Athens Journal of Social Sciences- Volume 13, Issue 1, January 2026 – Pages 29-48

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

By Nils Fischer[*] & Michael Schellenbach[±]

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

Artificial intelligence (AI) systems, particularly those based on machine learning (ML) and deep neural networks, are increasingly integrated into everyday technologies. From autonomous vehicles to clinical decision support tools, these systems promise efficiency and precision, but often at the cost of transparency. Users are asked to place trust in algorithms whose internal workings remain largely opaque (Carleton 2016). This challenge is commonly referred to as the "black box" problem. While developers may understand the overall architecture of a neural network, the detailed transformations that an input undergoes within the network often remain inaccessible. This lack of transparency is especially problematic in safety-critical domains such as healthcare or mobility, where decisions made by AI systems can have direct consequences for human well-being. In such contexts, physicians, engineers, or the public must be able to trust the system, even if they cannot fully comprehend its inner logic. This raises a fundamental question: What factors influence users' trust in AI, and how does this differ across different contexts?

To frame this inquiry, we define trust as a psychological state involving the willingness to accept vulnerability based on positive expectations of another's

[*]Lecturer, Hochschule Ruhr West University of Applied Sciences, Germany.
Research Associate, Hochschule Ruhr West University of Applied Sciences, Germany.

behavior (Mayer et al. 1995, Ferronato & Bashir 2020). In the context of AI, trust describes the extent to which users believe that a system will behave reliably, ethically, and competently, especially when its decisions are difficult to explain. Following Luhmann (1968), trust also serves a functional role: it reduces complexity and allows action in uncertain situations. The opaquer the system becomes, the more critical trust will become (Carleton 2016). Recent research has emphasized that trust in AI is multi-dimensional, encompassing technical as well as social aspects (Jacovi et al. 2021, Ikkatai et al. 2022). Technical trust refers to confidence in system functionality, stability, and accuracy. Social trust, by contrast, draws on interpersonal analogies and reflects whether the system appears accountable, transparent, or aligned with user values.

Explainable AI (XAI) has been proposed as one solution to the trust dilemma (Floridi et al. 2019, Goodman & Flaxman 2016). While providing algorithmic transparency may increase trust in principle, recent findings suggest that explainability alone is insufficient. Trust must also be calibrated based on user expectations, perceived risk, and social cues (Bonezzi et al. 2022, Li et al. 2024). Against this background, this study investigates which specific factors influence trust in AI systems and whether these factors differ depending on the application context, namely between safety-critical and non-safety-critical domains. This research aims to classify these factors into a hierarchical model, referred to as the Trust Pyramid and empirically examines how users evaluate trustworthiness based on varying system attributes and contexts.

Literature Review

The question of trust in artificial intelligence has become increasingly central in recent years, as AI systems are deployed in more complex and sensitive domains. In response, numerous academic studies have sought to define what makes AI trustworthy. One significant benchmark is the Ethics Guidelines for Trustworthy AI published by the European Commission (2019), which outlines seven key requirements: human agency, robustness, privacy, transparency, ethical behavior, sustainability, and accountability. These principles are not merely idealistic; they form the necessary foundation for building public trust in AI systems. Without such normative guidelines, trust is unlikely to emerge in the first place.

However, research also highlights other practical aspects that influence trust. Factors such as explainability, controllability, reliability, and usability have been shown to significantly affect how users perceive and engage with AI systems (Jacovi et al. 2021, Ikkatai et al. 2022). These dimensions extend beyond ethical design into everyday interaction, indicating that trust is not just designed into a system, it must also be earned through its behavior and performance in use.

To better understand the factors that shape trust, scholars often distinguish between two broad categories: technical trust and social trust. Technical trust involves confidence in a system's functionality, accuracy, and stability, while social trust is rooted in transparency, ethical alignment, and shared values (Hoff & Bashir 2015, Hancock et al. 2011). Interestingly, even though human-machine trust differs from interpersonal trust in important ways, it is still governed by many of the same

psychological principles. For instance, reliability and traceability play vital roles in both contexts.

A particularly important nuance in the literature is the distinction between trust and dependence. As Parasuraman and Riley (1997) observed, users may continue to rely on AI systems even after they fail, suggesting not genuine trust but rather dependency. This can be especially problematic in contexts where users have no alternative tools. Ferronato et al. (2020) emphasize that forced reliance without authentic trust can reduce safety and system effectiveness. Just as in human relationships, trust built on necessity rather than choice may lead to poor outcomes over time.

Trust is not a fixed attribute; rather, it evolves throughout the interaction with a system. Ferronato et al. (2020) introduce the concepts of primary trust, formed during the first encounter with a system and secondary trust, which develops with ongoing use. Initial impressions, shaped by prior experiences, public discourse, or even media representation, often have a disproportionate influence on future evaluations. This idea is echoed by Bonezzi et al. (2022), who describe trust development as a mix of rational shortcuts and emotional responses. Hancock et al. (2011) further conceptualize trust as a process involving three elements: a sender (the user), a receiver (the system), and a communication channel (the interface or environment). As AI systems become more autonomous, this communication becomes increasingly complex and more like interpersonal interactions.

A major challenge arises when AI systems appear competent but act unethically or unreliably. This phenomenon, often described as false trust, can emerge when persuasive design conceals system limitations (Devitt 2018, Hancock et al. 2011). Users may assume a system is trustworthy based on smooth interactions, even when its decisions are flawed or biased. In this context, the concept of functional trust becomes particularly relevant referring to the belief that a system can correctly carry out a specific task. Misplaced functional trust can be dangerous, especially when users overestimate a system's abilities or fail to scrutinize its outputs.

In line with this, Smithson (2018) draws an important distinction between mistrust and the absence of trust. Mistrust is based on active doubt, while absence of trust often results from insufficient information. For new or unfamiliar AI systems, the latter is more common and more manageable. Through measures such as transparency and user education, designers can help bridge this informational gap and cultivate trust where mistrust might otherwise take root (Floridi et al., 2019).

Taken together, the literature points to a multidimensional and context-sensitive view of trust in AI. Technical performance, ethical alignment, and social resonance must all come together to foster trust. Recent studies (Gu et al. 2024, Baron 2024) go even further, arguing that users seek not just clarity and transparency, but also emotional reassurance and alignment with their values. It is not enough to show that system workers must feel that it works for them. As such, AI developers must adopt trust-aware design strategies that go beyond functionality to include audience-specific expectations and experiences. Ultimately, trust in AI is not a static trait but a dynamic, evolving relationship between humans and machines, shaped by interaction, context, and the broader social environment.

The Human Black Box

While much of the literature emphasizes technical functionality and ethical alignment as pillars of trustworthy AI, a deeper paradox emerges when we examine the human tendency to ascribe trust differently to humans versus machines. In human-to-human interactions, trust is frequently placed in individuals whose internal decision-making processes remain largely opaque. Bonezzi et al. (2022) refer to this phenomenon as the "illusion of transparency". Although people may assume that other humans are more understandable than algorithms, this assumption is often based on projection: individuals use their own cognitive processes as a template to interpret and predict others' behavior. In contrast, AI systems are perceived as foreign and inaccessible, making it more difficult for users to map their own decision-making process onto the system.

This distinction plays a significant role in shaping user expectations. Humans tend to trust other people not necessarily because they understand them better, but because they empathize with them more readily. Empathy and familiarity create a perceived social proximity, which in turn facilitates trust, even when objective transparency is lacking. AI systems, however, do not elicit the same affective responses. The differences are magnified by the absence of shared experiences or emotional cues, thereby reinforcing user skepticism.

Interestingly, while regulations like the European Artificial Intelligence Act (European Commission, 2024) increasingly mandate transparency in algorithmic processes (Goodman & Flaxman 2016), no equivalent standards are applied to human decision-makers. Physicians, for instance, are often trusted without offering patients a detailed explanation of their diagnostic reasoning. Likewise, passengers routinely trust taxi drivers without full insight into their moment-to-moment decisions. Yet trust is still extended in these cases, often through indirect mechanisms such as institutional accountability, legal frameworks, or shared social norms. Kusche (2023) refers to this phenomenon as "bridging asymmetry" between decision-makers and those affected, a crucial element in negotiating trust in uncertain or high-stakes contexts like Luhmann's approach.

This asymmetry becomes even more pronounced in interactions with AI. Unlike humans, AI systems do not possess consciousness, intentionality, or emotional responsiveness. As such, users often struggle to form what could be called relational trust with these systems. Bonezzi et al. (2022) finds that while interpersonal closeness enhances trust between humans, no such relational shortcut exists with machines. Without the ability to empathize with a system, users are left to rely solely on external indicators, such as transparency reports, interface design, or third-party validation, to assess trustworthiness.

One intriguing proposition could be the use of a "white-box" algorithm as a reference point for understanding and trusting a more complex "black-box" system. If users can be provided with a simplified model that mimics or explains the behavior of an opaquer system, they may be more willing to accept its outputs. This conceptual shift moves beyond raw explainability toward cognitive modeling, in which the user builds a mental framework based on analogies and simplifications.

To explore this further, the present study distinguishes between three key categories of trust-influencing factors: technical, social, and evaluation factors. Technical and social factors have been widely documented in prior research. For example, Ikkatai et al. (2022) proposed an eight-dimensional trust framework that encompasses functionality, transparency, and fairness, among others. Similarly, Kaplan et al. (2021) and Bach et al. (2022) have developed taxonomies to classify attributes contributing to system trustworthiness expanding upon frameworks that defined definitions and measures of IT-Systems without an AI component (McKnight et al. 2011). These frameworks help to differentiate trust in technology from trust in people.

However, this study introduces evaluation factors as a third category, aimed at capturing the meta-cognitive and contextual influences that shape trust formation. These include users' prior experiences, perceived domain risk, media portrayals of AI, and the communicative framing of a system's purpose and capabilities. Yardi (2022) writes that the biggest problem that AI technology faces today is not that it is unethical but that it is being used by large and powerful corporations that have unethical business models. By incorporating evaluation factors, we aim to better understand how trust is not only built from system attributes but also dynamically shaped by user interpretation, situational context, and broader societal narratives.

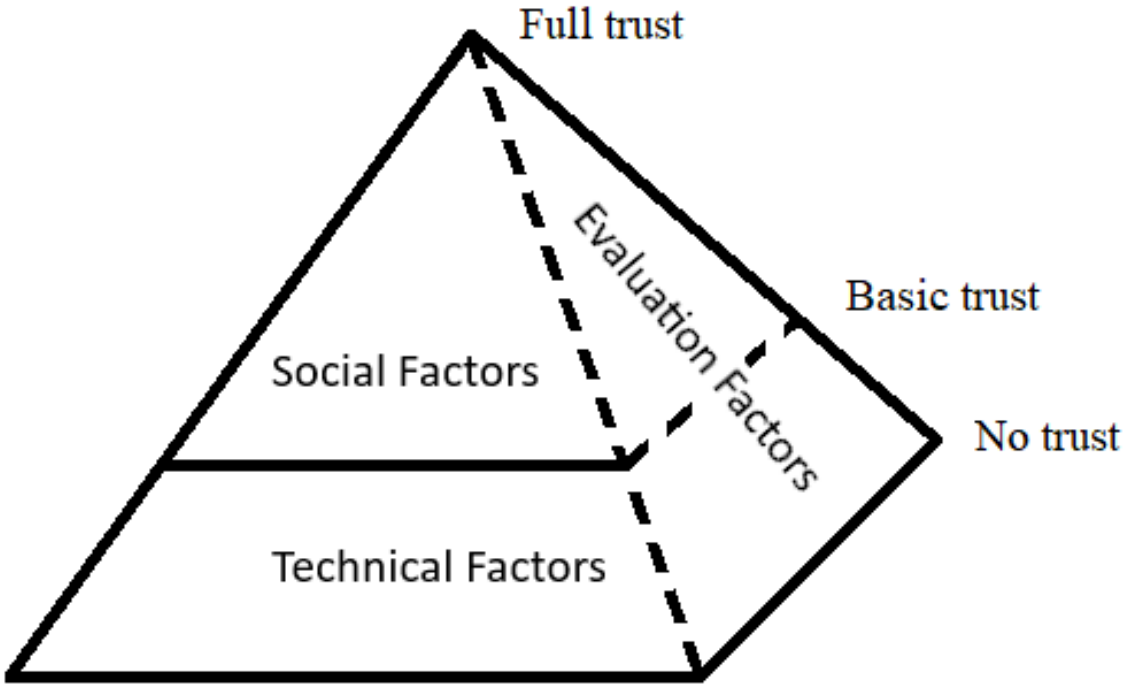
Introducing the Trust Pyramid as a hierarchy

After the analysis of other studies, we conducted our own to empirically assess the factors identified in the previous sections. This investigation was guided by two central research questions:

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

These questions are interdependent. While the first addresses the hierarchy and relative importance of various trust dimensions, the second introduces the variable of application context. The comparison distinguishes between scenarios where AI systems are deployed. On the one hand in safety-critical environments (autonomous driving, clinical diagnostics, etc.) and on the other hand in non-safety-critical settings (content recommendation, smart assistants, etc.). The findings are synthesized into a conceptual model, the Trust Pyramid, which classifies and organizes trust-influencing factors according to their role in building comprehensive trust in an AI system.

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



The Trust Pyramid is intended as a generalizable yet individualized heuristic. It assumes a user-centric perspective and applies to a single AI system evaluated by a single user. This model does not presume a uniform process of trust formation but instead acknowledges the deeply personal and contextual nature of trust. Everyone brings a unique epistemic baseline, shaped by prior experience, cognitive predispositions, and worldview. In cases where a user lacks familiarity with AI or has limited exposure to digital technologies, the trust formation process may diverge significantly from that of an experienced user. At the lowest level of the pyramid, trust is assumed to be absent reflecting the default state prior to any user interaction with the system.

The base layer of the pyramid encompasses technical trust factors. These include system stability, reliability, behavior, comprehensibility, and the accuracy of results. Trust at this level is grounded in the system's demonstrable capabilities. Without a minimum threshold of technical reliability, higher-order trust dimensions cannot be meaningfully engaged. A malfunctioning or unintelligible system erodes the foundations upon which trust must be built. Therefore, technical robustness is treated as a necessary, but not sufficient, condition for trust.

An important nuance is introduced through the dashed boundary that connects technical and evaluation factors. While a technically sound system may satisfy functional requirements, users also evaluate its trustworthiness based on extrinsic cues such as the reputation of its developer, the perceived intentions of the organization behind it, or sociopolitical narratives associated with its deployment. For example, the public stance of a CEO or controversies surrounding a tech company may influence perceptions independently of system performance. These contextual elements constitute the evaluation dimension of trust, which intersects with and overlays the technical foundation.

Building upon the technical base, the second tier of the pyramid introduces social trust factors, such as reliability, accountability, traceability, competence, and transferability. These attributes reflect the user's perception of whether the AI system behaves in ways that are socially and morally acceptable. Importantly, these factors gain relevance only when the underlying technical trust has been established. A user is unlikely to consider questions of fairness or accountability in a system they perceive as unreliable or dysfunctional. As such, the social layer depends structurally on the technical substrate.

The left face of the pyramid, comprising technical and social trust factors, represents the trust that can be cultivated solely through the system's architecture and behavior. This constitutes the maximal intrinsic trust attainable based on the system itself. By contrast, the right face of the pyramid captures trust shaped by evaluation factors, including user biases, media narratives, cultural framing, and perceived institutional legitimacy. These elements are often beyond the control of system developers but significantly shape how users interpret and respond to AI technologies.

At the top of the pyramid lies full trust, a state in which the system is not only technically and socially validated, but also contextually affirmed. Achieving this requires more than internal system quality. It demands trust in the actors behind the system and a favorable public discourse surrounding AI. Conversely, even a technically and socially sound system may fall short of this ideal if external signals, such as negative media coverage or prior incidents of misuse, undermine user confidence.

Taken together, the Trust Pyramid proposes a layered and conditional model of trust formation in AI, where technical factors serve as a prerequisite, social factors as a moral and relational extension, and evaluation factors as a contextual filter. This model emphasizes that trust is not merely a matter of system design but a holistic construct, shaped by both internal attributes and external narratives.

Methodology

To empirically ground the proposed Trust Pyramid and validate its underlying dimensions, the following section outlines the identification and operationalization of trust-related factors. These were derived through an integrative process that combined insights from existing literature with findings from a series of preliminary test runs. Specifically, nine pilot studies were conducted prior to the main survey to refine both the theoretical framework and the questionnaire structure. These trials served to ensure that the factors were both theoretically sound and practically comprehensible for participants, thus enhancing the interpretability and internal validity of the study.

The model distinguishes between three overarching categories of trust-building factors: technical, social, and evaluation-based. Each category is defined not only by its internal components but also by its relation to the others, forming a layered understanding of how trust is structured and shaped in user–AI interactions.

Technical factors refer to attributes intrinsic to the AI system itself. Unlike social or evaluation-based factors, technical elements are less amenable to interpersonal analogies and cannot be readily mapped onto human-to-human trust dynamics. Nonetheless, they serve as prerequisites for the emergence of higher-order trust components. Given their foundational role, this study includes privacy protection and ethical compliance under the technical domain, recognizing these as essential operational features that precede and enable social trust.

The technical dimension is composed of the following factors:

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

Social trust emerges once a sufficient level of technical trust has been established. These factors mirror those found in interpersonal relationships and relate to how users ascribe intent, reliability, and values to the AI system. While the system is still recognized as non-human, users may interact with it in ways analogous to human relationships, particularly in contexts where the system behaves autonomously or takes on roles traditionally associated with human decision-makers.

The social dimension is composed of the following factors:

- Traceability
- Reliability (Me/Others)
- Accountability
- Competence
- Transferability

Beyond system-specific attributes, users' trust is also shaped by broader contextual and experiential variables. These evaluation factors pertain to meta-level perceptions, including reputational cues and prior exposures to AI technologies. Unlike technical and social factors, which are often embedded in the design and behavior of the system itself, evaluation factors arise from external influences, media discourse, public sentiment, and institutional trust. These variables serve as interpretive lenses through which users assess whether an AI system is worthy of trust.

The evaluation dimension is composed of the following factors:

- Reputation (Media/Popular Culture)
- Reputation (Developer/Manufacturer)
- Experiences (Other AI Systems)
- Experiences (Other People)
- Testing Reports (Magazines/Consultations)

Study Design

To empirically assess the salience of these trust factors across different AI use contexts, a structured questionnaire was developed and administered. The design of the instrument was guided by the OECD Guidelines on Measuring Trust (OECD 2017), which provided a robust framework for translating abstract concepts into measurable items. The guidelines were written about interpersonal trust relationships and therefore adjusted to human-AI trust dynamics. As the study was conducted in a fully digital environment, particular attention was given to usability and response flow.

The data collection occurred over a 46-day period and involved 258 valid participants. An initial total of 294 individuals began the questionnaire; however, 19 responses were excluded due to incomplete submissions, and a further 17 were discarded after failing embedded attention-check items. The final sample consisted of 138 male, 118 female, 1 non-binary participant, and 1 participant who chose not to disclose their gender identity.

Each participant was presented with a series of contextualized AI use cases. These were situated in safety-critical (autonomous driving) and non-safety-critical settings (chatbot interactions). Individual survey items were directly linked to the identified trust factors, allowing for a granular assessment of how each category contributed to overall perceptions of system trustworthiness. The questionnaire items were structured based on a 5-point Likert scale, where always two questions represented one of the aforementioned factors.

The questionnaire required approximately 10 – 15 minutes to complete, a duration that was pre-tested for feasibility and confirmed via participant feedback. The pre-test was conducted with a separate sample of 10 individuals. It served as a mean to troubleshoot potential usability issues and to ensure the internal logic of the questionnaire. This included skip patterns and dependency conditions.

Participants were recruited through Survey Circle, an online research exchange platform affiliated with the Research Institute Mannheim. The platform incentivizes reciprocal participation across academic studies, thus ensuring a diverse and engaged sample pool. This approach not only facilitated data collection but also broadened the range of perspectives included in the study, thereby enhancing the generalizability of the results.

The comparison objectives of this study were to determine which factors most significantly contribute to the formation of trust in AI systems, and how these contributions vary depending on the context of use, the gender and the comparison between human-human and human-AI trust relationships.

Results

The empirical findings of this study provide initial answers to the central research question: What factors hold the highest priority when it comes to fostering trust in an AI system? The results underscore that trust is not governed by a single

determinant, but by a constellation of interrelated factors that vary in salience depending on the system context.

Table 1. *Results regarding the Technical Factors*

Technical Trust Factor	Does not apply	Does not apply %	Rather does not apply	Rather does not apply %	Partly / Partly	Partly / Partly %	Rather applies	Rather applies %	Applies	Applies %
Stability	2	0.78%	3	1.16%	7	2.71%	106	41.09%	140	**54.26%**
	2	0.78%	10	3.88%	50	19.38%	118	**45.74%**	78	30.23%
Applicability	1	0.39%	11	4.26%	41	15.89%	106	**41.09%**	99	38.37%
	4	1.55%	11	4.26%	52	20.16%	107	**41.47%**	84	32.56%
Behavior (Privacy)	4	1.55%	11	4.26%	17	6.59%	73	28.29%	153	**59.30%**
	3	1.16%	14	5.43%	56	21.71%	105	**40.70%**	80	31.01%
Behavior (Ethics)	6	2.33%	30	11.63%	56	21.71%	90	**34.88%**	76	29.46%
	26	10.08%	55	21.32%	80	**31.01%**	59	22.87%	38	14.73%
Comprehensibility	11	4.26%	42	16.28%	66	25.58%	92	**35.66%**	47	18.22%
	4	1.55%	19	7.36%	49	18.99%	102	**39.53%**	84	32.56%
Accuracy	2	0.78%	2	0.78%	26	10.08%	122	**47.29%**	106	41.09%
	2	0.78%	8	3.10%	24	9.30%	108	41.86%	116	**44.96%**

Table 1 shows that three factors, among the technical dimensions, emerged as particularly influential. Those factors are stability, applicability, and accuracy. These were consistently rated as central to establishing baseline trust across AI system types. Their prominence suggests that users prioritize predictable performance, usability, and dependable outcomes when interacting with AI. While comprehensibility was a little less frequently emphasized, its conceptual overlap with the social factor of traceability points to an important nuance: users often do not require full transparency or deep technical understanding of a system. Instead, the mere ability to query or retrospectively review decisions appears sufficient to foster confidence. Moreover, users may compensate for gaps in understanding by drawing on trust established through other dimensions, particularly reliability or accountability. While the factor behavior is more strongly rated towards privacy, the ethical component isn't viewed as important.

Table 2. *Results regarding the Social Factors*

Social Trust Factor	Does not apply	Does not apply %	Rather does not apply	Rather does not apply %	Partly / Partly	Partly / Partly %	Rather applies	Rather applies %	Applies	Applies %
Traceability	13	5.04%	37	14.34%	64	24.81%	96	**37.21%**	48	18.60%
	10	3.88%	16	6.20%	52	20.16%	116	**44.96%**	64	24.81%
Reliability (Me)	7	2.71%	16	6.20%	55	21.32%	122	**47.29%**	58	22.48%
	4	1.55%	2	0.78%	12	4.65%	91	35.27%	149	**57.75%**
Accountability	5	1.94%	12	4.65%	45	17.44%	121	**46.90%**	75	29.07%
	2	0.78%	5	1.94%	26	10.08%	94	36.43%	131	**50.78%**
Competence	2	0.78%	7	2.71%	15	5.81%	109	42.25%	125	**48.45%**
	2	0.78%	4	1.55%	29	11.24%	103	39.92%	120	**46.51%**
Transferability	28	10.85%	74	28.68%	103	**39.92%**	36	13.95%	17	6.59%
	15	5.81%	28	10.85%	52	20.16%	125	**48.45%**	38	14.73%
Reliability (Others)	17	6.59%	40	15.50%	87	**33.72%**	87	**33.72%**	27	10.47%
	2	0.78%	6	2.33%	26	10.08%	122	**47.29%**	102	39.53%

In the social domain, Table 2 shows that reliability, accountability, and competence emerged as the most critical factors. These factors reflect the human-like qualities users ascribe to AI systems, especially in scenarios where decision-making or autonomy is involved. The factor of traceability is valued a little bit less than the already mentioned factors. Although classified as social, traceability also links back to the technical infrastructure, especially in its dependency on the system's ability to make decision-making pathways visible. Interestingly, transferability, which captures the alignment between user values and system behavior, was more difficult for participants to conceptualize. Its relevance varied by context: it was more readily imagined in interactions with chatbots than in the case of autonomous vehicles. This

suggests that the degree to which users expect, or desire human-likeness is not universal but context-dependent, warranting further investigation.

Table 3. *Results regarding the Evaluation Factors*

Evaluation Trust Factor	Does not apply	Does not apply %	Rather does not apply	Rather does not apply %	Partly / Partly	Partly / Partly %	Rather applies	Rather applies %	Applies	Applies %
Reputation (Media / Popular Culture)	28	10.85%	43	16.67%	77	29.84%	80	**31.01%**	30	11.63%
	40	15.50%	77	**29.84%**	67	25.97%	60	23.26%	14	5.43%
Reputation (Developer / Manufacturer)	5	1.94%	12	4.65%	42	16.28%	146	**56.59%**	53	20.54%
	4	1.55%	16	6.20%	41	15.89%	135	**52.33%**	62	24.03%
Experiences (Other AI Systems)	9	3.49%	34	13.18%	81	31.40%	101	**39.15%**	33	12.79%
	4	1.55%	9	3.49%	41	15.89%	141	**54.65%**	63	24.42%
Experiences (Other People)	17	6.59%	36	13.95%	69	26.74%	85	**32.95%**	51	19.77%
	12	4.65%	18	6.98%	79	30.62%	113	**43.80%**	36	13.95%
Testing Reports (Magazines / Consultations)	4	1.55%	11	4.26%	31	12.02%	124	**48.06%**	88	34.11%
	2	0.78%	12	4.65%	44	17.05%	131	**50.78%**	69	26.74%

Table 3 shows that in terms of the evaluation factors, the most significant influences that were identified were the reputation towards developers/manufacturers, the personal experience regarding AI systems, and professional testing reports. While media portrayals, such as those found in films, games, or news stories, can shape initial perceptions, they appear to play a secondary role in trust formation. More significant are concrete, domain-specific evaluations from credible sources, along with users' own experiences. Similarly, anecdotal experiences from others (family or friends) were found to be less influential than firsthand interaction with AI systems. These findings emphasize the importance of direct engagement and institutional credibility in shaping trust.

A central insight of this study is that trust factors cannot be weighted uniformly across all AI systems. Different application domains elicit different expectations from users. For instance, in safety-critical contexts such as autonomous driving, stability and accuracy were prioritized, reflecting the need for operational reliability. In contrast, with systems such as chatbots, participants emphasized factors like privacy and ethical behavior, highlighting concerns over data handling and moral alignment. The results also indicate that users are more willing to share personal data with safety-critical systems, provided those systems meet higher standards of technical reliability. Conversely, non-safety-critical systems are often held to stricter standards of behavioral appropriateness despite posing fewer tangible risks.

This differentiation demonstrates that a singular, universal trust hierarchy is untenable without accounting for the use-case context. Therefore, any trust model must allow for context-sensitive weighting of factors to remain analytically valid and practically useful.

The proposed Trust Pyramid offers a preliminary tool for organizing the diverse influences on trust in AI. By distinguishing between technical, social, and evaluation factors, and by allowing for their dynamic interplay, the model expands upon existing frameworks proposed by Ikkatai et al. (2022), Kaplan et al. (2021), and Bach et al. (2022). While prior models primarily emphasized technical and social dimensions, the inclusion of evaluation factors in this study introduces a crucial third layer, capturing the external, contextual, and perceptual dynamics that often shape user trust in subtle yet powerful ways.

This enriched framework enables a more granular classification of trust dimensions and their interrelations, facilitating a better understanding of how users assess AI systems in practice. The findings underscore that trust is neither static nor

singular, but a layered construct shaped by a confluence of internal system properties, perceived social attributes, and external influences.

Discussion

The second research question of this study sought to determine whether the context in which an AI system operates plays a significant role in the formation of user trust. Specifically, the study investigated whether trust-related factors are consistent across different types of AI systems, or whether they vary based on the system's domain of application. To explore this question, two contrasting AI systems were selected for comparison: a chatbot, representing a non-safety-critical system, and an autonomous vehicle, representing a safety-critical system. These systems differ not only in their functions and use cases but also in the perceived risks and user expectations associated with them. This contrast enabled a nuanced analysis of context-specific trust dynamics.

To facilitate a direct comparison, 17 questions, each corresponding to one of the predefined trust factors, were selected from the main questionnaire. These questions were reformulated to apply to both the chatbot and the autonomous vehicle scenarios and aligned with one of the three overarching factor groups: technical, social, or evaluation factors.

Differences between Safety-critical and non-safety-critical Contexts

This section builds on several underlying assumptions. Safety-critical systems such as autonomous vehicles are typically perceived as more complex and risk-heavy due to their potential impact on physical safety. In contrast, chatbots generally involve lower stakes and are more familiar to users, particularly following the widespread adoption of systems like ChatGPT-3. As a result, trust in these two types of systems may be shaped differently based on factors such as familiarity, perceived risk, and media portrayal.

The 17 items presented to participants were as follows:

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

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

To determine whether significant differences exist between the two AI systems under investigation, several T-tests were conducted. A significance level of 0.001 was selected to ensure a detailed and reliable interpretation of the results. Table 1 presents an overview of the 17 paired comparisons, each corresponding to one of the 15 trust factors outlined in the methodology. Two factors, behavior, and experience were further divided to provide a more granular comparison. Specifically, the behavior factor was separated into privacy and ethical behavior, while the experience factor was split into personal experiences and experiences of other people.

Table 4. *T-test for comparing Safety-critical and non-safety-critical AI Systems*

Paired Samples T-Test

		Paired Difference							Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower value	Upper Value	T	df	One-tailed p	Two-tailed p
Pair 1	Stability	-,527	,819	,051	-,628	-,427	-10,340	257	<,001	<,001
Pair 2	Applicability	-,202	,798	,050	-,299	-,104	-4,059	257	<,001	<,001
Pair 3	Behavior (Privacy)	,105	1,091	,068	-,029	,238	1,540	257	,062	,125
Pair 4	Behavior (Ethics)	,124	1,051	,065	-,005	,253	1,895	257	,030	,059
Pair 5	Comprehensibility	-,643	1,179	,073	-,788	-,499	-8,766	257	<,001	<,001
Pair 6	Accurary	-,337	,710	,044	-,424	-,250	-7,625	257	<,001	<,001
Pair 7	Traceability	-,674	1,131	,070	-,813	-,536	-9,579	257	<,001	<,001
Pair 8	Reliability	-,260	,653	,041	-,340	-,180	-6,387	257	<,001	<,001
Pair 9	Accountability	-,667	1,061	,066	-,797	-,537	-10,092	257	<,001	<,001
Pair 10	Competence	-,333	,720	,045	-,422	-,245	-7,433	257	<,001	<,001
Pair 11	Transferability	-,159	1,387	,086	-,329	,011	-1,840	257	,033	,067
Pair 12	Reputation (Media)	-,593	1,113	,069	-,729	-,457	-8,560	257	<,001	<,001
Pair 13	Reputation (Developer)	-,779	1,220	,076	-,929	-,630	-10,259	257	<,001	<,001
Pair 14	Experience (AI systems)	-,632	1,029	,064	-,758	-,506	-9,864	257	<,001	<,001
Pair 15	Experience (Other)	-,481	1,014	,063	-,605	-,356	-7,611	257	<,001	<,001
Pair 16	Experience (Mine)	-,640	1,061	,066	-,770	-,509	-9,679	257	<,001	<,001
Pair 17	Testung reports	-,593	,874	,054	-,700	-,486	-10,900	257	<,001	<,001

The results presented in Table 1 indicate that significant differences were observed in nearly all trust factors when comparing chatbots and autonomous vehicles. Exceptions include the factors "Behavior (Privacy)", "Behavior (Ethics)", and "Transferability", for which no statistically significant differences were detected. This suggests that for these factors, participants expected both systems to operate under similar principles regardless of context. In contrast, all other factors yielded significant differences. Moreover, the negative T-values indicate that, for those factors, autonomous vehicles were rated more highly in terms of trust. This supports

the earlier assumption that trust in AI systems cannot be ranked in a universal hierarchy without considering the specific context in which a system operates.

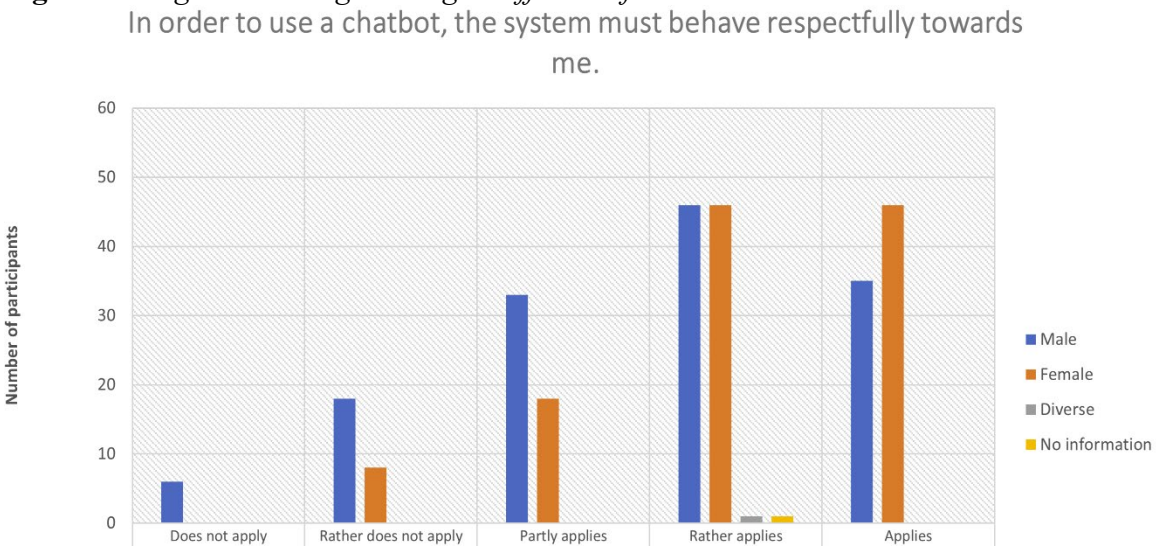
These findings highlight that the underlying nature of the AI system plays a decisive role in how trust factors are perceived and prioritized. As such, the concept of a "trust pyramid" remains useful, serving as a general framework for classifying trust factors into technical, social, and evaluation categories. However, it does not offer a fixed sequence of importance across different AI systems. The conclusion drawn from this analysis is clear: the context in which an AI system operates significantly influences the development of trust. Needed trust in safety-critical systems, such as autonomous vehicles, is generally stronger than in non-safety-critical systems like chatbots.

As a result, it is not feasible to establish a universal hierarchy of trust factors applicable to all AI systems. Instead, the prioritization of factors must be evaluated individually, depending on the nature of the specific system in question. Furthermore, since the trust pyramid also incorporates subjective, human-centered considerations, individual users may assign differing importance to the same factors even when evaluating the same AI system. A broader analysis that examines specific factor groups across multiple AI system types could eventually help identify patterns or categories of systems with similar trust profiles.

Gender-based differences between Female and Male Participants

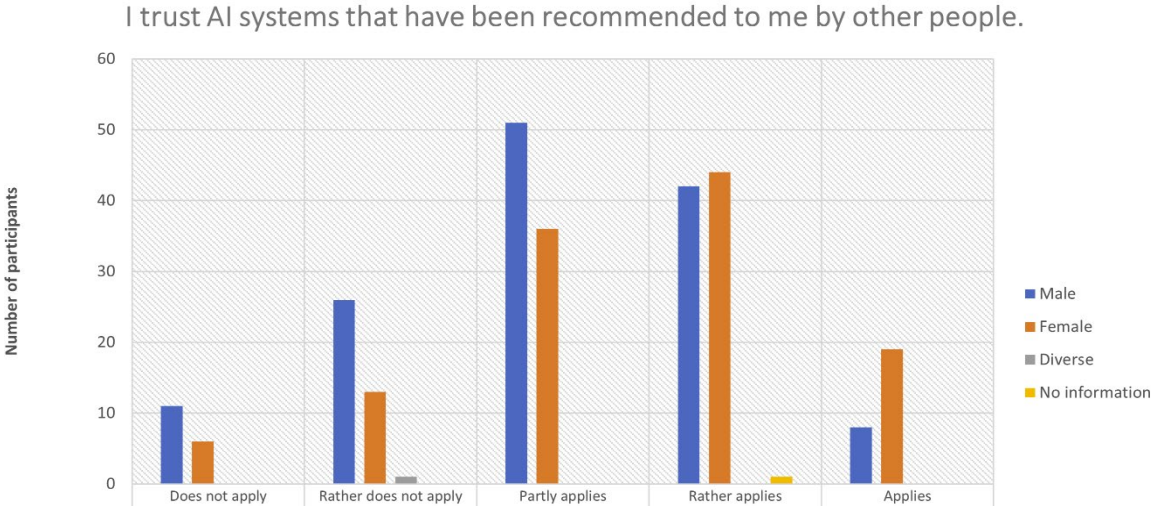
Given the nearly equal distribution of female and male participants, the study also examined potential gender-based differences in responses. Overall, the analysis revealed no substantial differences between the two groups. To assess any potential discrepancies, a separate T-test was conducted for each individual question, applying the same significance threshold of 0.001 as in the earlier comparison between safety-critical and non-safety-critical AI systems.

Figure 2. *Diagram showing the Single Difference for Technical Factors*



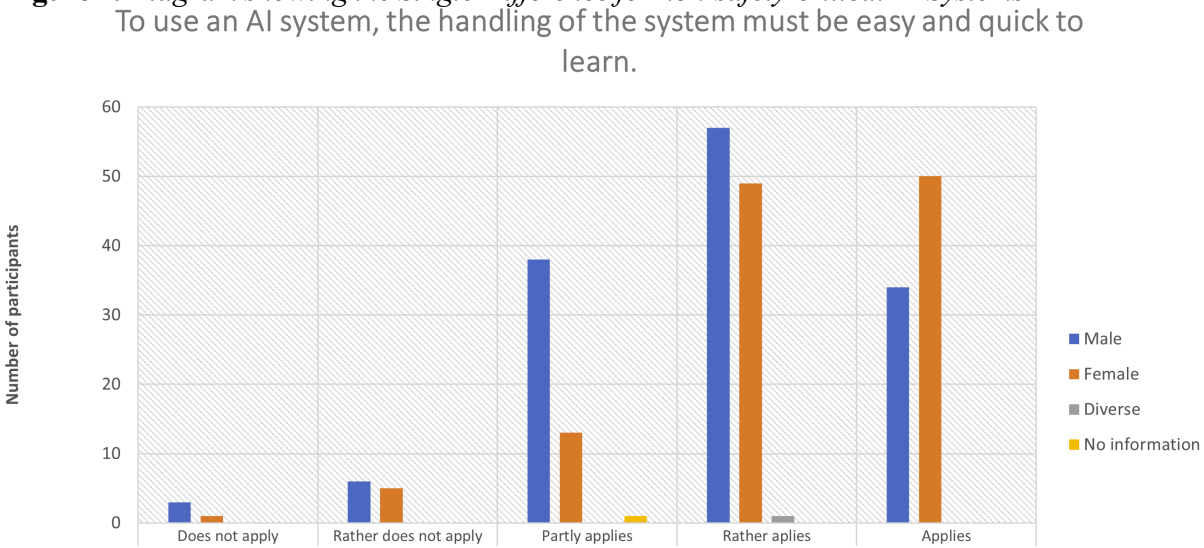
Only one significant difference was observed among the technical factors, as illustrated in Figure 2. This difference pertained to the importance of rapid learnability when interacting with an AI system. Female participants rated quick and easy handling as more important compared to their male counterparts, who assigned less importance to this aspect. While the remainder of the technical questions did not show statistically significant differences, this specific distinction highlights a gender-based divergence in expectations regarding user-friendliness.

Figure 3. *Diagram showing the Single Difference for Social Factors*



A similar pattern emerged among the social factors. As shown in Figure 3, only one question produced a significant difference between men and women. This question asked whether a recommendation from another person would influence trust in an AI system. Female participants demonstrated a greater tendency to value external recommendations as a basis for their own trust, while male participants were more inclined to rely primarily on their personal assessment of the system.

Figure 4. *Diagram showing the Single Difference for non-safety-critical AI Systems*



In the case of safety-critical AI systems, no statistically significant gender differences were identified for any of the questions. For non-safety-critical systems, however, one notable difference emerged. As depicted in Figure 4, the question concerned the importance of respectful behavior from a chatbot. Female participants placed more importance on respectful interaction, whereas male participants were generally less concerned with this attribute and were more likely to use the chatbot regardless of its demeanor.

Taken together, these findings indicate that gender has only a limited influence on the evaluation of trust in AI systems. While some isolated differences were observed, particularly in areas relating to usability, external influence, and respectful behavior, they were not widespread or consistent across factor groups. Therefore, it can be concluded that gender does not play a major role in shaping attitudes toward AI system trustworthiness in this context.

Differences in Trust Dynamics between Humans and AI Systems

At the beginning of the questionnaire, general questions were posed to gather insights into participants' overall attitudes toward trust. These questions focused on interpersonal trust and closely mirrored the social factors later used to assess trust in AI systems. Since the social factor items were originally inspired by dimensions of human-to-human trust, this allowed for a direct comparison between trust in people and trust in AI. To explore whether significant differences exist between these two types of trust, paired T-tests were conducted on matched questions that addressed similar constructs but varied in their reference, either to human beings or to AI systems.

Each pair of questions compared a statement about a human counterpart with a similarly worded statement about an AI system. In cases where the phrasing was identical for both contexts, the item was listed only once. The matched questions included elements of the social factors such as reliability, accountability, competence, traceability, and transferability.

The items presented to participants were as follows:

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

Table 2 summarizes the outcomes of these paired comparisons. The analysis revealed that significant differences existed in eight out of the twelve question pairs, indicating that trust in AI systems is not perceived in the same way as trust in humans. Notably, no significant differences were found for four specific pairs: the explanation of behavior (pair two), demonstrated reliability (pair four), comparability or perceived similarity (pair ten), and prior positive experiences (pair twelve). These exceptions point to certain dimensions of trust, such as the importance of past performance or perceived familiarity, that are similarly valued in both human and AI contexts.

Table 5. *T-test for comparing Trust in Humans and AI Systems*

Paired Samples T-Test

		Paired Differences							Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower Value	Upper Value	T	df	One-tailed p	Two-tailed p
Pair 1	Traceability	,589	1,327	,083	,426	,752	7,132	257	<,001	<,001
Pair 2	Traceability	,128	1,195	,074	-,019	,274	1,720	257	,043	,087
Pair 3	Reliability (Me)	,349	1,194	,074	,202	,495	4,691	257	<,001	<,001
Pair 4	Reliability (Me)	-,171	,975	,061	-,290	-,051	-2,808	257	,003	,005
Pair 5	Accountability	-,399	1,177	,073	-,543	-,255	-5,449	257	<,001	<,001
Pair 6	Accountability	-,380	1,023	,064	-,505	-,254	-5,966	257	<,001	<,001
Pair 7	Competence	-,368	1,105	,069	-,504	-,233	-5,351	257	<,001	<,001
Pair 8	Competence	-,562	1,054	,066	-,691	-,433	-8,565	257	<,001	<,001
Pair 9	Transferability	,857	1,335	,083	,693	1,020	10,310	257	<,001	<,001
Pair 10	Transferability	,066	1,275	,079	-,090	,222	,830	257	,204	,407
Pair 11	Reliability (Others)	-,326	,972	,060	-,445	-,206	-5,382	257	<,001	<,001
Pair 12	Reliability (Others)	,066	,904	,056	-,045	,177	1,171	257	,121	,243

Closer examination of these four non-significant pairs reveals that participants attributed equal importance to transparent behavior and reliable performance, regardless of whether the subject was a person or an AI system. The role of comparability also emerged as a shared aspect, though the basis differed: while interpersonal trust relied on perceived similarity to oneself, trust in AI systems was linked to resemblance to previously trusted technologies. Finally, the shared significance of past positive experiences underscores the broader importance of familiarity and reliability in fostering trust, regardless of the agent.

In conclusion, the findings demonstrate that while the social trust factors were derived from human interaction models, they do not fully translate to AI systems. The high number of significant differences between the two contexts suggests that trust in AI must be conceptualized differently. Nonetheless, the similarities identified in several areas highlight that past experiences, transparent reasoning, and performance-based reliability remain central pillars in the formation of trust, regardless of whether the counterpart is human or machine.

Conclusion

This study has shown that trust in AI systems arises from a multifaceted interplay of technical, social, and evaluation-related factors, each of which varies in relevance depending on the specific system context. Stability, applicability, and accuracy emerged as the most critical technical factors, laying the groundwork for functional confidence in AI. Among the social factors, reliability, accountability, and competence were identified as key, while evaluation factors such as trust in developers, firsthand experience, and credible testing reports strongly influenced user perception. Notably, trust is not distributed evenly across these categories—nor across AI system types. For instance, participants held safety-critical systems like autonomous vehicles to high standards of stability and precision, whereas in non-safety-critical systems such as chatbots, social and behavioral traits like respectful interaction and ethical alignment were more heavily weighted.

Context, therefore, proves to be a decisive factor in the formation of trust. The comparative analysis between chatbots and autonomous vehicles confirmed that user expectations shift significantly depending on perceived risk, familiarity, and system function. This reinforces the conclusion that a one-size-fits-all hierarchy of trust factors is neither realistic nor practical. Rather, the proposed Trust Pyramid serves best as a flexible, categorically organized model, guiding the context-sensitive assessment of trust.

Furthermore, the investigation into gender differences revealed only minimal variance, suggesting that while individual characteristics can play a role, gender alone does not substantially reshape trust perceptions. Similarly, the comparison between trust in humans and AI systems illustrated both overlap and divergence: while factors such as reliability, transparency, and traceability were valued in both domains, AI systems were not evaluated according to the same interpersonal criteria used for human trust. This underscores the need for tailored trust models that reflect the distinct nature of AI as a socio-technical agent.

Taken together, the study provides a foundational understanding of how different trust dimensions operate across AI contexts. Building on these findings, several promising avenues for future research have been identified. Further exploration should be directed at how trust hierarchies shift across varied AI system types, with particular attention to behavioral expectations and the concept of transferability. For instance, anthropomorphic traits may be beneficial in communication-focused AI but detrimental in safety-critical applications, a hypothesis that warrants empirical testing.

Additionally, the role of Explainable AI should be examined in greater depth. The current results suggest that users appreciate the option to access explanations rather than a constant flow of technical details. Future research could investigate the optimal balance between transparency and cognitive load, perhaps by testing different explanation models or user-triggered information mechanisms.

Finally, expanding the scope to include different age groups and user profiles may reveal further variations in trust formation. Since trust is inherently subjective, these nuances could guide more inclusive and adaptive AI design strategies.

AI systems should not adopt a uniform design strategy for trust-building. Developers must tailor system attributes, especially transparency, interaction behavior,

and explanation mechanisms, to the specific context in which the system will operate. For safety-critical systems, technical reliability, precision, and traceability should be emphasized. In contrast, for non-safety-critical systems, ethical behavior, respectful interaction, and value alignment become more salient.

The study found that users prefer to access explanations rather than be inundated with them. Developers should investigate implementing explainability as an on-demand feature, allowing users to trigger clarification when needed, thus reducing cognitive overload without sacrificing transparency.

By refining our understanding of trust in different contexts of use, this research presents an important step toward the development of more transparent, reliable, and user-aligned AI systems, ones that not only perform well but are actively trusted by those they aim to serve.

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.

Baron S (2024) Explainability and the limits of trust: A critical evaluation of XAI metrics. *Philosophy & Technology* 38(4).

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. Luxembourg.

European Commission: Regulation (EU) 2024/1689 of the European Parliament and of the Council (2024) European Artificial Intelligence Act (EU AI Act). *Official Journal of the European Union.* Publications Office. Luxembourg.

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.

Gu C, Zhang Y, Zeng L (2024) Exploring the mechanism of sustained consumer trust in AI chatbots after service failures: a perspective based on attribution and CASA theories. *Humanities and Social Sciences Communications* 11(1400).

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 KA, 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 P A (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.

Li Y, Wu B, Huang Y, Luan S (2024) A multidimensional framework of human–AI trust: Agent, receiver, and context. *Frontiers in Psychology*: 15.

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

McKnight DH, Carter M, Thatcher JB, Clay P (2011) Trust in a specific technology: An Investigation of its Components and Measures. *ACM Transactions on Management Information* Systems 2(2): 12-32.

Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. *Academy of Management Review* 20(3): 709–734.

Organisation for Economic Cooperation and Development (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.

Yardi MY (2022) ACM, Ethics, and Corporate Behavior. *Communications of the ACM* 65(3): 5.