Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

From Competence to Calibration: Modeling Cognitive Trust in Human-AI Collaborative Systems

Fahd Malik[1]✉, Muhammad Raza ul Haq[2], Zubair Shahid[3]
[1]*Department of Digital Business, Transformation & Innovation IE Business School, Madrid, Spain*
[2]*Department of Information Technology, Zain, Riyadh, Saudi Arabia*
[3]*Mavric Systems, Lahore, Pakistan*
Corresponding Author: Fahd Malik, **E-mail**: fahad.agmalik@gmail.com

| ABSTRACT
Human-AI collaboration is rapidly becoming embedded across domains: decision support, operations, creative work, quality assurance, and more. Yet what often limits effective collaboration is the human's cognitive trust in the AI system, the belief that the system is capable, reliable, understandable, and aligned with the user's goals. What this paper does is provide a conceptual model of how cognitive trust forms, evolves and influences behavior in human–AI collaboration. We synthesize trust antecedents (such as perceived competence, integrity, transparency, reliability, user disposition and contextual risk) with dual cognitive processing mechanisms (heuristic and systematic evaluation) to explain how users appraise an AI partner, adjust their trust level, and then behave (compliance, delegation, reliance). We also integrate a feedback loop by which outcomes of collaboration reshape cognitive trust and future appraisal. The contribution is two-fold: academically, we build a theory-driven framework linking psychological trust theory to AI system design; practically, we map design implications for AI systems that need calibrated trust rather than un-thinking over-trust or skeptical under-trust. We conclude by offering a set of testable propositions for subsequent empirical validation and highlight the implications for system designers, quality assurance professionals, and organizations adopting AI collaborations. The model opens the door for future studies on how to measure cognitive trust in AI settings, how to enable trust calibration, and how to build AI systems that align with human cognitive patterns of trust formation.

| KEYWORDS
Cognitive Trust, Human–AI Collaboration, Trust Formation, Cognitive Processing, Antecedents of Trust, Trust Calibration

| ARTICLE INFORMATION
ACCEPTED: 01 November 2025 **PUBLISHED:** 30 November 2025 **DOI:** 10.32996/jcsts.2025.7.12.23

1. Introduction

As artificial intelligence (AI) systems become integral in decision-making and collaborative workflows, understanding how humans establish trust in their AI partners gains critical importance. When end-users engage with intelligent systems, they not only assess whether the system works; they ask if they can rely on it, if it behaves consistently, and if it aligns with their goals in uncertain conditions. Trust that is miss-calibrated, either too high or too low, undermines effective human-AI collaboration. Over-trust may lead users to accept recommendations uncritically and thus expose themselves to risk. Under-trust may cause users to ignore or under-utilized valuable system support.

Trust in human-AI collaboration should be treated as a dynamic cognitive process rather than a static attribute. In this view, individuals arrive with initial beliefs shaped by perceptions of the system's attributes and their own predispositions. They then engage in ongoing evaluation of interaction outcomes, adjust their trust beliefs, and act accordingly, by delegating, monitoring,

or rejecting AI outputs. Prior empirical reviews show that users' perceptions of AI competence, transparency, reliability, and immediacy are central to developing cognitive trust [1]. Similarly, in human–computer interaction (HCI) research, trust models reflect multi-dimensional factors and emphasize the need to capture trust as evolving over time [2].

This dynamic process can be conceptualized via three linked domains: antecedents of trust, cognitive processing of trust signals, and behavioral outcomes (with feedback loops). First, antecedents set the stage before deep interaction unfolds. Users bring individual traits (for example a general propensity to trust technology), task and domain characteristics (for example perceived risk or stakes), and early perceptions of system qualities: perceived competence (can it perform the task well?), perceived integrity (does the system align with ethical or consistent behavior?), perceived transparency (can I understand how it works?), and perceived reliability (does it perform consistently?). Classic trust theory, such as [3] integrative model, indeed categorizes trust antecedents in terms of ability, integrity, and benevolence, setting a useful foundation for analyzing human-AI trust.

Second, as human-AI interactions occur, the human user engages in cognitive processing of system cues and outcomes. The user may rely on heuristic (fast, intuitive) judgments—"this system responded quickly so it must be good", or shift to systematic (deliberate, analytical) reasoning, "does the system's explanation make sense and align with my mental model?" Research on trust in automation highlights that trust formation involves analytic, analogical and affective processes under conditions of uncertainty and vulnerability [4]. The mode of processing depends on factors such as cognitive load, time pressure, domain expertise, and system complexity. The interplay of heuristic and systematic modes shapes how trust evolves over time.

Third, behaviors emerge from the trust beliefs: decisions about relying on the AI, delegating tasks, monitoring its outputs, or overriding them entirely. Critically, these behaviors generate outcomes, positive or negative, that feed back into the trust formation process. If the AI partner produces reliable, understandable, and useful results, trust may deepen; if it fails or behaves unpredictably, trust may erode or require recalibration. In essence, trust is not a one-off construct but a trajectory shaped by interaction, performance feedback, and evolving user perceptions.

Why this matters: From a design and deployment perspective, AI systems that neglect these cognitive dimensions risk failure. A technically capable AI that lacks transparency or produces unpredictable behavior may engender distrust. Conversely, an apparently transparent system with erratic performance may foster misplaced over-trust. From a theoretical standpoint, despite growing attention on trust in AI, there remains a gap in models that explicitly integrate human cognitive trust formation mechanisms within human-AI collaboration contexts. For example, [1] observe that much empirical research remains short-term, experimental, and limited in addressing long-term trust evolution. Additionally, trust measurement in HCI continues to struggle with conceptual and methodological heterogeneity [5].

In response, this paper proposes a unified conceptual model of cognitive trust in human-AI collaboration. It positions trust formation as a three-layered process of antecedents, cognitive processing, and behavioral outcomes, including a feedback loop that captures trust evolution over time. The contribution is two-fold: academically, by integrating psychological trust theory [3] and automation trust research [4] into the AI collaboration context; practically, by offering a blueprint for designers and human-AI teams to support calibrated trust rather than binary high/low trust states. Although this paper does not undertake empirical testing, it lays the groundwork for operationalization, measurement, and future validation of trust in human-AI systems.

In the remainder of the paper, we review key theoretical foundations of trust and cognitive processing, present the proposed conceptual model, discuss implications for design, human-AI team management, governance and future research, and conclude with a research agenda emphasizing measurement strategies, feedback mechanisms, and longitudinal analysis of trust dynamics.

2. Literature Review

Trust in automation research began by framing reliance problems in terms of use, misuse, disuse, and abuse, arguing that automation can fail not only through technical faults but through human miss-calibration of reliance. This perspective established a baseline for considering trust as a determinant of appropriate reliance and as a construct shaped by both system behavior and human cognition [6]. That early framing remains influential because it links trust to practical failure modes such as overreliance and disuse, which still surface in contemporary human–AI teamwork.

Subsequent empirical work quantified the reliance consequences of trust by showing that people's decisions to accept or reject automated advice track their trust beliefs, but often in biased ways. In controlled studies, participants initially displayed a positivity bias toward automated aids, then adjusted reliance downward following visible errors, revealing the asymmetry of trust

updates and the fragility of early trust gains [7]. This line of work demonstrated that trust is not simply a background attitude. It is a proximal driver of reliance choices that can either amplify or mitigate automation benefits.

Complementing these micro level findings, a large scale meta analysis in human robot interaction quantified how human factors, robot attributes, and environmental conditions contribute to perceived trust. The analysis showed that elements such as performance, reliability, and transparency consistently matter, while contextual moderators shift effect sizes across settings [8]. This quantitative evidence base helped move the field beyond narrative claims by providing pooled estimates and highlighting where heterogeneity persists across studies and domains.

Integrative reviews then organized the proliferating empirical results into conceptual scaffolds. Authors in [9] synthesized trust antecedents, dynamics, and outcomes into a three layer model and argued for designable levers that can adjust user trust over time. Their account emphasized variability in trust as a function of system cues and user traits, and it called for measures that separate trust from trustworthiness to avoid designing for blind acceptance.

As human–AI systems grew more opaque, scholarship on explanation and interpretability reframed part of the trust problem as one of sense making. Authors in [10] argued that explainable AI should draw from decades of social science on how humans generate and evaluate explanations, warning against assuming that any technical disclosure will increase user understanding or trust. The key insight is that explanations are social and pragmatic. People want contrastive, selective, and causal accounts that speak to their goals, not data dumps of model internals. This insight matters because superficial transparency can distort trust by creating either false reassurance or undue skepticism.

HCI work operationalized these concerns into actionable research agendas and patterns. In [11], authors mapped 289 papers to identify gaps at the intersection of explainability, accountability, and intelligibility, and set an agenda that concretizes how human factors methods can evaluate explanations in context. Their agenda made clear that explanation quality must be tested with people performing tasks and not inferred from algorithmic disclosure alone.

Experimental behavioral work complicated the intuition that more transparency always yields better trust. In [12], authors showed that the amount and framing of information about an algorithm shape trust in non monotonic ways. Moderate transparency improved trust, while too little or too much could backfire depending on expectation alignment. The implication is that designers must calibrate informational load to cognitive goals and task stakes, rather than chasing maximal disclosure.

Parallel streams in judgment and decision making clarified that trust is intertwined with preferences for human or algorithmic advice. In [13], authors documented algorithm aversion, where users abandon algorithms after observing even a single error, despite superior aggregate performance. Counterbalancing this, authors in [14] found algorithm appreciation in settings where algorithmic advice was framed as competent and objective, leading users to weigh it more heavily than human advice. Taken together, these results explain why human–AI collaboration often oscillates between under-trust and over-trust based on framing, first impressions, and salient mistakes.

Within AI assisted decision making, the quality of confidence cues and explanation design has direct effects on accuracy and trust calibration. In [15], it was demonstrated that showing model confidence and explanations can improve or degrade calibration depending on how these cues interact with user expertise and error distributions. The key contribution is to treat trust calibration as an outcome to be measured alongside accuracy, rather than assuming that more confidence information will automatically help users rely appropriately on models.

Measurement remains a perennial pain point. The field has relied heavily on scales that conflate trust, perceived trustworthiness, and related constructs. In [16], authors developed one of the earliest scales tailored to automated systems. Their instrument seeded a large measurement literature but also invited critiques about factor structure and construct purity in later work. This measurement history underscores why papers should report both subjective trust and behavioral reliance, and wherever possible tie them to ground truth system performance.

Recent surveys now focus explicitly on trust calibration as a design and evaluation goal. In [17], authors reviewed a substantial corpus and distinguished warranted trust from unwarranted trust, as well as static from adaptive calibration strategies. The survey argues for adaptive, context sensitive mechanisms that adjust cues and explanations based on signs of under or overreliance in the interaction, rather than one size fits all disclosures at the start or end of a task. This direction aligns with viewing trust as a dynamic cognitive process subject to feedback.

Across these streams, a few integrative themes emerge. First, trust is neither a simple trait nor a static property of the system. It is a trajectory that results from interactions among initial beliefs, observed performance, explanations, and contextual risk. Second, explanation quality is necessary but not sufficient. Explanations must be designed to answer the user's practical why and why not questions while avoiding cognitive overload. Third, calibration rather than maximization should be the design objective. The aim is to align user trust with actual system competence through confidence cues, targeted explanations, and interaction patterns that surface limitations at the right time. Fourth, measurement must mature. Instruments need to distinguish trust from trustworthiness and connect subjective trust to behavioral reliance and error detection. Finally, boundary conditions matter. Algorithm aversion and appreciation show that the same user population can move in opposite directions based on error salience, framing, and early experiences, which underscores the importance of longitudinal evaluation and adaptive trust interventions that respond to moment by moment signals.

What this literature collectively indicates for a cognitive model of trust formation in human–AI collaboration is straightforward. Antecedents such as perceived competence, integrity, transparency, and reliability set priors that are then filtered through dual processing routes. Heuristic cues like interface fluency and anthropomorphic framing shape fast judgments, while systematic processing engages when the task is high stakes, when explanations are actionable, or when confidence cues prompt deeper scrutiny. Outcomes such as compliance, delegation, and monitoring behavior both express trust and update it, especially when users encounter errors or mismatches between expectations and observed behavior. Designing for calibrated trust therefore requires a portfolio of mechanisms that manage first impressions, provide the right information at the right time, and make limitations legible without overwhelming the user. The reviewed evidence base offers clear leverage points for those mechanisms and a caution that more information alone is not a panacea.

3. A Cognitive Trust Formation Model for Human–AI Interaction

Trust in human–AI collaboration is not a single cognitive judgment but a dynamic process involving multiple theoretical dimensions. The framework proposed here integrates established trust and cognition theories into a single conceptual model that explains how humans form, evaluate, and adapt trust toward AI collaborators. This integration rests on four primary theoretical pillars: the Integrative Model of Organizational Trust [3], the Trust in Automation framework [4], the Heuristic–Systematic Model of information processing [18], and the Social Cognitive Theory [19].Each contributes a distinct explanatory lens for the antecedents, mechanisms, and outcomes of cognitive trust formation in human–AI interaction.

At its core, the model assumes that trust arises from an interplay between user cognition, perception of system attributes, and contextual influences. The process begins with antecedents that shape initial beliefs about the AI's competence and integrity. It proceeds through a cognitive processing stage, where heuristic and systematic reasoning determine how trust information is interpreted. The outcomes, trust calibration, compliance, reliance, and delegation represent behavioral manifestations of trust, which in turn feed back into the antecedent layer through ongoing experience. Table 1 summarizes the four theoretical foundations that inform the Cognitive Trust Formation Model.

Theoretical Lens	**Core Constructs / Principles**	**Relevance to Human–AI Trust Formation**	**Mapped Model Component**
Organizational Trust Theory (Mayer, Davis, & Schoorman, 1995)	Ability, integrity, benevolence as the basis for perceived trustworthiness	Defines *antecedent-level perceptions* (perceived competence, reliability, integrity) that shape initial cognitive trust toward AI systems	Antecedents
Heuristic–Systematic Model (Chaiken, 1980)	Dual processing of information: heuristic (quick, intuitive) vs. systematic (deliberate, analytical)	Explains how users process AI cues through fast intuitive or slow analytical reasoning, shaping ongoing trust calibration	Cognitive Processing
Social Cognitive Theory (Bandura, 1986)	Learning through observation, experience, and feedback loops (self-efficacy, adaptation)	Describes how users refine trust through feedback, experience, and adaptation over repeated interactions	Feedback Loop / Learning Mechanism
Trust in Automation Framework (Lee & See, 2004)	Appropriate trust, reliance, and compliance in human–machine systems	Defines behavioral outcomes of trust (calibration, overtrust, undertrust) and emphasizes balance between reliance and oversight	Behavioral Outcomes

Table 1: Theoretical Integration Underpinning the Cognitive Trust Formation Model for Human–AI Interaction

3.1. Antecedent Layer

The antecedent layer encapsulates the initial factors that predispose users to trust or distrust AI systems. These factors are largely perceptual and contextual. In [3], authors identified three foundational dimensions of trust i.e., ability, integrity, and benevolence, that together determine perceived trustworthiness. Translating this into AI contexts, the equivalents become perceived competence, perceived integrity, and perceived reliability. Competence refers to the AI's capability to perform tasks accurately and efficiently. Integrity represents the perceived consistency of behavior and adherence to explicit or implicit ethical standards. Reliability reflects consistent performance across time and conditions.

Additional antecedents include perceived transparency, user disposition to trust, and contextual risk. Perceived transparency describes how clearly users understand the AI's decision logic, while disposition to trust captures an individual's stable tendency to extend trust toward technological systems. Contextual risk refers to the perceived consequences of system failure within a given domain. These factors interact to set the cognitive baseline for initial trust evaluations.

In low-risk domains, heuristic judgments may suffice, whereas in high-risk settings, users require systematic reasoning and verifiable evidence of competence. The antecedent stage thus provides the foundation for cognitive processing by establishing initial expectations against which subsequent interactions are evaluated.

3.2. Cognitive Processing Layer

Once interaction begins, users engage in cognitive processing of trust cues. In [18], Heuristic–Systematic Model (HSM) provides the theoretical foundation for this stage. According to the HSM, individuals process information either through heuristic shortcuts, simple rules of thumb and affective cues, or through systematic analysis, which involves effortful cognitive elaboration. Applied to human–AI collaboration, heuristic processing might involve trusting an AI because its interface appears professional, its responses are immediate, or it communicates confidently. Systematic processing, by contrast, occurs when users scrutinize the AI's explanations, verify its outputs, or test its reasoning against their own domain knowledge.

These two processing modes interact dynamically. When cognitive load or time pressure is high, heuristic processing dominates. As familiarity increases or task stakes rise, systematic processing becomes more prominent. Social Cognitive Theory [19] reinforces this view by emphasizing observational learning and self-efficacy. Through repeated interactions, users learn the AI's behavior patterns, refine their understanding of its strengths and limitations, and adjust trust accordingly. Observed consistency, feedback quality, and corrective experiences inform cognitive schemas about when and how much to trust.

This dual-route processing explains variability in trust dynamics: users oscillate between fast intuitive reliance and slow analytical assessment, depending on situational and task demands. Over time, these processes converge into calibrated trust, trust that aligns closely with the AI's actual capabilities.

3.3. Behavioral Outcomes Layer

The behavioral outcomes layer reflects how trust translates into action. Model of Trust in Automation provides a key linkage by arguing that the goal of automation design should not be maximum trust but appropriate trust [4]. Users' behavioral expressions of trust include reliance (the degree to which they use AI outputs), delegation (the willingness to allow the AI to make autonomous decisions), and monitoring (the extent of human oversight). Trust calibration, the alignment between user trust and system performance, represents the ideal state.

However, two dysfunctional extremes can emerge: over-trust and under-trust. Over-trust occurs when users rely excessively on AI systems despite poor performance or limited understanding. Under-trust reflects skepticism or avoidance even when the AI is accurate and reliable. Both distort the efficiency and safety of human–AI collaboration. The behavioral outcomes layer therefore encompasses both desirable (calibrated trust, appropriate reliance) and undesirable (miss-calibrated trust) manifestations.

A feedback loop connects outcomes back to antecedents. Successful interactions reinforce perceived competence and reliability, strengthening trust antecedents. Conversely, failures weaken these perceptions, prompting reassessment. This cyclical mechanism ensures that trust is continuously updated rather than static. Figure 1 shows continuous feedback process with three interlinked layers: Antecedents, Cognitive Processing, and Behavioral Outcomes. This structure captures the iterative nature of human–AI trust evolution, reflecting how users' mental models are updated after each interaction.

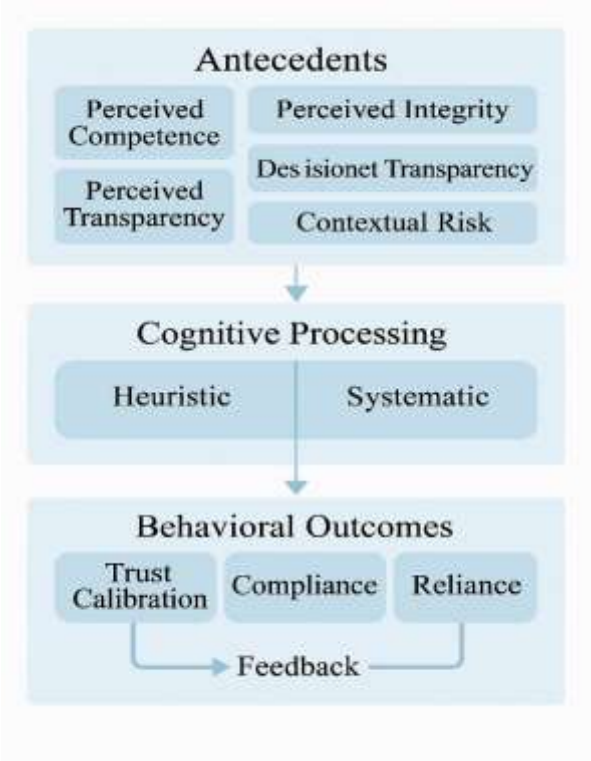

Figure 1: Cognitive Trust Formation Model for Human–AI Interaction

3.4. Integration of Theoretical Foundations

Integrating these theories yields a cohesive cognitive model of trust formation. In [3], authors contribute the structure for antecedent beliefs. In [18], authors explains the cognitive mechanisms through which users interpret trust-related information. In [19], learning and adaptation through repeated experience, accounting for feedback and self-efficacy was introduced. In [4], authors anchor behavioral outcomes, emphasizing calibration and reliance. Together, these frameworks describe trust as a recursive cognitive system: beliefs shape information processing, which drives behavior, which in turn reshapes beliefs.

The model also reconciles two seemingly opposing perspectives in trust research. The first treats trust as rational and evidence-based, assuming that users systematically evaluate system attributes. The second emphasizes affective and heuristic influences, highlighting that humans often rely on surface cues or analogies to human behavior. By embedding both routes within the cognitive processing layer, this framework acknowledges that trust in AI is neither purely rational nor purely emotional; it is situationally contingent and cognitively mediated.

3.5. Research Propositions

Building on this theoretical synthesis, several propositions can guide future empirical testing:

P1: Perceived competence, integrity, transparency, and reliability of an AI system positively influence initial cognitive trust.

P2: Contextual risk moderates the strength of the relationship between perceived system attributes and cognitive trust, such that higher risk contexts amplify the need for systematic processing.

P3: Heuristic processing dominates early or low-risk interactions, while systematic processing increases with task familiarity, cognitive involvement, and perceived consequence.

P4: Systematic processing leads to higher trust calibration than heuristic processing because it aligns trust more closely with system performance.

P5: Trust calibration positively influences adaptive behavioral outcomes such as appropriate reliance and delegation.

P6: Behavioral outcomes feed back into antecedent perceptions, such that positive experiences strengthen perceived competence and reliability, reinforcing future trust formation.

These propositions illustrate how trust unfolds as a learning process. Each stage informs the next, and the feedback loop ensures adaptation across time and context. By specifying these relationships, the model transitions from a conceptual abstraction to a testable theoretical structure.

The proposed framework extends existing literature in three key ways. First, it integrates fragmented theories from organizational behavior, psychology, and human factors into a single cognitive process model, providing a unified language for studying trust in human–AI systems. Second, it positions trust formation as iterative rather than linear, accounting for the continuous learning and recalibration inherent to AI interactions. Third, it reframes trust not as a design outcome but as a dynamic system property shaped by both human cognition and AI transparency. This perspective has implications for interface design, explainability strategies, and the governance of AI in safety-critical environments.

The theoretical synthesis also supports a more nuanced view of explainability and transparency. While these are often treated as static design attributes, the model suggests they should be adaptive, providing more detailed, systematic information when users engage in analytic reasoning, and concise heuristic cues when cognitive load is high. This adaptive transparency approach could optimize trust calibration and mitigate overreliance or rejection.

Finally, the feedback mechanism introduces a longitudinal dimension absent in many prior models. It acknowledges that trust in AI systems evolves with exposure that negative experiences have asymmetric effects compared to positive ones, and that organizational and social contexts shape this evolution. Empirical testing of these propositions across domains such as healthcare, autonomous vehicles, and decision-support systems can validate and refine the model.

4. Discussion and Implications

The Cognitive Trust Formation Model for Human–AI Interaction offers a structured way to understand how humans cognitively develop and adapt trust toward intelligent systems. The model reframes trust as a cyclical learning process rather than a static judgment, emphasizing that antecedents, cognitive processing, and behavioral outcomes interact continuously. This interpretation carries direct implications for AI system design, policy governance, and research agendas aimed at ensuring human–AI collaboration remains reliable, ethical, and adaptive.

At a theoretical level, the model clarifies that trust is not simply a by-product of technical performance or interface design. It emerges from how users interpret and internalize system cues over time. Perceived competence and integrity may initiate trust, but ongoing interactions, mediated by heuristic and systematic reasoning, determine whether that trust stabilizes, calibrates, or deteriorates. Recognizing this temporal dimension resolves the persistent tension in prior research between static trait-based and situational models of trust. It also underscores that cognitive feedback loops are central to maintaining appropriate reliance. When users experience consistency between expectations and outcomes, their mental representation of system reliability strengthens; when discrepancies arise, recalibration occurs. Designers and researchers must therefore treat trust as an evolving state that can be supported, disrupted, and repaired through design interventions.

4.1. ***Implications for Design***

For system designers, the framework suggests that building trustworthy AI is less about maximizing transparency and more about *calibrating* it to the user's cognitive state. Heuristic cues, such as tone, confidence indicators, and interface coherence should be intentionally crafted to establish positive initial impressions without creating false assurance. Once engagement deepens, systems should progressively enable systematic processing by revealing logic, uncertainty, and justification at the right level of granularity. Adaptive transparency mechanisms can monitor user engagement and modulate the depth of explanation dynamically. This ensures that novice users are not overwhelmed, while experienced users gain analytical insight necessary for accurate calibration.

Furthermore, design should support *trust repair*. When an AI makes an error or behaves unpredictably, immediate acknowledgment and contextual explanation can prevent disproportionate trust erosion. Interfaces can embed corrective

feedback channels that show learning or adjustment after failure. These small but visible signals help users perceive the AI as responsive, maintaining a sense of integrity and predictability.

The model also highlights the importance of *risk-sensitive design*. Trust cues that are acceptable in entertainment or information retrieval domains may be insufficient in high-stakes environments such as healthcare, aviation, or autonomous driving. Designers should match the transparency and validation level to the contextual risk profile, ensuring that cognitive workload and error tolerance are appropriately balanced.

4.2. ***Implications for Policy and Governance***

At the policy level, the model's emphasis on cognitive processing and feedback loops has consequences for how AI assurance, accountability, and certification are structured. Traditional governance frameworks often treat trust as compliance, ensuring the system meets technical or ethical checklists. However, the model shows that human trustworthiness perception depends equally on user experience, interpretability, and adaptive learning mechanisms. Regulators should therefore require that AI systems demonstrate not only performance accuracy but also explainability protocols that support both heuristic and systematic reasoning.

Policies could incorporate *trust calibration audits*, evaluating whether users' confidence aligns with actual system capability. Misalignment, such as persistent over-trust despite documented failure rates should trigger design modifications or user training requirements. Moreover, standards for AI transparency should shift from static disclosure documents to interactive, contextual transparency, interfaces that communicate uncertainty in real time.

From an organizational governance perspective, the model implies that accountability cannot rest solely on system developers. Organizations deploying AI systems must monitor post-deployment trust dynamics to detect signs of erosion or complacency. Establishing continuous feedback mechanisms, surveys, performance dashboards, and post-interaction trust metrics can help recalibrate both system design and user expectation.

4.3. ***Implications for Future Research***

The model opens several research trajectories. First, empirical work should operationalize the constructs defined in each layer i.e., competence, integrity, transparency, reliability, heuristic versus systematic processing, and trust calibration, and test their interactions longitudinally. Most existing trust research relies on single-session laboratory studies; the feedback loop proposed here demands time-series or field designs that capture adaptation.

Second, researchers should investigate how *adaptive transparency* and *dynamic explanation strategies* influence trust calibration across domains. For example, studies could compare static explanations to context-aware ones that change as user expertise grows. Eye-tracking, think-aloud protocols, or neurocognitive measures could reveal how users switch between heuristic and systematic reasoning during these interactions.

Third, the model invites exploration of *cultural and social moderators*. Disposition to trust and interpretation of system integrity may vary across cultural contexts, professional norms, or regulatory climates. Comparative research across geographies or sectors can test how these antecedents shift cognitive processing thresholds.

Fourth, there is a need to extend this framework to *multi-agent collaborations*. As AI systems increasingly work in teams, interacting with humans and other agents simultaneously, trust formation becomes distributed rather than individual. Understanding how collective trust emerges, diffuses, and recalibrates in such networks could advance both theory and practical governance.

Finally, future studies should link cognitive trust dynamics to measurable outcomes such as decision quality, task efficiency, and error recovery. Establishing these empirical links will demonstrate how calibrated trust directly contributes to system performance and human well-being.

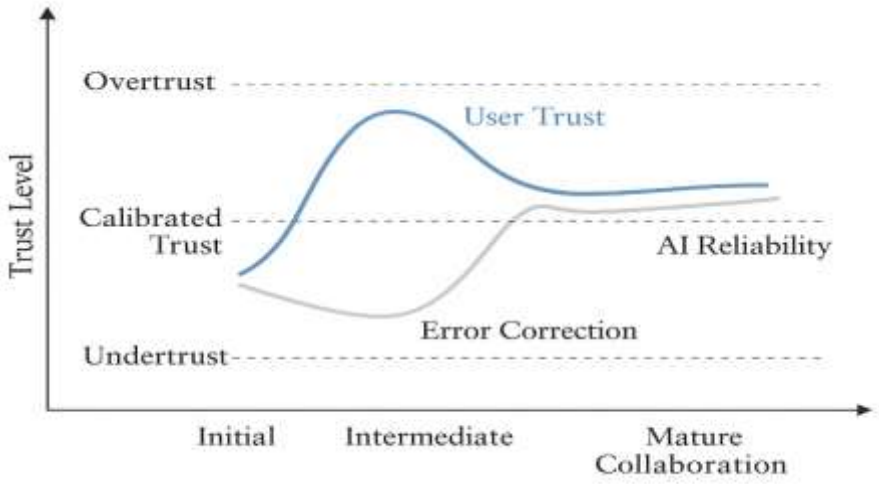

Figure 2: Trust Calibration Dynamics Over Time

5. Conclusion

This paper has developed the *Cognitive Trust Formation Model for Human–AI Interaction*, a theoretical framework explaining how humans form, calibrate, and adapt trust in intelligent systems. The model advances the understanding of trust as a cognitive and iterative process rather than a fixed disposition or one-time judgment. By integrating principles from organizational trust theory, social cognition, dual-process reasoning, and automation trust research, it offers a holistic perspective on how antecedents, cognitive processing mechanisms, and behavioral outcomes interact through continuous feedback.

The framework's key contribution lies in positioning trust as an evolving state shaped by both human cognition and system behavior. It challenges the traditional binary view of trust and distrust by introducing calibration as the optimal condition—trust that aligns with system capability and contextual risk. This shift has theoretical significance: it connects rational evaluation and affective intuition within a single cognitive continuum, bridging prior divides between psychological and engineering approaches to trust.

Practically, the model provides designers, developers, and policymakers with actionable insights. For design practice, it calls for adaptive transparency and dynamic explanation strategies that match user expertise and cognitive load. For governance and policy, it advocates the inclusion of trust calibration assessments and continuous post-deployment monitoring as part of AI assurance and certification frameworks. These implications extend the concept of "responsible AI" beyond compliance to include cognitive alignment between system and user.

From a scholarly perspective, the model outlines clear propositions for future validation, emphasizing longitudinal and context-sensitive methodologies. By doing so, it lays the foundation for an evidence-based understanding of trust evolution over time, something largely missing in existing cross-sectional research. The model's feedback structure also encourages researchers to explore how trust repair mechanisms, cultural norms, and social context influence adaptive trust formation.

Importantly, the framework invites a rethinking of what it means for AI to be "trustworthy." Trustworthiness is not solely an attribute of the system but an emergent property of the human–AI relationship. Systems cannot demand trust; they must sustain it through consistent, interpretable, and contextually appropriate behavior. Similarly, users must cultivate informed trust by engaging cognitively with explanations, limitations, and outcomes. This reciprocal relationship highlights the need for ongoing dialogue between design, ethics, and cognition in AI development.

In summary, the Cognitive Trust Formation Model reframes human–AI collaboration as a learning partnership grounded in cognitive adaptation. It provides the conceptual scaffolding for a next generation of research and practice that treats trust not as a static design goal but as a dynamic, measurable, and correctable process. Achieving this equilibrium—where human cognition and artificial reasoning evolve in concert—defines the future of responsible, effective, and truly collaborative AI.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] 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
[2] Gulati, S., McDonagh, J., Sousa, S., & Lamas, D. (2024). Trust models and theories in human–computer interaction: A systematic literature review. *Computers in Human Behavior Reports, 16*(3), 100495. https://doi.org/10.1016/j.chbr.2024.100495
[3] 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/10.5465/amr.1995.9508080335
[4] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392
[5] Kohn, S. C. (2021). Measurement of trust in automation: A narrative review. *Frontiers in Psychology, 12*, Article 604977. https://doi.org/10.3389/fpsyg.2021.604977
[6] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886
[7] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7
[8] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254
[9] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570
[10] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
[11] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Paper 582). Association for Computing Machinery. https://doi.org/10.1145/3173574.3174156
[12] Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2390–2395). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858402
[13] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033
[14] Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005
[15] Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372852
[16] Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04
[17] Wischnewski, M., Krämer, N. C., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state of the art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Paper 350). Association for Computing Machinery. https://doi.org/10.1145/3544548.3581197
[18] Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752
[19] Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Prentice-Hall.
[20] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392