Scoping Review

Matteo Pallomo, Eva Ledovskaia, Andrea Barbantani, Alessio Quaranta, Marco Farano, Elena Ruberto, Fiamma Pia Paternò, Michele Barale, Federico Ciociola, Flora D'Angelo.
Politecnico di Torino
Turin, Italy - 2026

1 Abstract

As artificial intelligence (AI) becomes increasingly embedded in professional workflows, understanding how humans develop and calibrate trust in these systems is critical. This paper presents a scoping review of the literature on trust in human–AI interaction, aiming to identify the key factors that determine whether trust is appropriately calibrated to a system's actual capabilities. A systematic screening process, applied to highly cited publications from 2015–2026, resulted in a final set of 47 relevant studies.

From this analysis, we derive a general, problem-agnostic framework that conceptualizes trust as a dynamic alignment between objective system properties and users' subjective perceptions. The framework is structured into five interdependent dimensions: Objective System Properties, Subjective Perception, Human Traits, Contextual Factors, and Adaptive Trust Dynamics. Together, these dimensions capture both the static and evolving nature of trust, including feedback loops, calibration processes, and temporal effects.

To demonstrate its applicability, the framework is applied to a use case in Advanced Driver Assistance Systems (ADAS) engineering, highlighting how trust is shaped in safety-critical environments. The results emphasize that trust extends beyond technical performance, emerging from a complex interplay of cognitive, individual, and contextual factors. This work provides both a conceptual foundation and a practical lens for analyzing and designing trust in AI-enabled systems.

2 Introduction

As Artificial Intelligence becomes increasingly integrated into the workplace, evolving from a support tool to an integral component of work, it is reshaping how tasks are performed and problems are solved. A key factor in this transformation is the relationship between workers and AI systems, particularly how *trust* is established and maintained.

This work aims to identify *which factors determine whether a worker's trust in an AI system is well-calibrated for a given task*. To this end, we systematically cluster the most relevant factors shaping trust between a *worker* and an *AI system augmenting or automating the worker's task*. This distinction is important, as trust is not a static construct: evidence shows that it evolves through cognitive, emotional, and organizational pathways, and sustained use depends on more than technical performance alone.

Our goal is to develop a general (i.e., problem-agnostic) framework to assess trust between these components. To demonstrate its practical relevance, we subsequently apply the framework to a specific use case.

3 Methods

In order to review current literature we decided to analyze the key keywords of our research question and then, based on those keywords, to run a query on *Scopus* to retrieve the results. The full query is given in *Appendix A*, however it is worth noticing the following remarks:

- The query was tailored to provide results between 2015 and 2026.
- The query was tailored to only provide results in English language.
- The query was tailored to only return the following types of publications: *articles, conference papers, book chapters, reviews, conference reviews and short surveys.*

The query returned approximately 11,000 results. To obtain a manageable subset for initial screening, results were sorted in descending order of number citations. From this ranked list, the first 600 papers were selected. This threshold corresponded approximately to the point at which publications had fewer than 100 citations, thereby prioritizing studies with higher academic impact. The selected records were exported into CSV document and brought to a shared spreadsheet. We firstly researched duplicates papers. This resulted in the removal of a single paper (same publication appearing under different journals), resulting in 599 unique papers to start our selection process from.

Following the duplicate removal, we proceeded with a title screening phase. The dataset was divided among 10 reviewers, each assessing approximately 60 papers. In this phase, our inclusion/exclusion criteria were the following:

- Titles where trust did not appear to be a primary or central concept.
- Titles that were overly domain-specific (e.g., focused exclusively on marketing, healthcare, or other narrow applications).
- Titles suggesting a purely technical focus without a human–AI interaction component.

This process reduced the number of candidate papers to 265.

Subsequently, the remaining papers underwent abstract screening. In this phase, all reviewers contributed collaboratively using a shared spreadsheet. A locking mechanism was adopted, whereby each paper was reviewed by a single researcher at a time, who recorded their decision in a dedicated column. Although this approach does not ensure independent double screening, it enabled efficient allocation of effort across the team.

In this phase, the following overall inclusion/exclusion criteria were defined:

- Inclusion of studies explicitly addressing trust in AI, automation, or human–AI interaction.

- Exclusion of studies lacking a clear human-centered perspective (e.g., purely algorithmic or technical contributions).
- Exclusion of studies where trust was only marginally mentioned and not analytically developed.

Papers that met the inclusion criteria after abstract screening were retained and constituted the final set of studies included in the review. Whilst reviewing each paper we performed the following actions:

- Mark the paper as *theoretical, empirical or practical.*
- Give each paper a unique reference (internal to the group). This enabled to automate some processes (*e.g.* citations references generation).

The table 1 summarizes the overall results (and a PRISMA flowchart can be found in Appendix A):

Stage	Pass	Fail
Total Papers	599 (100%)	
Title Screening	266 (44.4%)	333 (55.6%)
Abstract Screening	47 (17.7%)$^{+}$	218 (82.3%)$^{+}$
Final Classification (n = 47)		
Empirical	24 (51.1%)*	
Theoretical	22 (46.8%)*	
Practical	1 (2.1%)*	

$^{+}$ Percentages for abstract screening are relative to papers after title screening ($n = 265$).

* Percentages are relative to the final selected papers ($n = 47$).

Table 1: Paper Selection and Classification Summary

During the review process each researchers noted down keywords and definitions from the paper he/she was reviewing. When the process was complete, a framework clearly emerged from the interaction of those keywords and concepts. The framework will be the scope of our next sections.

4 Framework Overview

At its core, the framework conceptualizes trust as a *dynamic alignment problem*. Specifically, trust is defined as the degree to which a user's subjective perception of an AI system corresponds to the system's actual, objective capabilities. This condition is commonly referred to as *trust calibration* [4, 31]. Misalignment leads to overtrust or undertrust, both of which can degrade performance and safety.

This alignment is not fixed, but evolves through interaction, feedback, and learning. Users continuously update their mental models based on observed system behavior, forming a feedback loop between experience and expectation [11, 43].

The framework is organized into five interdependent macro-categories. *Objective System Properties* properties represent the ground truth of the system's capabilities. *Subjective Perception* captures how these capabilities are interpreted by users. *Human Traits* influence how perception is formed, while *Context* shapes both the relevance and consequences of trust. Finally, *Adaptive Trust Dynamics* describes how trust evolves over time through calibration, feedback, and learning mechanisms [3, 5].

A summary of the framework is given in figure 1, and table 2. Moreover, in *Appendix B*, the full summary table can be found.

Figure 1: Framework Summary — Four main dimensions are involved in the formation of trust: *Objective System Properties, Subjective Perception, Human Traits, and Context.* These dimensions evolve over time through *Adaptive Trust Dynamics.*

5 Objective System Properties

Objective system properties correspond to measurable and verifiable characteristics of the AI system. These properties define the system's actual trustworthiness independently of user perception.

A primary dimension is *performance*, which includes correctness, reliability, error rate, efficiency, functionality, and output quality [3, 5, 17, 34, 38]. These factors determine whether the system can achieve its intended task accurately and consistently.

Beyond raw performance, *technical behavior* plays a critical role. Consistency, predictability, and responsivity enable users to form stable expectations and reduce uncertainty [15, 19, 31]. Systems that behave in a stable and interpretable manner facilitate trust calibration by making their behavior easier to anticipate.

Additional mechanisms such as explicit communication of uncertainty and system confidence provide important calibration cues, helping users understand when reliance is appropriate and when caution is required [18, 36, 46].

Finally, trust is strongly influenced by *governance and ethical properties.* Privacy, integrity, fairness, and institutional safeguards extend trust beyond technical performance into broader socio-technical considerations [5, 23, 26, 28]. Data integrity and robustness are also critical, as biased or unreliable data sources undermine system credibility [2, 44].

Layer	Key Factors	Interpretation
Objective System Properties	Performance; Technical Behavior; Governance & Ethics	Represents the system's actual capabilities and trustworthiness, including correctness, reliability, consistency, transparency, privacy, and institutional safeguards.
Subjective Perception	Perceived Performance; Interpretability; Social Perception	Captures how users interpret the system, including perceived usefulness and ease of use, explainability and transparency, as well as social and affective responses such as credibility, empathy, and bias.
Human Traits	Disposition; Experience; Personality	Individual characteristics (e.g., trust propensity, expertise, familiarity, technophobia) that influence how objective evidence is interpreted and translated into trust.
Contextual Factors	Task; Risk; Environment; Governance Context	Situational conditions shaping trust, including task criticality, uncertainty, human oversight, organizational constraints, and broader institutional trust.
Adaptive Trust Dynamics	Calibration; Feedback; Learning; Temporal Evolution	Describes how trust evolves over time through interaction, including feedback loops, trust calibration, asymmetry effects, and mechanisms such as trust transfer and mediation.

Table 2: Summary of trust framework dimensions and their roles

6 Subjective Perception

While objective properties define what the system is capable of, trust ultimately depends on how these capabilities are *perceived* by users. Subjective perception encompasses cognitive evaluation, interpretation, and affective response.

A key component is *perceived performance*. Constructs such as perceived usefulness, perceived ease of use, and confidence directly influence willingness to rely on AI systems [3, 5, 17, 42]. Importantly, perceived performance may diverge from actual performance, leading to miscalibrated trust.

Interpretability is another central dimension. Transparency, explainability, understandability, and controllability allow users to make sense of system outputs and decisions, reducing ambiguity and supporting informed reliance [5, 25, 37, 38]. Perceived goal alignment further strengthens trust by ensuring that users believe the system operates according to their intentions [38].

In addition, *social and affective perception* plays a significant role. Anthropomorphism, perceived benevolence, credibility, and emotional responses influence trust independently of objective performance [5, 7, 15, 47]. Perceived social biases and relatability further shape how users evaluate AI systems [38, 47]. These elements highlight that trust is not purely rational, but is profoundly influenced by social interpretation and emotional involvement. [4, 38, 47]

Finally, subjective perception directly influences *behavioral outcomes*, such as intention to use and actual system usage, highlighting its central role in determining effective human–AI interaction [3].

7 Human Traits

Human traits represent relatively stable individual characteristics that influence how users perceive and interact with AI systems. These traits act as a cognitive filter through which objective system properties are interpreted.

Key factors include trust propensity, trusting stance, technophobia, expertise, and familiarity [7, 16, 29, 34, 39]. Individuals with a higher propensity to trust or a positive attitude toward technology are more likely to rely on AI systems, whereas technophobia or, in general, a limited propensity towards technology, introduces skepticism and resistance.

Experience and expertise are particularly important for trust calibration. Users with higher competence are better able to assess system limitations and interpret outputs critically, reducing the likelihood of overreliance [2, 45]. Conversely, limited expertise may lead to blind trust or excessive caution.

Importantly, these traits do not directly determine trust, but influence how objective evidence is translated into subjective perception.

8 Contextual Factors

Trust is inherently context-dependent. The same system may be trusted or distrusted depending on task characteristics, risk levels, and environmental conditions.

Task characteristics and criticality strongly influence trust decisions. Users tend to rely more on AI in structured and analytical tasks, while exhibiting greater caution in high-stakes or socially complex scenarios [25, 30]. Trust is closely linked to vulnerability and risk, as reliance on AI exposes users to potential negative consequences [4, 11].

Other factors such as environmental and organizational may further shape trust. For example, these may include the presence of human oversight, facilitating conditions, readiness mismatches, and perceived threats to professional identity [7, 8, 47].

Broader socio-economic and institutional conditions also play a role. Governance frameworks, corporate trust, and societal perceptions of AI influence baseline trust levels and adoption willingness [5, 44]. Additionally, concerns such as overreliance risk and human–AI substitution introduce context-specific trust dynamics, particularly in safety-critical or high-responsibility domains [3, 28].

9 Adaptive Trust Dynamics

Trust is not static but evolves over time through interaction. This macro-category captures the mechanisms governing the temporal evolution of trust.

At the center of this process lies *trust calibration*, defined as the alignment between perceived and actual system performance [4]. Calibration is supported by feedback loops, where interaction outcomes continuously update user beliefs [11, 43].

This process is inherently asymmetric: failures have a stronger negative impact on trust than successful interactions have in increasing it [6]. As a result, trust can degrade rapidly after errors, while recovery requires sustained evidence of reliability.

Over time, trust evolves through mechanisms such as trust repair, trust dampening, and broader temporal adaptation [9, 11]. Users gradually refine their mental models, improving their ability to calibrate reliance decisions.

Additional mechanisms further enrich this dynamic perspective. Trust can transfer from related entities (e.g., developers or institutions) to AI systems [5]. Moreover, trust often influences behavior through mediation processes, where it shapes intention, which in turn drives actual usage [3].

Finally, system-side adaptation also contributes to trust dynamics. Continuous learning and performance evolution modify future interactions, further influencing how trust develops over time [13].

10 Worked Use Case

10.1 Application to ADAS System Engineering

We apply the proposed framework to the context of a *System Engineer* working on *Advanced Driver Assistance Systems (ADAS)*, focusing on an AI-based tool supporting requirement generation and traceability.

Objective System Properties. In this domain, trust is strongly grounded in measurable system properties. Core performance dimensions include **correctness**, **reliability**, **error rate**, and **output quality** of generated requirements [34, 38]. In particular:

- **Requirement correctness**: whether generated requirements accurately reflect system design intent. Incorrect outputs directly undermine trust and increase verification burden.
- **Traceability completeness**: whether requirements are properly linked to upstream and downstream artifacts (e.g., system elements, test cases). Missing links reduce auditability and certification readiness.
- **Requirement testability**: whether requirements are expressed in a way that enables objective validation through test cases.

Beyond performance, **consistency** and **predictability** are critical technical behavior properties [15, 31]. Engineers rely on stable outputs to form reliable workflows and mental models. Additionally, mechanisms such as **uncertainty communication** (e.g., confidence scores) can support trust calibration by indicating when outputs require additional scrutiny [36, 46].

Finally, **governance and compliance** aspects are central in ADAS. Concerns such as **data integrity**, **traceability transparency**, and **accountability** are critical for certification processes (e.g., ISO-26262, UNECE homologations, NCAP standards), and strongly influence trust independently of raw performance [2, 3].

Subjective Perception. Trust is mediated by how engineers perceive system outputs during interaction. **Perceived usefulness**, **confidence**, and **perceived ease of use** influence whether the system is considered practically valuable [3, 5, 42].

Several perception-specific factors are particularly relevant:

- **Readability**: clarity and structure of generated requirements. Poor readability increases cognitive load and reduces trust.
- **Alignment with expectations**: degree to which outputs match the engineer's mental model and design intent. Misalignment can lead to distrust even when outputs are objectively correct.
- **Explainability**: ability to understand why a requirement or trace link was generated. Limited explainability reduces willingness to rely on outputs [25].

Interpretability dimensions such as **transparency** and **understandability** further support trust by making system behavior cognitively accessible [5, 37].

Additionally, subjective perception influences **behavioral outcomes**, such as intention to use and actual reliance. Even high-performing systems may be underutilized if they are perceived as unreliable or difficult to interpret [3].

Human Traits. Individual characteristics influence how system behavior is interpreted. **Expertise**, **familiarity**, and **technical competence** enable engineers to better assess system outputs and limitations, supporting more accurate trust calibration [2, 29, 45].

Conversely, traits such as **technophobia** or low trust propensity may lead to conservative reliance, even when system performance is adequate [34, 39]. However, in the context of a System Engineer working in the Autonomous Driving field, might be highly unlikely to find people reluctant to use/accept autonomous technologies. Nevertheless, these traits might act as a filter between objective system properties and subjective perception, shaping how evidence is interpreted rather than directly determining trust.

Contextual Factors. ADAS development is inherently safety-critical, which amplifies the role of **risk** and **vulnerability** in trust decisions [11]. Engineers operate under strict validation and certification constraints, leading to low tolerance for errors.

Key contextual factors include:

- **Task criticality**: requirements directly impact safety-relevant functions (e.g., braking, steering), leading to conservative trust strategies [30].
- **Organizational processes**: strong validation cultures and formal verification workflows may limit reliance on AI-generated artifacts [47].
- **Human oversight**: the presence of a human-in-the-loop mitigates risk but may also signal limited system autonomy [8].

Broader concerns such as **professional accountability**, **certification acceptance**, and **automation-related job impact** further influence trust independently of system performance [3]. An engineer might be reluctant to generate outputs with an AI system because, his/her colleagues might judge him and lazy and/or incompetent, introducing also a layer of *social perception*.

Adaptive Trust Dynamics. Trust evolves through repeated interaction. Engineers continuously update their beliefs about system capabilities through **feedback loops**, where observed outcomes influence future reliance decisions [11, 43].

A critical property of this evolution is **asymmetry**: failures have a stronger negative impact on trust than successful interactions have in increasing it [6]. In safety-critical domains, even a single critical error can significantly reduce trust.

Relevant dynamic factors include:

- **Consistency and repeatability**: stable outputs enable the formation of accurate mental models [15].
- **Error experience**: unexpected or severe failures trigger rapid trust degradation.
- **Trust repair**: recovery from failures requires sustained evidence of reliability over time [11].

Over time, engineers develop calibrated trust, learning when to rely on the system and when to intervene. This process may also involve **trust transfer** from institutional or organizational trust, and **mediation effects** where trust influences intention and usage behavior [3, 5].

Finally, system-side evolution—such as continuous improvement of AI models—further shapes trust trajectories by modifying future interaction outcomes [13].

11 Conclusions

This work proposed a structured framework to analyze trust in human–AI interaction as a *calibration problem*, defined by the alignment between a system's objective capabilities and the user's subjective perception of those capabilities. By synthesizing evidence from the literature, we identified five interdependent macro-categories shaping this alignment: *Objective System Properties*, *Subjective Perception*, *Human Traits*, *Contextual Factors*, and *Adaptive Trust Dynamics*.

A key contribution of this work is the explicit integration of these dimensions into a unified model that captures both static and dynamic aspects of trust. In particular, the framework highlights that trust cannot be reduced to system performance alone, but emerges from the interaction between measurable properties, cognitive interpretation, individual differences, and situational constraints. Furthermore, by modeling trust as a temporal process, the framework accounts for feedback loops, asymmetries in trust evolution, and the progressive calibration of user expectations.

The application to an ADAS system engineering use case demonstrates the practical relevance of the framework. In safety-critical domains, trust is strongly influenced not only by objective performance metrics such as correctness and reliability, but also by interpretability, validation effort, and organizational constraints. The results show that even highly capable systems may be used conservatively when contextual risks and accountability concerns are high, reinforcing the need to consider trust as a socio-technical phenomenon.

Overall, this work contributes a generalizable and problem-agnostic perspective on trust in AI systems, providing both a conceptual foundation and a practical lens for analyzing trust calibration in real-world settings.

12 Gaps and Future Work

While the proposed framework provides a structured view of trust in human–AI interaction, several limitations and open research directions remain.

First, the framework is primarily conceptual and does not provide a formal or quantitative model of trust calibration. Future work should aim to operationalize the relationships between dimensions and develop measurable indicators to enable empirical validation.

Second, the literature selection introduces potential bias. Only English-language publications were considered, which may exclude relevant contributions from other research communities. Moreover, prioritizing highly cited articles risks systematically excluding more recent studies (e.g., published between 2024 and 2026) that have not yet accumulated sufficient citations—an important limitation in a rapidly evolving field such as artificial intelligence. Additionally, the analysis may reflect a *technical bias*, as the authors' background is predominantly in engineering, potentially emphasizing system-centric aspects of trust over social or organizational perspectives. Finally, only one practical source has been included in the review.

Furthermore, a subset of the analyzed literature focuses on *Autonomous Driving*. This introduces a potential domain bias that may skew the identified factors toward this application area. While this is beneficial for the ADAS use case, it may reduce the generality of the framework.

Finally, the framework has been illustrated through a single domain-specific use case (ADAS). Further validation across multiple domains and contexts is necessary to assess its generalizability and refine its applicability.

Addressing these limitations would support the transition from a descriptive framework to a more robust and generalizable model for analyzing and designing trust in AI systems.

13 Appendix A - Selection Process Details

13.1 Scopus Query

Query was run on: 15/04/2026 (dd/mm/yyy).

```
(
    "trust in artificial intelligence"
    OR "trust in automation"
    OR "trust calibration"
) AND
(
    LIMIT-TO ( DOCTYPE , "ar" )
    OR LIMIT-TO ( DOCTYPE , "cp" )
    OR LIMIT-TO ( DOCTYPE , "re" )
    OR LIMIT-TO ( DOCTYPE , "ch" )
    OR LIMIT-TO ( DOCTYPE , "cr" )
    OR LIMIT-TO ( DOCTYPE , "sh" )
) AND
(
    LIMIT-TO ( LANGUAGE , "English" )
) AND
PUBYEAR > 2015 AND PUBYEAR < 2026
```

13.2 PRISMA Diagram

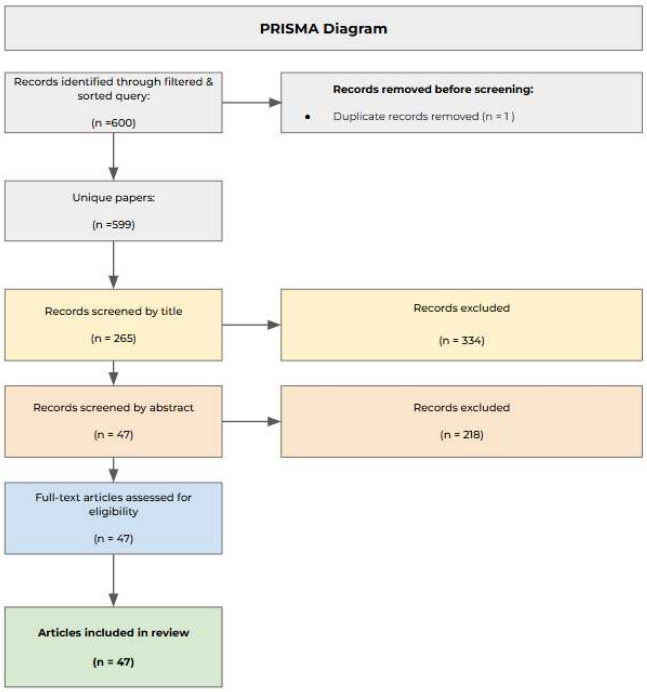

Figure 2: PRISMA Diagram — diagram for our selection process.

14 Appendix B - Summary Tables

Reference	Factor	Layer	Subdivision	Interpretation
[4]	Warranted Trust (Calibration)	Adaptive Trust Dynamics	Calibration	Alignment between perceived and actual trustworthiness
[31]	Calibration	Adaptive Trust Dynamics	Calibration	Understanding system capabilities for correct reliance
[11]	Feedback loop	Adaptive Trust Dynamics	Feedback	Outcomes influence future trust
[13]	Continuous learning	Adaptive Trust Dynamics	Learning	System adapts over time
[13]	Evolution	Adaptive Trust Dynamics	Learning	System evolves and improves performance
[9]	Static vs. Dynamic Trust	Adaptive Trust Dynamics	Temporal Evolution	Trust changes over time through interaction
[5]	Trust Transfer	Adaptive Trust Dynamics	Transfer Mechanism	Process by which trust in humans extends to AI systems they design
[3]	Mediation (Trust → Intent → Use)	Adaptive Trust Dynamics	Temporal Evolution	Process where intent to use mediates the effect of trust on actual usage behavior
[25]	Task Characteristics	Context	Task	Trust varies depending on task type
[4]	Vulnerability / Risk	Context	Risk	Trust involves exposure to potential failure
[11]	Vulnerability / Risk	Context	Risk	Trust involves exposure to potential failure
[8]	Human in the loop	Context	Environment	Presence of human oversight
[7]	Facilitating conditions	Context	Environment	Availability of supporting resources
[47]	Readiness Mismatch	Context	Environment	Mismatch between system and user readiness
[20]	Threat to Professional Identity	Context	Environment	Employees fear losing their jobs due to automation potential.
[1]	Threat to Professional Identity	Context	Environment	Employees fear losing their jobs due to automation potential.
[30]	Uncertainty Management	Context	Task	Trust helps workers manage the uncertainty of relying on imperfect AI suggestions in high-stakes tasks.
[5]	Corporate Distrust	Context	Governance Trust Context	Degree to which users distrust corporations responsible for AI systems
[3]	Vulnerability / Risk	Context	Risk	Awareness that relying on ChatGPT involves potential negative consequences especially in high-stakes contexts
[28]	Intent Framing (Purpose Communication)	Context	Environment	Effect of communicated benefits (e.g. 24/7 access, uniformity) on trust formation
[41]	User Propensity	Human Traits	Disposition	Baseline tendency to trust technology
[16]	Trusting stance	Human Traits	Disposition	General belief that technology leads to positive outcomes
[39]	Technophobia	Human Traits	Disposition	Fear or resistance toward AI
[7]	Familiarity	Human Traits	Experience	Experience reduces uncertainty and increases trust

Reference	Factor	Layer	Subdivision	Interpretation
[42]	First Hand Knowledge	Human Traits	Experience	Technical knowledge before interaction
[5]	Trusting stance	Human Traits	Disposition	General tendency of individuals to trust others and technological systems
[5]	User Expertise	Human Traits	Experience	Level of user knowledge or familiarity with technology influencing trust
[2]	Technical Expertise	Human Traits	Experience	The worker's level of competence influences his ability to calibrate trust (avoid overtrust).
[28]	Privacy	Objective Sys. Properties	Governance & Ethics	Protection of personal and sensitive data
[23]	Privacy	Objective Metrics	Governance & Ethics	Protection of personal and sensitive data
[38]	Privacy	Objective Sys. Properties	Governance & Ethics	Protection of personal and sensitive data
[38]	Trust through Governance	Objective Sys. Properties	Governance & Ethics	Institutional mechanisms ensuring reliability
[28]	Performance (Technical Competence)	Objective Sys. Properties	Performance	Ability of the chatbot to provide correct and useful answers to citizen enquiries
[28]	Process Transparency	Objective Sys. Properties	Governance & Ethics	Understanding of how the chatbot works and how decisions are generated
[38]	Goal Alignment	Subjective Perception	Interpretability	Perceived alignment between the goal of the trustoor and trustee.
[4]	Goal Alignment	Subjective Perception	Interpretability	Perceived alignment between the goal of the trustoor and trustee.
[38]	Transparency	Subjective Perception	Interpretability	Perceived visibility of system logic
[25]	Transparency	Subjective Perception	Interpretability	Perceived visibility of system logic
[10]	Transparency	Subjective Perception	Interpretability	Perceived visibility of system logic
[24]	Explainability	Subjective Perception	Interpretability	Ability of system to explain decisions
[12]	Explainability	Subjective Perception	Interpretability	Ability of system to explain decisions
[38]	Explainability	Subjective Perception	Interpretability	Ability of system to explain decisions
[27]	Explainability	Subjective Perception	Interpretability	Ability of system to explain decisions
[37]	Understandability	Subjective Perception	Interpretability	Clarity of explanations from user perspective
[33]	Understandability	Subjective Perception	Interpretability	Clarity of explanations from user perspective
[42]	Perceived usefulness (PU)	Subjective Perception	Perceived Performance	Belief system improves job performance
[42]	Perceived ease of use (PEOU)	Subjective Perception	Perceived Performance	Perceived effort required to use system

Reference	Factor	Layer	Subdivision	Interpretation
[17]	Confidence	Subjective Perception	Perceived Performance	User confidence in system outputs
[7]	Anthropomorphism	Subjective Perception	Social Perception	Perception of human-like characteristics
[34]	Agent Personality	Subjective Perception	Social Perception	Perceived personality traits
[15]	Benevolence	Subjective Perception	Social Perception	Perceived goodwill of the system
[38]	Benevolence	Subjective Perception	Social Perception	Perceived goodwill of the system
[38]	Social Bias	Subjective Perception	Social Perception	Perceived social biases of the system.
[47]	Relatability	Subjective Perception	Social Perception	User emotional connection with system
[47]	Perceived credibility	Subjective Perception	Social Perception	Perceived expertise and honesty
[11]	Relationship Equity	Subjective Perception	Social Perception	is an emotional resource that predicts the degree of goodwill between two actors
[32]	Safety Beliefs	Subjective Perception	Cognitive	Pedestrians and cyclists intellectually associate AV adoption with improved road safety and better traffic flow.
[32]	Interaction Anxiety	Subjective Perception	Affective	Direct interaction experience with AVs significantly reduces user fear and increases emotional comfort levels.
[19]	Responsivity	Objective Sys. Properties	Technical Behavior	Adaptation to environment and user cues
[6]	Asymmetric Reliability Impact	Adaptive Trust Dynamics	Asymmetry	Failures reduce trust more than success increases it
[6]	Human Propensity to Trust	Human Traits	Disposition	Internal user traits and prior experience with technology serve as a baseline for initial trust before any interaction with the specific AI occurs.
[43]	Feedback loop	Adaptive Trust Dynamics	Feedback	Outcomes influence future trust
[46]	Explainability	Subjective Perception	Interpretability	Ability of system to explain decisions
[26]	Trust through Governance	Objective Sys. Properties	Governance & Ethics	Institutional mechanisms ensuring reliability
[22]	AI-Human Teaming Dynamics	Context	Governance Trust Context	Trust shifts from interpersonal reliability to system-based reliability when AI is integrated as a functional member of a professional team.
[45]	Domain Expertise	Human Traits	Experience	A worker's domain expertise determines how effectively AI explanations can help them calibrate trust.
[44]	Technological Integrity	Objective Sys. Propertiess	Technical Behavior	Data biases and privacy vulnerabilities act as primary technical barriers to establishing system trust.

Reference	Factor	Layer	Subdivision	Interpretation
[44]	Algorithmic Transparency	Objective Sys. Properties	Governance & Ethics	Algorithmic "black boxes" hinder cognitive trust by reducing system explainability and predictability.
[44]	Psychological Barrier	Human Traits	Disposition	Fear of job displacement and perceived lack of accountability directly diminish affective trust in AI.
[44]	Socio-Economic Divide	Subjective Perception	Social Perception	Global wealth inequality and the digital divide foster systemic distrust toward AI-driven institutions.
[44]	Institutional Governance	Objective Metrics	Technical Behavior	Regulatory frameworks and ethical codes serve as essential environmental safeguards that facilitate public trust.
[21]	Correctness	Objective Sys. Properties	Performance	Objective correctness of the generated output.
[35]	Emotional Accuracy	Subjective Perception	Performance	Social Perception recognize emotions in artificial agents through body movements and postures alone
[18]	System Reliability	Objective Sys. Properties	Performance	Reliability varies greatly depending on the initial conditions.
[18]	Trust Calibration Cues	Adaptive Trust Dynamics	Calibration	Specific cognitive signals used as an active feedback mechanism to push the user to recalibrate their trust
[29]	User Expertise	Human Traits	Experience	Prior knowledge and experience with system
[14]	Proactive Personality	Human Traits	Disposition	A worker's proactive personality acts as a key moderator that, alongside trust in AI, allows them to leverage AI-supported autonomy into innovative performance
[36]	Transparency	Subjective Perception	Interpretability	Visual verification of the system's environmental perception.
[40]	Perceived Empathy	Subjective Perception	Social Perception	Perceived ability of the AI to understand human emotional and contextual nuances.

Table 3: Trust framework mapping

15 Appendix C: AI usage disclosure

In the writing of this document, the following LLM have been used to help the team. In the following table it is possible to find which LLM were used and for what purpose, please reference 4:

LLM Name	Usages
ChatGPT	• LaTeX formatting help. • Papers summarization. • Image generation. • Document writing (i.e. text enhancement). • Python automations.
Gemini	• LaTeX formatting help. • Document writing (i.e. text enhancement).

Table 4: Overview of LLMs and Their Usages

16 Appendix D: Team Members & Contacts & Useful Links

Name Surname	Student ID	Email
Matteo Pallomo	**s337855**	**s337855@studenti.polito.it**
Eva Ledovskaia	s310269	s310269@studenti.polito.it
Andrea Barbantani	s342878	s342878@studenti.polito.it
Alessio Quaranta	s351986	s351986@studenti.polito.it
Marco Farano	s337643	s337643@studenti.polito.it
Elena Ruberto	s336356	s336356@studenti.polito.it
Fiamma Pia Paternò	s342654	s342654@studenti.polito.it
Michele Barale	s341813	s341813@studenti.polito.it
Federico Ciociola	s341238	s341238@studenti.polito.it
Flora D'Angelo	s336616	s336616@studenti.polito.it

Table 5: Team Members - In bold the name of the team leader.

All materials used for the development of this document is available at this Git Repository: Submission Package Link. The access to this repository is *public*, so it should accessible to anybody with the link. In case of issue, please contact one of the team leader.

References

[1] Arslan A., Cooper C., Khan Z., Golgeci I., and Ali I. 2022. Artificial intelligence and human workers interaction at team level: a conceptual assessment of the challenges and potential HRM strategies. *International Journal of Manpower* 43, 1 (2022), 75–88. doi:10.1108/IJM-01-2021-0052

[2] Aldoseri A., Al-Khalifa K.N., and Hamouda A.M. 2023. Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Applied Sciences (Switzerland)* 13, 12 (2023). doi:10.3390/app13127082

[3] Choudhury A. and Shamszare H. 2023. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research* 25 (2023). doi:10.2196/47184

[4] Jacovi A., Marasović A., Miller T., and Goldberg Y. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 624–635. doi:10.1145/3442188.3445923

[5] Schepman A. and Rodway P. 2023. The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human-Computer Interaction* 39, 13 (2023), 2724–2741. doi:10.1080/10447318.2022.2085400

[6] Kaplan A.D., Kessler T.T., Brill J.C., and Hancock P.A. 2023. Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors* 65, 2 (2023), 337–359. doi:10.1177/00187208211013988

[7] Albahri A.S., Duhaim A.M., Fadhel M.A., Alnoor A., Baqer N.S., Alzubaidi L., Albahri O.S., Alamoodi A.H., Bai J., Salhi A., Santamaría J., Ouyang C., Gupta A., Gu Y., and Deveci M. 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion* 96 (2023), 156–191. doi:10.1016/j.inffus.2023.03.008

[8] Young A.T., Amara D., Bhattacharya A., and Wei M.L. 2021. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *The Lancet Digital Health* 3, 9 (2021), e599–e611. doi:10.1016/S2589-7500(21)00132-1

[9] Zhang A.X., Muller M., and Wang D. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020). doi:10.1145/3392826

[10] Liu B. 2021. In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human-AI Interaction. *Journal of Computer-Mediated Communication* 26, 6 (2021), 384–402. doi:10.1093/jcmc/zmab013

[11] de Visser E.J., Peeters M.M.M., Jung M.F., Kohn S., Shaw T.H., Pak R., and Neerincx M.A. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12, 2 (2020), 459–478. doi:10.1007/s12369-019-00596-x

[12] Vilone G. and Longo L. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106. doi:10.1016/j.inffus.2021.05.009

[13] Kim H., So K.K.F., and Wirtz J. 2022. Service robots: Applying social exchange theory to better understand human–robot interactions. *Tourism Management* 92 (2022). doi:10.1016/j.tourman.2022.104537

[14] Kong H., Yin Z., Chon K., Yuan Y., and Yu J. 2024. How does artificial intelligence (AI) enhance hospitality employee innovation? The roles of exploration, AI trust, and proactive personality. *Journal of Hospitality Marketing and Management* 33, 3 (2024), 261–287. doi:10.1080/19368623.2023.2258116

[15] Zhang H., Wu B., Yuan X., Pan S., Tong H., and Pei J. 2024. Trustworthy Graph Neural Networks: Aspects, Methods, and Trends. *Proc. IEEE* 112, 2 (2024), 97–139. doi:10.1109/JPROC.2024.3369017

[16] Tussyadiah I.P., Zach F.J., and Wang J. 2020. Do travelers trust intelligent service robots? *Annals of Tourism Research* 81 (2020). doi:10.1016/j.annals.2020.102886

[17] Dudley J.J. and Kristensson P.O. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (2018). doi:10.1145/3185517

[18] Okamura K. and Yamada S. 2020. Adaptive trust calibration for human-AI collaboration. *PLoS ONE* 15, 2 (2020). doi:10.1371/journal.pone.0229132

[19] Yam K.C., Tang P.M., Jackson J.C., Su R., and Gray K. 2022. The Rise of Robots Increases Job Insecurity and Maladaptive Workplace Behaviors: Multimethod Evidence. *Journal of Applied Psychology* 108, 5 (2022), 850–870. doi:10.1037/apl0001045

[20] Buckley L., Kaye S.-A., and Pradhan A.K. 2018. Psychosocial factors associated with intended use of automated vehicles: A simulated driving study. *Accident Analysis and Prevention* 115 (2018), 202–208. doi:10.1016/j.aap.2018.03.021

[21] Chong L., Zhang G., Goucher-Lambert K., Kotovsky K., and Cagan J. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022). doi:10.1016/j.chb.2021.107018

[22] Larson L. and DeChurch L.A. 2020. Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *Leadership Quarterly* 31, 1 (2020). doi:10.1016/j.leaqua.2019.101377

[23] Nazar M., Alam M.M., Yafi E., and Su'Ud M.M. 2021. A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques. *IEEE Access* 9 (2021), 153316–153348. doi:10.1109/ACCESS.2021.3127881

[24] Rubagotti M., Tusseyeva I., Baltabayeva S., Summers D., and Sandygulova A. 2022. Perceived safety in physical human–robot interaction—A survey. *Robotics and Autonomous Systems* 151 (2022). doi:10.1016/j.robot.2022.104047

[25] Mathur M.B. and Reichling D.B. 2016. Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition* 146 (2016), 22–32. doi:10.1016/j.cognition.2015.09.008

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

[27] De Graaf M.M.A. and Malle B.F. 2017. How people explain action (and autonomous intelligent systems should too). *AAAI Fall Symposium - Technical Report* FS-17-01 - FS-17-05 (2017), 19–26.

[28] Aoki N. 2020. An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly* 37, 4 (2020). doi:10.1016/j.giq.2020.101490

[29] Martinez-Martin N., Luo Z., Kaushal A., Adeli E., Haque A., Kelly S.S., Wieten S., Cho M.K., Magnus D., Fei-Fei L., Schulman K., and Milstein A. 2021. Ethical issues in using ambient intelligence in health-care settings. *The Lancet Digital Health* 3, 2 (2021), e115–e123. doi:10.1016/S2589-7500(20)30275-2

[30] Asan O., Bayrak A.E., and Choudhury A. 2020. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research* 22, 6 (2020). doi:10.2196/15154

[31] Carsten O. and Martens M.H. 2019. How can humans understand their automated cars? HMI principles, problems and solutions. *Cognition, Technology and Work* 21, 1 (2019), 3–20. doi:10.1007/s10111-018-0484-0

[32] Penmetsa P., Adanu E.K., Wood D., Wang T., and Jones S.L. 2019. Perceptions and expectations of autonomous vehicles – A snapshot of vulnerable road user opinion. *Technological Forecasting and Social Change* 143 (2019), 9–13. doi:10.1016/j.techfore.2019.02.010

[33] Vaithilingam P., Zhang T., and Glassman E.L. 2022. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. *Conference on Human Factors in Computing Systems - Proceedings* (2022). doi:10.1145/3491101.3519665

[34] Binns R., Van Kleek M., Veale M., Lyngs U., Zhao J., and Shadbolt N. 2018. 'It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018). doi:10.1145/3173574.3173951

[35] Hortensius R., Hekele F., and Cross E.S. 2018. The Perception of Emotion in Artificial Agents. *IEEE Transactions on Cognitive and Developmental Systems* 10, 4 (2018), 852–864. doi:10.1109/TCDS.2018.2826921

[36] Haeuslschmid R., Von Buelow M., Pfleging B., and Butz A. 2017. Supporting trust in autonomous driving. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2017), 319–329. doi:10.1145/3025171.3025198

[37] Iyer R., Li Y., Li H., Lewis M., Sundar R., and Sycara K. 2018. Transparency and Explanation in Deep Reinforcement Learning Neural Networks. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018), 144–150. doi:10.1145/3278721.3278776

[38] Afroogh S., Akbari A., Malone E., Kargar M., and Alambeigi H. 2024. Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications* 11, 1 (2024). doi:10.1057/s41599-024-04044-8

[39] Choi S., Jang Y., and Kim H. 2023. Influence of Pedagogical Beliefs and Perceived Trust on Teachers' Acceptance of Educational Artificial Intelligence Tools. *International Journal of Human-Computer Interaction* 39, 4 (2023), 910–922. doi:10.1080/10447318.2022.2049145

[40] Tong S., Jia N., Luo X., and Fang Z. 2021. The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal* 42, 9 (2021), 1600–1631. doi:10.1002/smj.3322

[41] You S., Yang C.L., and Li X. 2022. Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems* 39, 2 (2022), 336–365. doi:10.1080/07421222.2022.2063553

[42] Zhang T., Tao D., Qu X., Zhang X., Zeng J., Zhu H., and Zhu H. 2020. Automated vehicle acceptance in China: Social influence and initial trust are key determinants. *Transportation Research Part C: Emerging Technologies* 112 (2020), 220–233. doi:10.1016/j.trc.2020.01.027

[43] Bach T.A., Khan A., Hallock H., Beltrão G., and Sousa S. 2024. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human-Computer Interaction* 40, 5 (2024), 1251–1266. doi:10.1080/10447318.2022.2138826

[44] Wang W. and Siau K. 2019. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management* 30, 1 (2019), 61–79. doi:10.4018/JDM.2019010104

[45] Wang X. and Yin M. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2021), 318–328. doi:10.1145/3397481.3450650

[46] Buçinca Z., Lin P., Gajos K.Z., and Glassman E.L. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2020), 454–464. doi:10.1145/3377325.3377498

[47] Xu Z., Zhang K., Min H., Wang Z., Zhao X., and Liu P. 2018. What drives people to accept automated vehicles? Findings from a field experiment. *Transportation Research Part C: Emerging Technologies* 95 (2018), 320–334. doi:10.1016/j.trc.2018.07.024