negative T-value indicates that all significantly different values regarding autonomous vehicles were rated more strongly. This confirms the previously stated assumption that a clear hierarchy between trust factors cannot be established independent of the AI system considered. Only the underlying AI system provides the basis for considering the factors and enables individual prioritization of each factor. Therefore, the trust pyramid can be used as a reference model to allow for a rough categorization of the three groups of factors.

Based on the results of the T-test, the second question can be conclusively answered with "Yes". The context in which an AI system makes its decisions plays a crucial role. Trust in a safety-critical system is rated higher than in a non-safety-critical system. Establishing a hierarchy of factors, as attempted at the end of the previous section, is thus not possible and must be done individually when considering a specific AI system. However, since the trust pyramid also includes human considerations, a different order of factors can be assumed between two different individuals for the same AI system. A specific examination of individual factor groups concerning more than two AI systems could provide an approximation of an order. This could identify types of AI systems that are like each other.

Differences between female and male participants

Due to the very similar number of female and male participants, differences between these two groups will also be addressed. In general, it can be observed that no significant differences were found between the two groups. To assess the differences, a separate t-test was conducted for each group on an individual question. Like the previous comparison between safety-critical and non-safety-critical AI systems, a significance level of 0.001 was chosen.

Figure 2. *Diagram showing the single difference for technical factors*

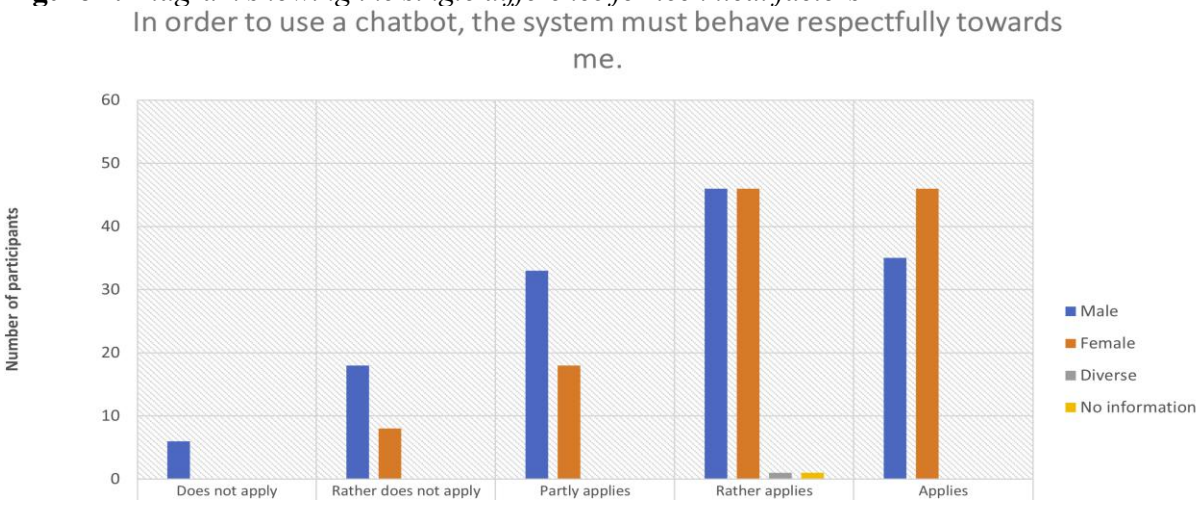
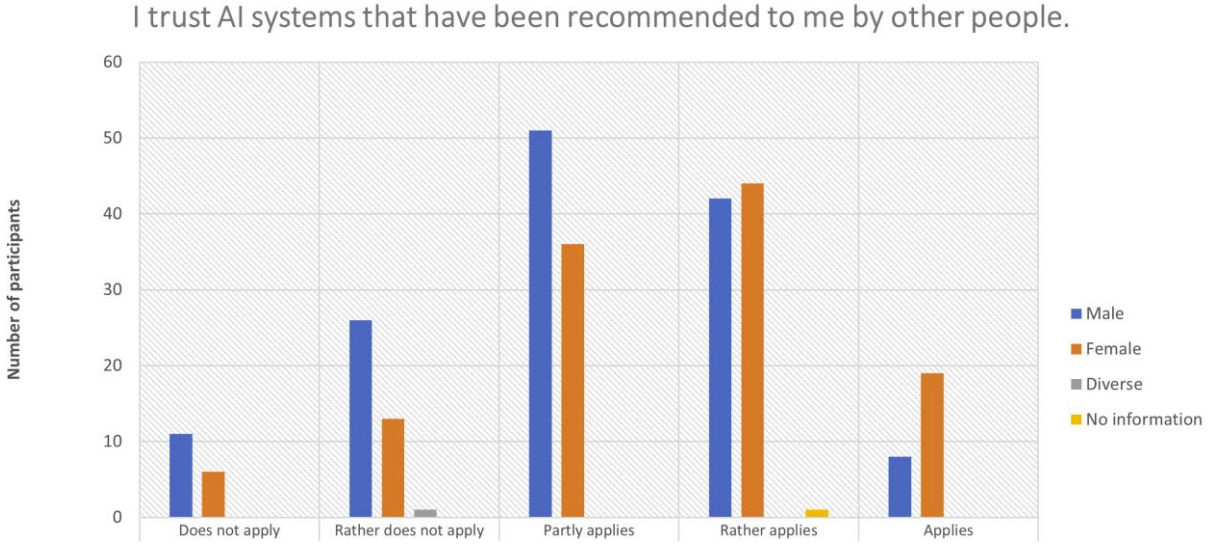


Figure 2 shows the only significant difference in the responses between men and women regarding the technical factors. This question concerned the rapid learnability of handling a system. While it was more important for female participants in the study that the handling of an AI system is quickly learnable, male participants indicated that it is not as important to them compared to females.

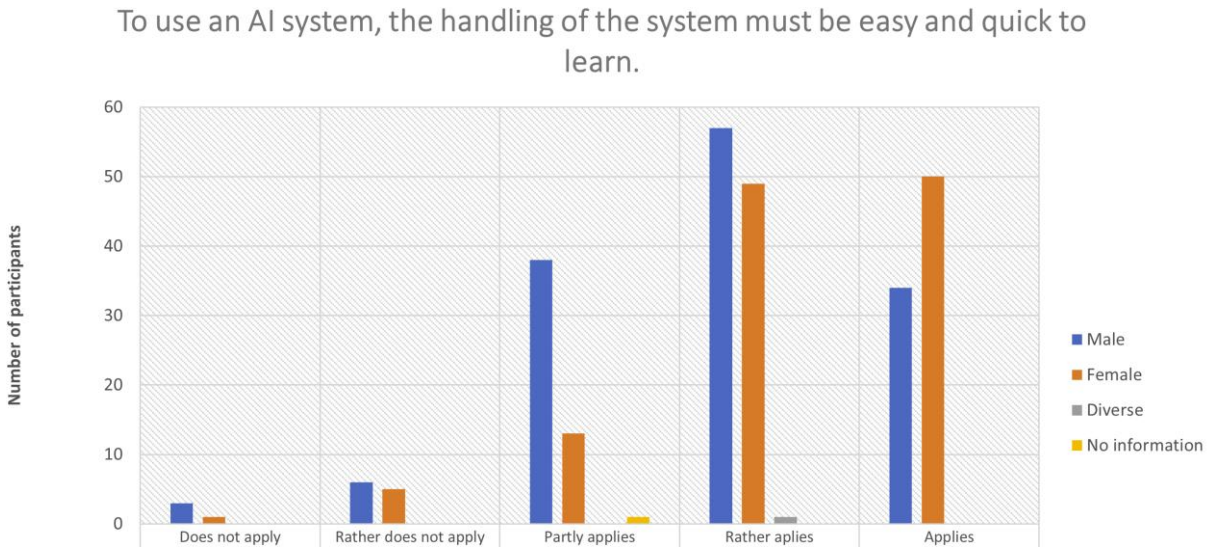
However, the other questions do not show significant differences between the two comparison groups, which is why this difference warrants special mention.

Figure 3. *Diagram showing the single difference for social factors*



Similarly, regarding the social factors, only one question exhibited a significant difference between the genders. This difference is illustrated in figure 3. The disparity between the two groups is evident in this question, which explored whether trust in an AI system was influenced when it was recommended by another person. Here, there was a stronger inclination towards reluctance among male participants compared to female participants. Women participating in the study were more inclined to consider others' trust as an important indicator for their own trust, whereas men regarded their own trust in an AI system as more crucial.

Figure 4. *Diagram showing the single difference for non-safety-critical AI systems*



Regarding questions related to safety-critical AI systems, no significant differences between men and women were observed for any of the questions. Similarly, for questions concerning non-safety-critical AI systems, only one question yielded a significant difference. This question addressed the respectful behavior of a chatbot towards its users. Here, female participants found it more important than male participants for the chatbot to behave respectfully. A larger proportion of male participants indicated that the chatbot's behavior towards the user was less important, as more individuals stated they would still use the chatbot. Figure 4 shows these observations. These slight discrepancies between the two groups suggest that gender does not have huge implications on responses regarding AI systems.

Differences between Trust in Humans and in AI Systems

At the beginning of the questionnaire, general questions about the trust of the study participants were recorded. These questions related to the previously mentioned questions about the social factors. They were derived from interpersonal trust between two individuals. The questions about interpersonal trust as well as the questions about social factors can accordingly be compared to determine a difference between trust in humans and AI systems. As with the comparisons between autonomous vehicles and chatbots previously, paired T-tests were conducted for both appropriate questions to identify significant differences between the responses. The following list shows the compared questions, as they differ slightly in formulation, on the left side of the hyphen, the reference is to humans, while on the right side, the reference is to AI systems. If the question is identical for humans and AI systems, it was listed only once. This is the case for the pairs six, eleven, and twelve.

1. whose actions I can comprehend. - whose algorithms I can comprehend.
2. who explains their behavior to me. - whose decision-making process I can perceive.
3. whom I have known for a long time. - with whom I have interacted frequently.
4. who have acted reliably before. - which deliver reliable results.
5. who are responsible for their actions. - whose responsibilities I can assess.
6. who proceeds with great accuracy and care.
7. whom I perceive as competent. - which competently accomplish assigned tasks.
8. who have previously executed a given task properly. - which have previously delivered correct results.
9. who are like me. - which behave similarly to humans.
10. in whom I can empathize. - which are like AI systems I trust.
11. who have been recommended to me by other individuals.
12. whom I have trusted in the past.

Table 2 provides an overview of the paired T-tests. From this table, it is evident that there are significant differences between trust in humans and AI systems. The analysis revealed that there are only four questions where no significant differences were found. It can be observed that no significant differences were found for pairs two, four, ten, and twelve. Consequently, eight questions, concerning social factors, differ in terms of trust between humans and AI systems. This indicates that a different type of trust can be assumed when a person interacts with another person than with an AI system.

Table 2. *T-test for comparing trust in humans and AI systems*

Pairs	T	One-Sided p	Two-Sided p
Pair 1	7.132	<.001	<.001
Pair 2	1.720	.043	.087
Pair 3	4.691	<.001	<.001
Pair 4	-2.808	.033	.005
Pair 5	-5.449	<.001	<.001
Pair 6	-5.966	<.001	<.001
Pair 7	-5.351	<.001	<.001
Pair 8	-8.565	<.001	<.001
Pair 9	10.310	<.001	<.001
Pair 10	.830	.204	.407
Pair 11	-5.382	<.001	<.001
Pair 12	1.171	.121	.243

Upon closer examination of the questions where no difference was detected, it can be noted that understanding the behavior of both humans and AI systems is similarly important. Likewise, past experiences in dealing with both humans and AI systems hold a similar significance for the participants of this study. Also, comparability and thus proximity to known and established trust structures influence interactions with other humans or other AI systems. When interacting with other humans, one's own person serves as a point of comparison, whereas with AI systems, similar other AI systems are used. The last question reaffirms the importance of experiences, as positive experiences in the past can have a positive impact on trust in the future. All these questions reaffirm the previously mentioned importance of trust. Previous experiences, similar systems, reliable results, and a transparent decision-making process are key establishing trust in an AI system.

In summary, the high number of significant differences between humans and AI systems suggests that trust must be viewed differently. Although the social factors were developed with the idea of trust between two people, this comparison illustrates well that while there are some similarities, a different trust relationship must be assumed. However, the relationship between two people and the experiences gained from it still provide clues to attempt an approximation of trust between humans and AI systems.

Conclusion

Based on the results of the study and the subsequent analysis, a closer examination of the individual factor groups should be the next step. Here, different types of AI systems are particularly important. It might be possible to establish hierarchies of factors based on the type of system. This would provide a good overview of the key factors for specific types of AI systems. Furthermore, another study with a more detailed examination of age groups could provide additional information regarding younger and older users.

Further interesting starting points can be found in the factors of transferability and behavior. Regarding transferability, it could be investigated whether human-like chatbots tend to increase trust, while human-like autonomous vehicles might decrease it. Additionally, a general direction of trust for the behavior of different AI systems to foster trust could be identified. In terms of behavior, aspects of privacy and ethics were examined. However, trust in an AI system would need to be assumed here to ensure that data misuse is avoided.

Furthermore, specific questions could be asked regarding the topic of Explainable AI. Especially the results from the comparison of human trust relationships to the trust relationship between a human and a given AI system yielded an emphasis on a transparent decision-making process. Here, respondents could first be provided with more detailed information and approaches through Explainable AI algorithms before questions are asked. This could help assess the effectiveness of specific methods. It would also be interesting to learn how explainability and transparency are desired by users. The study revealed that the option for reference was already sufficient for a large portion of the participants, indicating that an information overload may not be desired. This would mean that only in cases of uncertainty the results of an algorithm would be verified to uncover potential problems or errors in the algorithm's solution process.

The analysis of trust between humans and AI systems generally offers great potential and allows for a better assessment of which aspects of a system should be given special attention in the development process. This is especially important for technologies intended for commercial use. Different people have different notions of trust, and therefore, consideration should be given to how trust in a new technology can be established or improved. As mentioned at the beginning of this paper, it cannot be assumed that all users fully understand the technology. Trust is needed to fill this "knowledge gap", enabling the use of technology without concerns.

References

Alarcon GM, Gibson AM, Jessup SA, Capiola A, Raad H, Lee MA (2020) Effects of Reputation, Organization, and Readability on Trustworthiness Perception of Computer Code. *Human-Computer Interaction: Human Values and Quality of Life:* 367-825.

Bach TA, Khan A, Hallock H, Beltrão G, Sousa S (2022) A Systematic Literature Review of User trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human-Computer Interaction* 40(5): 1251-1266.

Bonezzi A, Ostinelli M, Melzner J (2022) The Human Black-Box: The Illusion of Understanding Human Better Than Algorithmic Decision-Making. *Journal of Experimental Psychology General* 151(9): 2250-2258.

Carleton RN (2016) Into the unknown: A review and synthesis of contemporary models involving uncertainty. *Journal of Anxiety Disorders*: 30-43.

Devitt SK (2018) Trustworthiness of Autonomous Systems, Foundation of Trusted Autonomy. *Foundations of Trusted Autonomy:* 161-184

European Commission: Directorate-General for Communications Networks, Content and Technology (2019) *Ethics guidelines for trustworthy AI*. Publications Office.

Ferronato P, Bashir M (2020) An Examination of Dispositional Trust in Human and Autonomous System Interactions. *Human-Computer Interaction: Human Values and Quality of Life, Thematic Area, HCI 2020: Proceedings Part III*: 420-435

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2019) AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28: 689-707.

Goodman, B. & Flaxman S. (2016). EU regulations on algorithmic decision making and a "right to explanation": *AI Magazine* 38(3): 50-57.

Hancock PA, Billings DR, Schaefer KE (2011) Can You Trust Your Robot?. *Ergonomics in Design The Quarterly of Human Factors Applications* 19(3): 24-29.

Hoff, K. A. & Bashir, M. (2015) Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust. *Human Factors: The Journal of the human Factors and Ergonomics Society* 57(3): 403-434.

Ikkatai Y, Hartwig T, Takanashi N, Yokoyama HM (2022) Octagon Measurement: Public Attitudes toward AI Ethics. *International Journal of Human-Computer Interactions* 38(17): 1589-1606.

Jacovi A, Marasovic A, Miller T, Goldberg Y (2021) Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 624–635.

Kaplan AD, Kessler TT, Brill JC, Hancock PA (2021) trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors: The Journal of the Human Factors and Ergonomic Society* 65(2).

Kusche I (2024) Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk. *Journal of Risk Research*: 1-14.

Luhmann N (1968) Trust. A Mechanism for the Reduction of Social Complexity. (5th ed). UVK Wiesbaden.

OECD (2017) OECD Guidelines on Measuring Trust, *OECD Publishing,* Paris.

Parasuraman R, Riley V (1997) Human Use of Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39(2): 230-253.

Smithson M (2018) Trusted Autonomy Under Certainty. *Foundation of Trusted Autonomy: Studies in Systems, Decision and Control* 117: 185-201.