International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Three Layers of Trust in AI Interfaces: Interface, Behavior, and Organization

Ravi Palwe

Abstract: ***As artificial intelligence (AI) systems become deeply integrated into sectors such as healthcare, finance, and autonomous services, establishing and maintaining user trust has emerged as a critical determinant of system adoption and sustained engagement. Despite increasing attention to trust in AI, existing models often treat it as a singular construct, neglecting the multifaceted ways in which user interfaces shape the perception and evolution of trust. This paper introduces a novel three-layered model of trust in AI interfaces- Interface Trust, Behavior Trust, and Organizational Trust-to reflect the dynamic and layered nature of trust formation. Drawing upon interdisciplinary literature and domain-specific case studies in healthcare and finance, we demonstrate how the AI interface serves as the primary channel through which users engage with system logic, evaluate behavior, and perceive organizational credibility. Through this model, we argue that trust begins with interface design and transparency, is sustained through reliable AI behavior, and is ultimately anchored in the perceived integrity and accountability of the organization behind the system. This paper provides theoretical foundations and practical design guidelines to foster trust across these three layers, offering insights into how trust can be engineered, measured, and maintained throughout the lifecycle of AI-human interaction.***

Keywords: Trust in AI, Human-AI Interaction, Interface Design, Behavioral Trust, Organizational Trust, Explainable AI, UX in AI Systems, AI Ethics, Transparency, User-Centered Design

1. Introduction

The integration of artificial intelligence (AI) systems into critical domains such as healthcare, finance, and autonomous technologies has significantly transformed how decisions are made, information is processed, and services are delivered. As AI becomes increasingly responsible for supporting high-stakes decisions-ranging from clinical diagnoses to financial risk assessments-user **trust** emerges as a vital factor in determining the acceptance, adoption, and ethical deployment of these systems (Glikson & Woolley, 2020; Henrique et al., 2024).

However, despite a growing body of literature on trust in AI and algorithmic transparency, much of the existing work tends to conceptualize trust as a **singular or static construct**, often decoupled from the human–computer interaction (HCI) interface through which users actually experience AI (Ueno et al., 2022; Yang & Wibowo, 2022). Such approaches overlook the fact that trust in AI systems is **not solely determined by the underlying algorithms or models**, but is often initiated and shaped by how the AI system is presented, interacted with, and explained through its **interface** (Ribes Lemay et al., 2021; Böckle et al., 2021). In this context, the **user interface becomes the first site of cognitive appraisal and emotional response**, forming the foundation upon which deeper behavioral and organizational trust can develop- or erode.

This paper introduces a **three-layered model of trust in AI interfaces** to address this oversight and better reflect the dynamic, multi-dimensional nature of trust development. The model includes:

- **Interface Trust**- the trust formed through interaction design, usability, and transparency in the user interface (Zerilli et al., 2022);
- **Behavior Trust**- the trust derived from the perceived performance, consistency, and intelligibility of the AI's decisions (Papenmeier et al., 2019);
- **Organizational Trust**- the broader institutional trust based on perceived ethical standards, privacy protections, and regulatory compliance communicated through the interface (Li et al., 2024; Gulati et al., 2025).

This framework aims to fill a gap in the literature by moving beyond narrow conceptions of trust that focus solely on either AI behavior (e.g., explainability, accuracy) or organizational ethics (e.g., privacy policies), and instead showing how these layers are **interconnected and interface-mediated**. For example, a highly intuitive healthcare AI interface may facilitate early trust, but inconsistent diagnostic performance may cause users to lose confidence. Conversely, a system with solid performance but a confusing or opaque interface may fail to gain user acceptance.

2. Conceptual Framework

2.1 Defining Trust in AI Interfaces

Trust is a foundational concept in human–AI interaction, often defined as a psychological state involving the willingness of a user to rely on a system under conditions of uncertainty or risk (Mayer et al., 1995; Glikson & Woolley, 2020). In the context of AI, trust is neither uniform nor static; it develops dynamically based on user experience, system performance, and contextual cues (Ueno et al., 2022; Bach et al., 2022). However, much of the existing literature on AI trust emphasizes algorithmic behavior (e.g., accuracy, explainability) or system-level ethics (e.g., fairness, transparency) without fully recognizing the role of the **user interface** as the initial and ongoing touchpoint through which these factors are interpreted and internalized by the user (Ribes Lemay et al., 2021; Böckle et al., 2021).

This paper defines **trust in AI interfaces** as a **multi-layered construct** that emerges progressively through three interrelated channels:

1) **Interface Trust** – user perceptions based on visual design, usability, and transparency;
2) **Behavior Trust** – trust derived from the AI's actions, performance, and decision consistency;
3) **Organizational Trust** – broader trust informed by perceived institutional ethics, compliance, and data practices.

This layered model reflects the interdependence of **design, performance, and credibility**, positioning the interface not merely as a functional surface, but as an **epistemic and affective medium** that mediates the user's trust journey from initial engagement to sustained interaction.

2.2 The Three Layers of Trust

2.2.1 Interface Trust

Interface Trust refers to the trust users form **at the point of interaction** with an AI system's visual, informational, and navigational elements. It is influenced by design factors such as layout clarity, information architecture, interactive feedback, and visual indicators of transparency (Zerilli et al., 2022; Papenmeier et al., 2019). For instance, interfaces that clearly disclose AI decision rationales (e.g., "why this diagnosis was suggested") and allow users to explore alternative outcomes promote a greater sense of control and understanding (Dhiman et al., 2023). Studies in explainable AI (XAI) show that **interface-based explanations**, when tailored to user expertise, can increase perceived trustworthiness and reduce mental workload (Sunny, 2025).

Interface Trust also includes emotional and aesthetic dimensions. Research shows that **visually coherent and intuitively designed interfaces** are more likely to be perceived as credible, even if the AI behind them is less accurate (Böckle et al., 2021). This introduces the risk of **over-trust**-where users may trust a system due to interface fluency rather than real performance. Therefore, **Interface Trust must be designed to calibrate expectations**, not simply to gain attention or acceptance.

2.2.2 Behavior Trust

Behavior Trust is grounded in the **observed performance and decision-making behavior of the AI system**, as interpreted through the interface. Users assess whether the AI is consistent, reliable, accurate, and intelligible over time (Glikson & Woolley, 2020). Behavior Trust depends not just on whether the AI is "correct," but whether users can understand and predict its actions. This is closely tied to **explainability, consistency, and feedback**.

Studies indicate that users often prioritize **performance over explanation** when deciding whether to trust an AI system (Papenmeier et al., 2019). However, explanation quality and relevance- especially when embedded in the interface- can modulate trust significantly (Ribes Lemay et al., 2021). For example, finance users tend to trust AI trading assistants that provide **traceable justifications** for predictions and highlight model confidence levels (McGrath et al., 2025).

Behavior Trust is dynamic and may increase or deteriorate with experience. Minor inconsistencies or decision reversals, if not explained properly, can lead to **behavioral distrust**, even if overall system accuracy remains high (Ueno et al., 2022).

2.2.3 Organizational Trust

Organizational Trust involves users' trust in the institution or entity behind the AI system, including its ethical standards, data governance policies, compliance with regulations, and commitment to responsible AI practices (Li et al., 2024; Henrique et al., 2024). Unlike Interface or Behavior Trust, this layer is often **indirectly communicated** through the interface- via privacy notices, ethical disclosures, regulatory compliance badges, or accessible user policies (Zerilli et al., 2022).

In domains like healthcare and finance, where decisions carry legal and ethical consequences, Organizational Trust becomes critical. Users may trust a system more if it is backed by a **reputable hospital network or financial institution** with transparent data usage policies (Tun et al., 2025; Sagona et al., 2025). The perception of organizational accountability can act as a **trust buffer-** protecting user confidence during temporary system failures or updates (Wong et al., 2025).

2.3 Interrelationships Between the Trust Layers

The three trust layers- Interface, Behavior, and Organizational- are **not independent**, but **mutually reinforcing**. Interface Trust often acts as the **entry point**, shaping initial perceptions and setting expectations. It directly influences how users interpret system behavior and perceive organizational credibility. Conversely, negative experiences with AI performance (Behavior Trust) or institutional misconduct (Organizational Trust) can retroactively reduce trust in the interface itself, regardless of design quality (Gulati et al., 2025; Montag, 2024).

These interrelations underscore the need to design AI interfaces **holistically**, acknowledging that trust cannot be engineered at a single layer. Rather, it emerges through **interactions across layers**, mediated by the interface as the communication bridge.

3. Trust in AI Interfaces Across Domains

Trust in AI systems is highly contextual, shaped by domain-specific expectations, perceived risks, and user familiarity. While trust formation follows a similar layered structure- starting with interface interaction and deepening through observed behavior and organizational signals-its manifestation varies across industries. This section examines the application of the three-layered trust model in two critical and contrasting domains: **finance** and **healthcare**.

3.1 Finance Domain

In finance, users interact with AI systems in the form of robo-advisors, algorithmic trading platforms, fraud detection tools, and personalized financial planning assistants. These systems often operate in environments characterized by high-frequency data, probabilistic forecasts, and complex risk modeling.

Interface Trust in Finance

Financial AI systems must present **dense, abstract, and often high-risk decisions** in a way that is clear, intuitive, and accessible. Trust in these systems often begins with visual design elements-such as clean dashboards, real-time updates, and interactive simulations-that enable users to feel in control and informed (Ribes Lemay et al., 2021). Effective interfaces in finance leverage **visual transparency** (e.g., trend graphs, confidence intervals) and **narrative explanations** (e.g., "Your portfolio was adjusted due to market volatility") to build trust (Böckle et al., 2021).

Behavior Trust in Finance

Behavior Trust is primarily influenced by the **AI's predictive accuracy, consistency, and rationality** in decision-making. Users form expectations about the system's reliability over time by observing how it reacts to market changes and whether its recommendations align with financial goals (Papenmeier et al., 2019). A system that provides **explanation-backed decisions** (e.g., "This stock was excluded due to ESG concerns") reinforces perceived behavioral transparency. However, performance failures-especially unaccompanied by justification-can rapidly erode trust.

Organizational Trust in Finance

Trust in financial AI is strongly shaped by the reputation of the organization behind the technology. Indicators such as **compliance with financial regulations (e.g., GDPR, MiFID II)**, **visible privacy settings**, and **third-party security certifications** (e.g., ISO 27001) serve as proxies for organizational trust (Henrique et al., 2024). Financial institutions often signal trustworthiness by integrating **disclaimers, risk warnings**, and access to human oversight directly within the interface. Institutional branding and customer support availability also reinforce trust, particularly when users face financial losses or volatility.

3.2 Healthcare Domain

In healthcare, AI systems support diagnostics, treatment recommendations, patient monitoring, and administrative triage. The stakes are especially high due to the direct impact on human health and well-being, demanding both technical precision and ethical rigor.

Interface Trust in Healthcare

Trust in healthcare AI begins with interfaces that are **clinically interpretable**, **accessible**, and **aligned with medical standards**. For both patients and professionals, transparency in how diagnoses or recommendations are generated is essential. Features such as **step-by-step rationale**, **visual annotation of scans**, and **confidence levels** in diagnoses foster trust (Dhiman et al., 2023). Interface elements must also accommodate diverse user types- including clinicians, patients, and caregivers-each with varying levels of expertise and cognitive load (Jermutus et al., 2022).

Behavior Trust in Healthcare

In healthcare, **behavioral trust is heavily influenced by diagnostic accuracy, error handling, and consistency over time**. AI systems that fail to deliver consistent outcomes or provide unreliable recommendations-even if well-designed-are quickly distrusted, especially by clinicians who must justify medical decisions (Tun et al., 2025). Furthermore, explainability becomes vital: clinicians often seek **clinically grounded justifications**, not generic model explanations (McGrath et al., 2025). In this context, **false positives and negatives** carry serious consequences, making behavioral transparency a critical component of sustained trust.

Organizational Trust in Healthcare

Organizational Trust in healthcare depends on perceived **ethical integrity, data stewardship, and regulatory compliance** (Sagona et al., 2025; Wong et al., 2025). Trust is strengthened when the AI system is developed or endorsed by **reputable medical institutions** and demonstrates compliance with health data regulations (e.g., HIPAA, GDPR). Interfaces that clearly disclose **privacy protections, auditability, and access control** help reinforce the institution's credibility. In times of system failure or ethical controversy, organizational transparency often determines whether trust is restored or permanently lost.

3.3 Comparative Analysis Across Domains

While both domains involve high-stakes decisions, the **dimensions of trust are weighted differently** depending on user priorities and contextual factors.

Dimension	Finance	Healthcare
Interface Trust	Emphasis on **clarity, interactivity**, and control	Focus on **clinical interpretability** and **accessibility**
Behavior Trust	Users value **accuracy + explainability** for risk decisions	Users prioritize **consistency and diagnostic safety**
Organizational Trust	Driven by **regulatory compliance** and **brand trust**	Grounded in **ethical standards** and **medical oversight**

In finance, users may tolerate occasional prediction failures if the **institution is perceived as secure** and **the interface provides sufficient transparency**. In contrast, healthcare users may value **behavioral reliability** and **organizational integrity** more highly, especially in life-critical scenarios.

For example, a user may trust a fintech AI that made one inaccurate stock forecast but still provides detailed reasoning and is operated by a known bank. However, a patient is unlikely to continue using a diagnostic AI that misidentifies symptoms-even with a clean, user-friendly interface-if the system lacks credible backing or medical accountability.

4. Developing the Layered Trust Model

The development of trust in AI interfaces is not a linear process; rather, it is dynamic, recursive, and shaped by repeated user interaction. This section expands upon the proposed three-layer model-**Interface Trust**, **Behavior Trust**, and **Organizational Trust**-by explaining how these layers interact, evolve, and influence each other throughout the user experience. Trust is understood here as both a **temporal phenomenon** and a **multi-dimensional construct**, subject to change based on cumulative

experiences, feedback loops, and contextual factors (Glikson & Woolley, 2020; McGrath et al., 2025).

4.1 User Perception and Initial Trust Formation

At the outset, users form impressions based on **interface cues**, including visual design, layout coherence, transparency mechanisms, and interaction flow. This process, often described as **surface-level cognitive trust formation**, is critical for establishing an initial willingness to engage with the system (Zerilli et al., 2022; Ribes Lemay et al., 2021). Elements such as **onboarding clarity**, **explanation availability**, and **predictive feedback** can build trust quickly- or provoke skepticism if they are confusing or absent.

For instance, in financial AI applications, an interface that explains how portfolio risk levels are calculated and allows the user to simulate different investment scenarios fosters **Interface Trust**. In healthcare, a diagnostic tool that uses visual annotations and natural language summaries of medical scans can establish early trust among clinicians or patients (Dhiman et al., 2023).

However, these initial perceptions are **fragile**. Trust formed at the interface level must be validated by the system's behavior to sustain long-term user confidence (Papenmeier et al., 2019).

4.2 Trust Evolution Over Time

Trust in AI interfaces evolves as users interact with the system and accumulate evidence of its **reliability, predictability, and ethical alignment**. This evolution often follows a **three-phase trajectory**:

1) **Initiation Phase**: Users form early impressions based on interface quality and initial interactions (Interface Trust).
2) **Validation Phase**: Users test system behavior through repeated use, evaluating consistency, error handling, and explainability (Behavior Trust).
3) **Consolidation Phase**: Users form holistic judgments based on long-term system performance and organizational assurances (Organizational Trust).

As trust matures, it can either **stabilize**, **deepen**, or **decline**. For example, a user might initially trust a healthcare AI system due to its transparent interface but lose trust if the system misdiagnoses a common illness. Alternatively, a user might be skeptical at first but build trust as the system consistently performs and the organization provides ethical guarantees, such as transparent audit trails or data ownership disclosures (Henrique et al., 2024; Li et al., 2024).

4.3 Feedback Loops in Trust Formation

Trust development is influenced by **feedback loops**-reciprocal mechanisms by which system responses shape user expectations, and user behavior in turn shapes system adaptation or perception. These loops can be **positive** (trust reinforcement) or **negative** (trust erosion), and they often cut across the three layers:

- A **positive loop** may occur when a transparent interface leads to correct understanding of AI behavior, which then matches user expectations and is supported by the organization's reputation. This increases trust across all layers.
- A **negative loop** may begin with a confusing interface that leads to misinterpretation of an AI decision. Even if the decision is technically correct, the user's misunderstanding can result in dissatisfaction and doubts about the system's reliability or the institution's credibility.

Research shows that trust violations are particularly **difficult to repair** unless explicit trust-rebuilding mechanisms (e.g., apology, explanation, interface revision) are implemented (Zerilli et al., 2022; Montag, 2024). Hence, the design of **adaptive interfaces-** ones that respond to trust breaches by offering contextual explanations or escalation pathways-can serve as critical feedback regulators.

4.4 Cross- Layer Impacts and Fragility

The layered nature of trust also introduces **dependency risks**-where failure at one layer can cascade into others. For instance:

- A poorly designed interface may cause users to misunderstand AI outputs, thereby reducing Behavior Trust.
- A system may perform well technically but fail to disclose its data-sharing practices, damaging Organizational Trust and, retroactively, user confidence in the interface.
- Even with high organizational transparency, persistent prediction errors may cause users to question both the behavior and legitimacy of the system.

This **fragility underscores the interdependence** of the three layers. Trust cannot be "fixed" at one layer while neglecting the others. High-functioning AI systems require **synchronization of design, behavior, and institutional practices**, all made visible and actionable through the user interface (De Silva et al., 2025; Gulati et al., 2025).

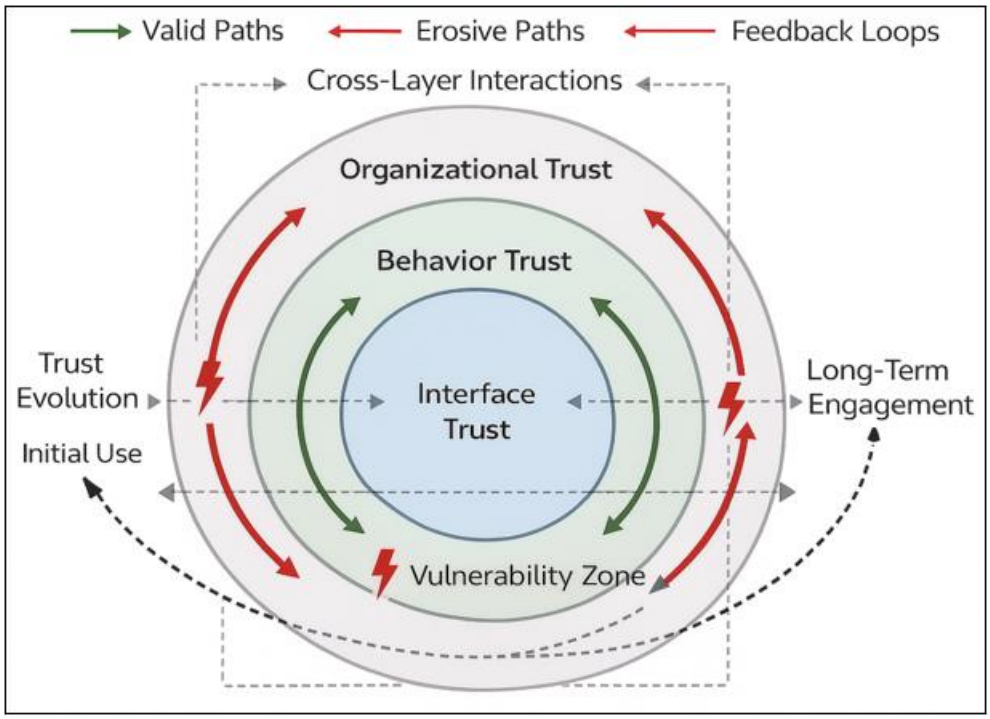

Figure 1: Multi Layered Trust Model in AI Interfaces

Figure 1 illustrates a multi-layered trust model in AI interfaces, showing how Interface, Behavior, and Organizational Trust interact and evolve through user engagement, feedback loops, and vulnerability points.

5. Case Studies and Scenarios

This section illustrates the application of the three-layered trust model in real-world AI deployments. By analyzing empirical cases from **healthcare** and **finance**, and exploring trust breakdown and recovery through constructed scenarios, we demonstrate how trust evolves across Interface, Behavior, and Organizational layers. These cases support the model's validity and its relevance to both domain-specific and generalizable trust challenges in human–AI interaction.

5.1 Case Study 1: IBM Watson Health in Clinical Decision Support

Interface Trust
IBM Watson Health, now operating under the brand Merative, was initially introduced as a clinical AI tool to assist in cancer diagnosis and treatment recommendations. The system featured a **visual dashboard** for oncologists to interact with, including patient summaries, evidence-backed treatment options, and comparative visuals (Henrico Dolfing, 2022). These interface features fostered early Interface Trust, particularly in U.S. hospitals, where Watson's brand and design lent credibility and usability.

Behavior Trust
Despite its polished interface, the system encountered major challenges in **Behavior Trust**. Watson for Oncology was criticized for recommending inappropriate treatments in certain international contexts, partly due to its **training data being heavily U.S.-biased** (ForeSee Medical, 2023). In some cases, recommendations conflicted with local clinical practices, undermining physicians' trust in the system's competence. These incidents revealed a misalignment between AI behavior and domain-specific expectations.

Organizational Trust
Initially, Watson's backing by IBM conferred high Organizational Trust. However, as clinical concerns grew and results failed to meet expectations, hospitals and practitioners began questioning IBM's **development process, transparency, and governance model**. Ultimately, Watson Health was sold and restructured, illustrating how Organizational Trust, once compromised, can deteriorate rapidly-even if Interface and Behavior Trust were partially intact at launch.

5.2 Case Study 2: Robo-Advisors in Financial Services

Interface Trust
Robo-advisors like **Wealthfront** and **Betterment** are AI-powered investment platforms offering users algorithm-driven portfolio management. Their user interfaces are designed to be clean, interactive, and informative-displaying real-time portfolio breakdowns, market performance visualizations, and risk-adjusted investment simulations (Wealthfront, 2024). Research shows that these interface design choices significantly influence **user confidence and initial adoption**, particularly among digital-native investors (Wang et al., 2023).

Behavior Trust
Long-term Behavior Trust in robo-advisors is based on **portfolio performance, explanation of rebalancing decisions**, and AI reaction to market changes. Users tend to trust the system when changes are **accompanied by timely explanations**, such as "Increased bond allocation due to recent market volatility" (Alshurideh et al., 2023). Conversely, when recommendations lack context or appear inconsistent, users express behavioral skepticism-even if performance is objectively stable.

Organizational Trust
The trustworthiness of the institutions behind robo-advisors is a **critical adoption factor**. Empirical research confirms that users are more likely to accept AI financial recommendations when the platform is **regulated**, displays **certifications**, and offers **transparent privacy and risk disclosures** (Gupta & Singh, 2024). This reinforces the need for Organizational Trust, especially when large financial decisions are delegated to AI systems.

5.3 Scenario-Based Analysis: Trust Breakdown and Recovery

Scenario A: Interface Trust Breakdown in Clinical AI
A diagnostic imaging tool crashes without explanation during a consultation. The interface provides only a generic error message with no further guidance. Despite the AI's previous accuracy, the physician is now uncertain whether the system failed internally or the data input was flawed. This single incident creates **cross-layer trust damage**, reducing both Behavior and Organizational Trust due to a lack of transparency.
Recovery Strategy: Introduce robust error messaging, contextual alerts, and direct access to support or audit logs. Provide users with recovery pathways and logs to rebuild Interface Trust and demonstrate organizational accountability.

Scenario B: Behavior Trust Erosion in Finance
A user notices frequent, unexplained changes in their investment portfolio allocation by a robo-advisor. While the interface remains sleek, no decision rationale is provided. As performance fluctuates, the user begins doubting whether the system understands market dynamics or their personal goals.
Recovery Strategy: Provide retrospective reports on decision logic, personalized investment rationale, and AI model updates to rebuild Behavior Trust. Integrate explainability modules to support transparency without overwhelming novice users.

Scenario C: Organizational Trust Collapse
A privacy leak exposes user data from a healthcare AI vendor. Although the system's interface and diagnostic logic remain unchanged, users lose confidence in the organization's ethical handling of sensitive information. The breach results in mass uninstalls and termination of hospital contracts.

Recovery Strategy: Transparent crisis communication, third-party audits, revised governance policies, and clear display of remediation measures (e.g., encryption updates, access controls) are essential to recover Organizational Trust.

6. Implications for Design and Development of AI Systems

The three-layered model of trust- comprising Interface Trust, Behavior Trust, and Organizational Trust- offers actionable insights for the design and deployment of AI systems. Building trust is not solely a technical challenge; it is a **design, behavioral, and institutional endeavor** requiring coordination across user experience, system performance, and organizational accountability (Gulati et al., 2025; De Silva et al., 2025). This section presents design guidelines for fostering trust at each layer, with an emphasis on interface-mediated strategies.

6.1 Best Practices for Interface Design (Interface Trust)

The user interface is the primary surface through which AI systems communicate intent, logic, and limitations. Trust can be strengthened by embedding transparency and control into the design.

- **Ensure Transparency by Design**: Use layered explanations and confidence indicators to convey how AI decisions are made (Ribeiro et al., 2016; Ribes Lemay et al., 2021). Visual disclosure of system logic helps users calibrate their trust appropriately.
- **Prioritize Usability and Accessibility**: Clear visual hierarchy, intuitive interaction flows, and accommodation for diverse cognitive and cultural backgrounds are essential for fostering initial trust (Bach et al., 2023).
- **Enable User Control and Customization**: Allow users to select their level of detail for explanations, modify decision parameters, and provide feedback on system performance (Shneiderman, 2020). Such control mechanisms improve perceived autonomy and trust.
- **Provide Guided Onboarding**: First-time interaction tutorials or contextual help systems support trust calibration by aligning user expectations with system capabilities (Zerilli et al., 2022).
- **Visualize Uncertainty and Errors**: Communicating limitations-such as prediction uncertainty or system downtime- builds credibility by reducing over-trust and clarifying boundaries of automation (Papenmeier et al., 2019).

6.2 Enhancing Trust Through AI Behavior (Behavior Trust)

Behavior Trust is influenced by perceived system intelligence, consistency, and logic over time.

- **Adopt Explainable AI Methods**: Use inherently interpretable models or supplement black-box models with post-hoc explainability tools (Guidotti et al., 2019). Explanations should be tailored to user expertise.
- **Support Traceability and Auditing**: Decisions should be linked to the data inputs and model logic that generated them. Traceability fosters trust, especially in regulated industries like healthcare and finance (Sagona et al., 2025).
- **Implement Feedback Loops**: Systems should allow users to flag errors, refine outputs, or contribute to ongoing learning. This builds interactive trust and enables behavioral tuning over time (McGrath et al., 2025).
- **Monitor Behavioral Consistency**: Users expect consistency unless otherwise indicated. Notify users of significant AI updates or retraining events that could affect outputs (Henrique et al., 2024).

6.3 Communicating Organizational Integrity Through Interfaces (Organizational Trust)

Trust in the organization behind the AI system can be either reinforced or undermined by how the interface communicates institutional commitments.

- **Surface Ethical and Legal Guarantees**: Prominently display compliance credentials (e.g., HIPAA, GDPR), third-party audits, and AI ethics policies through interface components (Li et al., 2024).
- **Disclose Data Practices**: Use interactive data dashboards or summaries to communicate what data is collected, how it is used, and by whom. Transparency enhances perceived accountability (Glikson & Woolley, 2020).
- **Offer Access to Human Oversight**: Embed human support options, appeal mechanisms, or explainability dashboards to ensure users feel protected from total automation (De Silva et al., 2025).
- **Clarify System Boundaries**: Indicate what the AI is and is not authorized or capable of doing. Misaligned expectations often originate from ambiguous system roles (Zerilli et al., 2022).

6.4 Toward Trust by Design

The convergence of these guidelines reflects the emerging paradigm of **"Trust by Design"**- an interdisciplinary approach that integrates **human-centered design, technical robustness, and institutional governance** (Shneiderman, 2020; Li et al., 2024). This paradigm calls on designers to **embed trust into the architecture** of AI systems from the outset, rather than retrofitting trust mechanisms after deployment.

By aligning Interface, Behavior, and Organizational Trust strategies, developers can create AI systems that are not only functional but **socially sustainable**, **ethically defensible**, and **personally meaningful** to users. This layered approach supports the development of trust as a **distributed and evolving property**, rather than a one-time assurance.

7. Challenges and Limitations

While the proposed three-layered model of trust in AI interfaces offers a structured and integrative framework for understanding how users engage with AI systems, it is not without limitations. These challenges reflect both **conceptual boundaries** and **practical constraints** in applying the model across diverse domains, user groups, and system types.

7.1 Challenges in Measuring Trust Across Layers

Trust remains an inherently **subjective, context-dependent, and temporally dynamic** construct, making it difficult to measure consistently across individuals and applications. Although the model distinguishes among Interface, Behavior, and Organizational Trust, operationalizing these dimensions in empirical studies is complex. Existing measurement tools often conflate different aspects of trust, or focus on a single layer-such as perceived accuracy or usability-without capturing the interaction between layers.

Furthermore, the **longitudinal nature of trust** presents additional complexity: trust may grow, decline, or recover based on cumulative interactions and contextual shifts. Capturing these dynamics requires **long-term user studies**, which are resource-intensive and difficult to generalize across domains.

7.2 Subjectivity and Cultural Variability

Users' trust perceptions are influenced by **personal, cultural, and situational factors**, including prior experiences with technology, domain knowledge, risk tolerance, and cultural attitudes toward automation. For example, a feature that promotes transparency in one cultural context (e.g., detailed audit logs) may be perceived as overwhelming or unnecessary in another.
This variability limits the **universal applicability** of specific design strategies and suggests that **localized adaptations** of the trust model may be necessary. Without accounting for socio-cultural diversity, AI systems risk alienating users or miscalibrating trust expectations.

7.3 Ambiguities in Layer Distinction

Although the model conceptually distinguishes among the three trust layers, in practice these boundaries are often **blurry and overlapping**. For instance, an interface that displays ethical commitments (e.g., data privacy badges) could be seen as both an **Interface Trust** and **Organizational Trust** cue. Similarly, behavior-based trust may rely on how well explanations are presented through the interface, merging elements of all three layers.

These overlaps raise challenges in **design attribution** and **intervention analysis**. When trust fails or succeeds, it may not be immediately clear **which layer is responsible**, complicating evaluation and refinement efforts.

7.4 Limitation of Scope to Interface-Mediated Systems

The model assumes that trust is formed primarily through **interface-mediated interactions**. While this is true for most AI systems, it may not fully apply to:

- **Embedded or ambient AI** (e.g., smart homes or autonomous vehicles),
- **Multi-agent systems** where decision-making is distributed, or
- **Backend AI systems** with limited or no direct user interface.

In such cases, trust may depend more heavily on **indirect signals** (e.g., system reputation, peer endorsement) or organizational guarantees that are not directly interface-visible. This calls for extension or modification of the model for non-interface-dominant AI contexts.

7.5 Potential for Overemphasis on Interface Design

While the model emphasizes interface as the first point of trust formation, there is a risk of **overemphasizing superficial interface features** at the expense of deeper issues, such as model fairness, robustness, and institutional accountability. A well-designed interface can **mask**

underlying algorithmic or ethical flaws, leading to misplaced or "false" trust- a phenomenon known as **overtrust** or **trust miscalibration**.

This risk underscores the need for **substantive alignment** between what is shown at the interface and what is true in system behavior and organizational practice. Otherwise, the interface becomes a veneer rather than a trust-enabling channel.

7.6 Limited Empirical Validation of Layered Dynamics

Although grounded in literature and case examples, the model has yet to undergo **comprehensive empirical validation**. Future research should assess:

- How trust transitions across layers over time,
- Which interventions most effectively support trust repair,
- And whether the model holds across novel domains (e.g., education, law, social robotics).

Such empirical work is essential to strengthen the model's predictive power and guide practical design implementations.

8. Future Research Directions

The proposed three-layered trust model-Interface Trust, Behavior Trust, and Organizational Trust-offers a conceptual framework for understanding how users build, maintain, or lose trust in AI systems. However, realizing its full potential requires ongoing refinement through empirical testing, cross-domain validation, and methodological innovation. This section outlines key areas where future research can deepen understanding and enhance the model's applicability across AI ecosystems.

8.1 Trust Across Diverse AI Technologies and Modalities

While this paper has focused on AI systems with decision-support interfaces in domains such as healthcare and finance, AI technologies are increasingly found in **non-traditional or low-interface modalities**, including:

- **Conversational agents** (e.g., ChatGPT, voice assistants),
- **Autonomous systems** (e.g., drones, self-driving cars),
- **Social robotics** and **ambient AI**.

Future research should investigate whether the layered trust model holds in these contexts or requires adaptation. For instance, in embodied AI (e.g., caregiving robots), Interface Trust may blend with physical cues and behavior-based signals, altering how users perceive and calibrate trust.

8.2 Longitudinal and Dynamic Trust Studies

Trust in AI is not static; it evolves with continued interaction, feedback, and contextual change. Yet, many existing studies rely on **cross-sectional or short-term experimental data**, which limits understanding of **trust trajectories** over time.
Future research should prioritize:

- **Longitudinal user studies** that track trust development and decay across use phases,
- **Trust calibration mechanisms**, including recovery from errors or failures,
- **Modeling trust feedback loops**, where trust in one layer influences or compensates for deficits in another.

Such research would strengthen the model's temporal dimension and guide developers in designing for trust **adaptation and resilience**.

8.3 Cultural, Demographic, and Contextual Sensitivity

Trust is influenced by cultural norms, user experience levels, and societal values. For instance, attitudes toward automation, institutional credibility, or privacy expectations vary across geographies and user demographics. Current trust models often adopt a **universalist approach**, which risks oversimplifying user diversity.

Research should explore:

- **Cross-cultural studies** comparing how trust forms across different populations,
- **User segmentation models** based on age, profession, or prior AI exposure,
- **Localized interface adaptations** and ethical framing strategies that reflect regional expectations.

These efforts can help refine the model to accommodate **trust pluralism** and reduce design bias.

8.4 Operationalizing the Trust Layers in Evaluation

Although the model distinguishes among three trust layers, empirical evaluation tools often conflate trust into a single construct or lack clear instrumentation for each dimension. Future work should focus on:

- Developing **validated scales** for Interface, Behavior, and Organizational Trust,
- Designing **diagnostic tools** that help researchers and practitioners identify which layer is most responsible for trust successes or failures,
- Integrating trust measurement into **UX and system performance metrics** for iterative design evaluation.

This would enable more **granular trust diagnostics** and inform targeted interventions.

8.5 AI Transparency, Ethics, and Regulatory Implications

As regulatory frameworks for AI governance evolve (e.g., EU AI Act, U.S. AI Bill of Rights), Organizational Trust is becoming not only a design concern but a **compliance mandate**. Future research should explore how the layered trust model aligns with:

- **Emerging AI regulations and policy frameworks**,
- **Ethical AI development standards** (e.g., IEEE, ISO, UNESCO),
- **Organizational transparency practices**, such as impact assessments and algorithmic audits.

Scholars and practitioners must co-develop tools that translate **ethical obligations into user-facing interface features** that can be understood and trusted.

8.6 Generalization Across Domains

While this study focused on healthcare and finance- both high-stakes, regulated domains- future work should examine:

- Application of the model in **low-stakes domains** (e.g., entertainment, education),
- Adaptability to **AI-mediated social systems**, such as hiring platforms or predictive policing,
- Use in **hybrid systems**, where AI collaborates with humans in real-time teams.

Such expansions can test the **generality and boundaries** of the model and help refine its theoretical underpinnings.

9. Conclusion

As artificial intelligence systems become increasingly integrated into domains that shape health, financial security, and daily life, the challenge of building and maintaining user trust is more urgent than ever. This paper has introduced a **three-layered model of trust in AI interfaces**-comprising **Interface Trust**, **Behavior Trust**, and **Organizational Trust**-to offer a nuanced and structured understanding of how trust is initiated, reinforced, or compromised throughout the user's journey with AI.

Through this model, we emphasize that trust in AI is not a monolithic construct, nor can it be reduced to technical performance alone. Instead, trust emerges from the interplay of design transparency, behavioral consistency, and institutional credibility-all of which must be communicated effectively through the AI interface. Drawing from empirical case studies in healthcare and finance, we illustrated how trust is shaped at each layer, how breakdowns propagate across layers, and how carefully designed interfaces can mediate trust repair and recovery.

This framework contributes to academic discourse by bridging gaps between human–computer interaction, AI ethics, and organizational accountability. It also provides practical guidance for developers, system architects, and policymakers, underscoring the importance of **"trust by design"** as a foundational principle for AI deployment. The model serves not only as an analytical lens but also as a design blueprint for creating AI systems that are ethically aligned, contextually sensitive, and socially robust.

Moving forward, interdisciplinary collaboration will be essential to validate, extend, and operationalize the layered trust model. By anchoring trust in the interface while accounting for deeper behavioral and institutional dynamics, we can begin to design AI systems that do more than function- they can be trusted, understood, and embraced in ways that support long-term human–AI collaboration.

References

[1] Bach, P., Diehl, D., & Wibowo, S. (2023). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction, 39*(5), 421–440. https://doi.org/10.1080/10447318.2022.2138826

[2] De Silva, D., Halloluwa, T., & Vyas, D. (2025). A multi-layered research framework for human-centered AI. *arXiv preprint arXiv:2504.13926*. https://arxiv.org/abs/2504.13926

[3] Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

[4] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black-box models. *ACM Computing Surveys, 51*(5), Article 93. https://doi.org/10.1145/3359786

[5] Henrique, J., Lima, G., & Costa, L. (2024). Trust in AI: Literature review and challenges. *Information Systems Frontiers, 26*(1), 45–62. https://doi.org/10.1007/s10796-022-10365-3

[6] Li, Y., Hao, Y., & Yang, M. (2024). Developing trustworthy AI: Insights from social cognition theory. *Frontiers in Psychology, 15*, 1382693. https://doi.org/10.3389/fpsyg.2024.1382693

[7] Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.2307/258792

[8] McGrath, A., Wang, L., & Kohn, N. (2025). Collaborative human–AI trust process framework. *Information Processing & Management, 62*(1), 102878. https://doi.org/10.1016/j.ipm.2024.102878

[9] Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence trust. *arXiv preprint arXiv:1907.12652*. https://arxiv.org/abs/1907.12652

[10] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

[11] Ribes Lemay, C., Jatowt, A., & Tanaka, K. (2021). Trust indicators and explainable AI: A study on user perceptions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3411764.3445679

[12] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1738096

[13] Ueno, M., Maldonado, S., & Murray, J. (2022). Trust in human–AI interaction: Models, measures, and methods. *arXiv preprint arXiv:2205.00189*. https://arxiv.org/abs/2205.00189

[14] Yang, Y., & Wibowo, S. (2022). User trust in artificial intelligence: A comprehensive conceptual framework. *Journal of Information Technology Management, 33*(3), 1–15.

[15] Zerilli, J., Knott, A., & Gavaghan, C. (2022). How transparency modulates trust in artificial intelligence. *Philosophy & Technology, 35*(2), 31–57. https://doi.org/10.1007/s13347-022-00502-5