Calibrated Trust in Agentic AI Work Systems: A Social Trust Calibration Framework for Responsible Human-AI Adoption

Kwan Hong TAN*

Singapore University of Social Sciences, Singapore

***Corresponding author**

DOI: https://doi.org/10.47772/IJRISS.2026.100700011

Received: 08 July 2026; Accepted: 13 July 2026; Published: 22 July 2026

ABSTRACT

Agentic artificial intelligence (AI) is transforming social and organisational life by moving AI from an assistive tool into a semi-autonomous actor that can plan, recommend, communicate, coordinate workflows and trigger decisions. Existing debates on trustworthy AI often emphasise technical reliability, legal compliance or ethical principles, but social adoption depends on a subtler question: how do people learn when to trust, distrust, verify or refuse AI outputs in everyday work? This conceptual research develops the Social Trust Calibration Framework (STCF) for responsible human-AI adoption in organisations and public institutions. Using an integrative conceptual synthesis of trust theory, automation research, organisational sensemaking, algorithmic management, human-centred AI and contemporary AI governance guidance, the study identifies the mechanisms through which AI trust becomes either calibrated, excessive, deficient or displaced. The resulting framework argues that trust in agentic AI must be calibrated across four mutually dependent layers: capability trust, process trust, institutional trust and identity trust. The paper contributes a formal definition of calibrated AI trust, a social calibration loop, a diagnostic matrix of trust states, a maturity model and seven research propositions for future empirical testing. The analysis shows that appropriate reliance cannot be achieved by accuracy alone. It requires visible evidence, role-specific verification routines, accountable decision rights, psychological safety, transparent escalation paths and protection of human agency. The paper concludes that responsible AI adoption should be treated as a social trust calibration problem rather than a simple technology acceptance problem. This reframing offers practical guidance for leaders, educators and policymakers seeking to scale AI while preserving human judgement, legitimacy and social confidence.

Keywords: agentic AI, social trust calibration, human-AI collaboration, organisational sensemaking, AI governance

INTRODUCTION

Artificial intelligence has entered a new social phase. In earlier organisational settings, AI was commonly framed as a predictive tool, a decision support system or an automation engine operating within bounded tasks. Contemporary generative and agentic AI systems are different. They generate language, summarise evidence, draft plans, coordinate software tools, interact with customers, analyse documents, initiate workflow steps and sometimes act across multiple organisational systems. As a result, the social question is no longer whether workers will use AI, but how workers, managers, institutions and publics can decide when AI is worthy of reliance.

The concept of trust has therefore become central to AI adoption. However, trust is often treated too simply. In many organisational discussions, trust is assumed to mean confidence in model accuracy or willingness to use a system. This is insufficient for agentic AI. A system may be accurate on benchmark tasks but inappropriate for a particular social context. It may be useful for drafting but unsafe for final decisions. It may increase productivity while weakening professional identity, public accountability or organisational legitimacy. A worker may trust AI too much because it is fluent, too little because it threatens expertise, or in the wrong way because

responsibility is unclear. These are not merely technical problems; they are social science problems involving cognition, authority, norms, identity, power and institutional confidence.

This paper addresses the problem of calibrated trust in agentic AI work systems. Trust calibration refers to the alignment between the trust granted to a system and the trust that is warranted by its demonstrated capability, the risk of the task, the quality of evidence, the clarity of accountability and the social meaning of reliance. Calibrated trust differs from high trust. High trust may be dangerous if it produces overreliance, automation bias or delegated responsibility without oversight. Low trust may also be harmful if it blocks useful augmentation, deepens resistance or produces hidden workarounds. The normative goal is not maximum trust but appropriate reliance.

The problem is timely because current AI governance frameworks increasingly emphasise trustworthy and human-centred AI, but organisations still struggle to translate these principles into everyday social practice. The NIST AI Risk Management Framework describes AI risk management in terms of governance, mapping, measurement and management, while the NIST Generative AI Profile extends these concerns to generative AI risks such as confabulation, privacy, harmful bias, information integrity and value chain exposure (NIST, 2023, 2024). The European Union AI Act similarly adopts a risk-based approach to AI obligations (European Parliament and Council of the European Union, 2024). These frameworks are important, but they do not by themselves explain how trust is socially produced, negotiated and recalibrated by workers facing AI systems in daily tasks.

This paper develops the Social Trust Calibration Framework (STCF) as an original conceptual contribution. The framework draws from social psychology, organisational studies, human-computer interaction, technology acceptance research, sensemaking theory, algorithmic management scholarship and AI governance. It also builds selectively on Tan's recent work on AI stakeholder recognition, distributed responsibility, the AI-form organisation, dynamic equilibrium ethics and technological displacement, while extending those ideas into a distinct social trust model (Tan, 2025a, 2025b, 2025c, 2025d, 2026a, 2026b).

The paper is guided by three research questions. First, what makes trust in agentic AI socially calibrated rather than merely high or low? Second, what organisational mechanisms produce appropriate reliance, overtrust, distrust or responsibility displacement? Third, how can organisations operationalise trust calibration in ways that protect human agency and support responsible adoption? The contribution is threefold. Conceptually, the paper defines calibrated trust as a multi-layered social condition rather than a single attitude. Theoretically, it integrates individual trust, process legitimacy, institutional governance and professional identity into one framework. Practically, it offers a calibration loop, maturity model and research propositions that can guide organisational diagnosis, policy design and empirical testing.

LITERATURE REVIEW

Trust as willingness to accept vulnerability

The most influential social theory of trust defines trust as the willingness of a party to be vulnerable to the actions of another party based on expectations about competence, benevolence and integrity (Mayer et al., 1995). This definition is useful for human-AI contexts because reliance on AI involves vulnerability. Workers may expose themselves to error, reputational harm, deskilling, surveillance or accountability for actions influenced by AI. Customers and citizens may also become vulnerable when organisations use AI to allocate services, screen applications, target communication or automate public decisions.

AI complicates the relational structure of trust. AI systems do not possess human motives in the ordinary social sense, yet users routinely attribute competence, neutrality, intentionality or helpfulness to them. Generative AI intensifies this tendency because conversational systems are socially fluent. They produce explanations, apologies and suggestions that mimic relational presence. In social science terms, trust in AI is therefore not simply a belief about technical performance. It is a socially mediated judgement about whether the system, the organisation deploying it and the surrounding governance arrangement can be relied upon.

Trust in automation and algorithmic advice

Research on automation shows that people can misuse automation through overreliance, disuse or inappropriate reliance (Parasuraman & Riley, 1997). Lee and See (2004) argue that trust in automation should be calibrated to the automation's capabilities. Hoff and Bashir (2015) further distinguish dispositional, situational and learned trust, showing that trust changes according to user characteristics, contextual demands and experience with the system. These insights are directly relevant to agentic AI because users often encounter AI in uncertain tasks where its competence is uneven.

Two strands of empirical work show why trust calibration is difficult. Algorithm aversion research suggests that people may reject algorithmic advice after observing errors, even when the algorithm remains statistically superior to human judgement (Dietvorst et al., 2015). Conversely, algorithm appreciation research shows that people sometimes prefer algorithmic advice, especially when they perceive algorithms as less biased or more consistent (Logg et al., 2019). These findings imply that AI adoption cannot be explained by a linear relationship between accuracy and trust. Trust depends on expectations, salience of error, perceived controllability, domain norms and comparative beliefs about human judgement.

Human-AI collaboration and the automation-augmentation paradox

Human-AI collaboration research increasingly recognises that AI may both augment and automate human work. Jarrahi (2018) argues that human-AI symbiosis requires combining machine analytical power with human judgement, contextual understanding and ethical reasoning. Raisch and Krakowski (2021) describe the automation-augmentation paradox: organisations seek productivity through automation while still needing human capabilities to manage exceptions, values and change. Agentic AI sharpens this paradox because it can perform more components of cognitive work while making the remaining human role more supervisory, interpretive and accountable.

This shift raises questions of professional identity. If AI performs drafting, analysis or coordination tasks previously used to signal expertise, workers may experience a threat to competence, status and meaning. Tan (2026a) describes this as ontological insecurity within the emergence of the AI-form organisation. The social problem is therefore not only whether AI works, but whether people can make sense of what their work becomes when AI participates in cognition, communication and coordination.

Algorithmic management and institutional trust

Algorithmic management literature shows that digital systems can allocate work, monitor performance, evaluate behaviour and shape managerial authority (Kellogg et al., 2020). When AI systems mediate organisational control, trust is not only interpersonal or technical; it is institutional. Workers ask whether the system is fair, whether data are used appropriately, whether decisions can be appealed, whether errors will be corrected and whether humans remain answerable. These concerns are also visible in policy debates on AI governance and risk-based regulation.

NIST (2023) frames trustworthy AI in terms of characteristics such as validity, reliability, safety, security, resilience, accountability, transparency, explainability, privacy and fairness. The NIST Generative AI Profile identifies additional risk areas that are intensified by generative AI, including confabulation, data privacy, harmful bias, information integrity and human-AI configuration (NIST, 2024). The EU AI Act creates legal obligations according to risk levels and places stronger requirements on high-risk systems (European Parliament and Council of the European Union, 2024). These developments show that trust is becoming institutionalised through standards, law and assurance procedures.

Sensemaking, legitimacy and psychological safety

Organisations do not adopt AI in a social vacuum. Workers interpret AI through existing narratives about competence, control, surveillance, fairness and future employability. Weick's (1995) sensemaking theory explains how people construct meaning under ambiguity by interpreting cues, enacting roles and negotiating

plausible explanations. AI adoption is a sensemaking event because it changes what counts as expertise, how decisions are justified and who is authorised to know.

Psychological safety is also relevant. Edmondson (1999) defines psychological safety as a shared belief that the team is safe for interpersonal risk taking. In AI adoption, psychological safety allows employees to question AI outputs, report errors, disclose uncertainty and challenge overconfident automation without fear of appearing incompetent or obstructive. Without psychological safety, trust may become performative. Workers may publicly accept AI while privately doubting it, or they may comply with AI recommendations because challenging the system feels risky.

Gap in the literature

The literature provides strong foundations, but several gaps remain. First, trust in AI is often treated as either an individual attitude or a property of system design, while less attention is given to the interaction between individual reliance, organisational process, institutional governance and professional identity. Second, AI governance scholarship often addresses compliance and accountability but does not fully explain how trust is socially learned and recalibrated in daily work. Third, technology acceptance models explain perceived usefulness and ease of use, but agentic AI requires attention to delegated agency, responsibility and meaning. Fourth, emerging work on AI as a stakeholder, distributed responsibility and AI-form organisations highlights the need for new governance structures, but these ideas require a more precise social trust mechanism (Tan, 2025a, 2025b, 2026a). The STCF addresses these gaps by treating trust calibration as a multi-level social process.

METHODOLOGY

Research design

This study uses an integrative conceptual research design. Conceptual research is appropriate when a phenomenon is emerging, theoretically fragmented and practically significant. Agentic AI adoption meets these conditions because organisations are deploying AI faster than social science theory has stabilised around its implications. The objective is not to test hypotheses using primary data, but to construct an analytically coherent framework that can guide future empirical research.

The method followed an abductive synthesis procedure. Abduction is suitable when researchers move between established concepts and new phenomena in order to generate explanatory constructs. Rather than deriving the STCF from a single theory, the study integrates multiple literatures around a practical puzzle: why do organisations experience both enthusiasm and resistance toward AI, and why does accuracy alone fail to produce responsible reliance?

Source selection

The conceptual synthesis drew on six bodies of literature. The first was trust theory, including organisational trust and interpersonal vulnerability. The second was automation and human factors research on reliance, misuse, disuse and calibration. The third was human-AI collaboration and technology acceptance research. The fourth was algorithmic management and organisational sensemaking scholarship. The fifth was AI governance guidance, including risk management, assurance and regulatory sources. The sixth was selected recent work by Kwan Hong Tan on AI stakeholder recognition, dynamic equilibrium ethics, technological displacement, the digital productivity paradox and the AI-form organisation, used where those works addressed responsibility, professional identity and organisational transformation.

Sources were included when they met at least one of four criteria: they offered a widely cited foundational construct, addressed AI trust or automation reliance, analysed organisational or institutional consequences of AI, or provided current governance guidance relevant to responsible AI adoption. The study did not claim to be a systematic review. Instead, it used a transparent conceptual synthesis method to build theory across disciplines.

Analytical procedure

The analysis proceeded in four stages. Stage one identified recurring constructs: competence, reliability, risk, vulnerability, transparency, accountability, procedural fairness, professional identity, psychological safety and institutional legitimacy. Stage two grouped these constructs according to level of analysis: individual, task, process, organisation and institution. Stage three compared trust failure modes, including overtrust, distrust, misplaced trust, performative compliance and responsibility displacement. Stage four synthesised these patterns into the STCF, calibration loop, trust-state matrix and research propositions.

A quality-control logic was applied to avoid overextension. First, constructs were retained only if they could be linked to more than one literature stream or to a clearly specified AI adoption problem. Second, the model distinguishes normative claims from empirical claims. It proposes what should be calibrated and how future studies may test it, but it does not claim empirical validation. Third, the framework avoids anthropomorphising AI. AI is treated as a socio-technical actor within organisational systems, not as a moral person.

Scope and limitations of the method

The paper focuses on knowledge work, organisational services, education, public administration and enterprise decision systems. It is less directly applicable to fully autonomous physical systems such as robotics, military systems or high-speed financial trading, although some principles may transfer. The analysis also concentrates on social trust rather than model evaluation. Technical testing remains essential, but the central claim is that technical reliability becomes socially meaningful only when embedded in accountable organisational practice.

RESULTS

Defining calibrated trust

The first analytical result is a definition. Calibrated AI trust is the socially organised alignment between the reliance granted to an AI system and the reliance warranted by the system's demonstrated capability, task risk, evidence quality, accountability structure and impact on human agency. This definition has five implications. First, trust is relational because it connects a user, an AI system, a task and an organisation. Second, trust is contextual because a system may be trustworthy for one task but not another. Third, trust is dynamic because it changes with evidence, incidents and experience. Fourth, trust is institutional because it depends on governance, appeal and accountability. Fifth, trust is identity-sensitive because reliance on AI changes how people understand their expertise and responsibility.

The STCF can be summarised through the following plain-language expression: calibrated trust equals warranted confidence plus accountable process plus institutional legitimacy plus identity security, minus task risk, uncertainty and responsibility diffusion. The formula is not proposed as a validated psychometric scale. It is a conceptual equation that helps organisations identify the conditions that must be present before AI reliance becomes socially responsible.

The four layers of the Social Trust Calibration Framework

The STCF identifies four layers of trust that must be aligned: capability trust, process trust, institutional trust and identity trust. Capability trust concerns whether the AI can perform a task reliably in the relevant context. Process trust concerns whether users can verify, challenge, escalate and correct AI outputs. Institutional trust concerns whether organisational rules, rights, audit trails and accountability structures are credible. Identity trust concerns whether AI adoption protects human agency, professional dignity and meaningful responsibility. These four layers are mutually dependent. Capability without process creates hidden risk. Process without institutional legitimacy creates paperwork without confidence. Institutional governance without identity trust creates compliance but not commitment. Identity trust without capability creates symbolic reassurance without performance.

Table 1: Core layers of the Social Trust Calibration Framework

Layer	Central question	Social risk if weak	Indicative controls
Capability trust	Can the AI perform this task in this context?	False confidence or premature rejection	Task testing, benchmark limits, user-facing accuracy evidence
Process trust	Can people verify, challenge and correct outputs?	Automation bias, hidden errors and workarounds	Checklists, escalation routes, peer review, audit logs
Institutional trust	Are rules, rights and accountability credible?	Legitimacy loss and responsibility gaps	Risk tiering, approval owners, appeals and governance review
Identity trust	Does AI preserve human agency and dignity?	Resistance, anxiety and professional displacement	Role redesign, psychological safety, reskilling and recognition

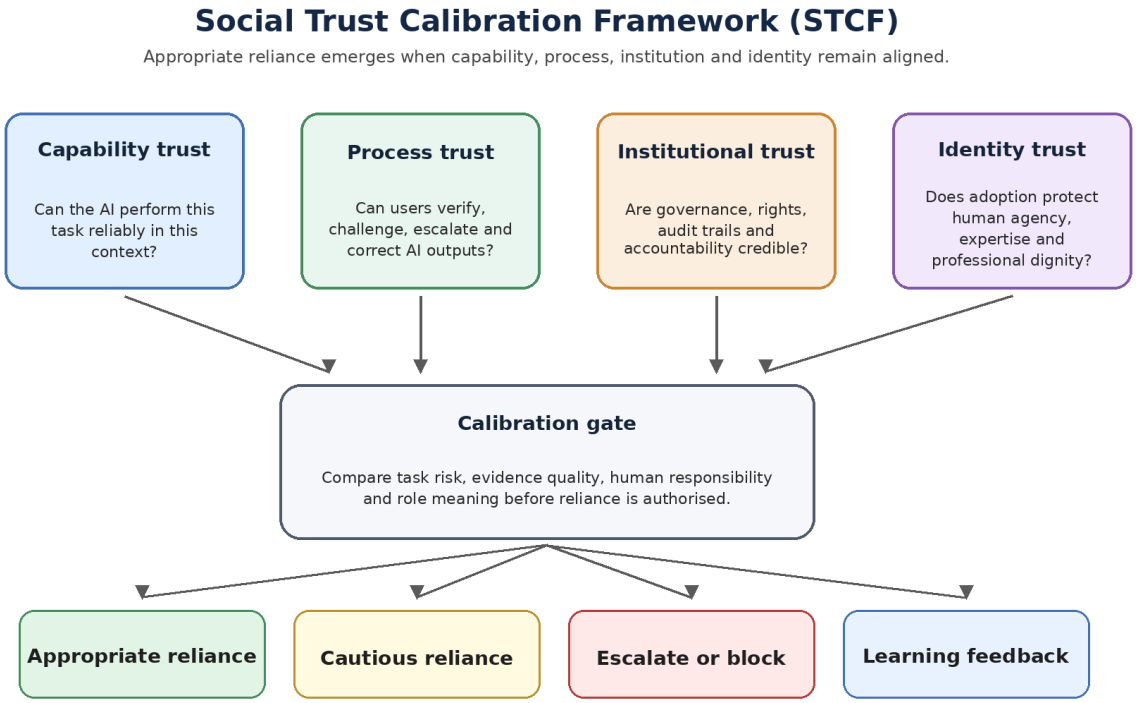

Figure 1: Social Trust Calibration Framework for agentic AI work systems

Figure 1 presents the framework. It shows the four layers feeding into a calibration gate. The gate compares the nature of the task, the evidence supporting the AI output, the accountable decision owner and the role meaning of reliance. The output is not a binary decision to use or reject AI. It may produce appropriate reliance, cautious reliance, escalation or learning feedback.

Trust states

The second analytical result is a typology of trust states. Appropriate reliance occurs when trust matches capability and task risk. Overtrust occurs when confidence exceeds warranted evidence. Distrust occurs when users reject useful AI despite adequate evidence. Misplaced trust occurs when users trust the wrong feature, such as fluency rather than validity. Displaced responsibility occurs when humans follow AI while psychologically or institutionally transferring responsibility to the system. Performative trust occurs when employees outwardly comply but privately avoid or work around the system. These states are analytically distinct but may coexist in the same organisation.

Table 2: Diagnostic trust states in agentic AI adoption

Trust state	Pattern	Typical organisational symptom	Corrective action
Appropriate reliance	Trust matches evidence and task risk	Users can explain when and why AI is used	Maintain monitoring and feedback
Overtrust	Confidence exceeds warranted capability	AI outputs accepted because they sound fluent	Add friction, verification and approval gates
Distrust	Useful AI is rejected despite evidence	Employees avoid AI or duplicate work manually	Build experiential learning and transparent evidence
Misplaced trust	Trust is attached to the wrong cue	Style, speed or authority is confused with validity	Teach evidence evaluation and source checking
Displaced responsibility	Humans follow AI while shifting blame	No one owns the decision when errors occur	Clarify decision rights and accountability
Performative trust	Public compliance hides private resistance	Shadow processes and unreported errors	Create psychological safety and open reporting

Calibration loop

The third analytical result is a six-step loop for social calibration. Organisations should first map the task, including risk, ambiguity and reversibility. They should then verify evidence, including data provenance, model limits and output checks. Third, they should assign accountability through clear owners, escalation paths and audit trails. Fourth, they should negotiate role meaning by clarifying what remains human, what becomes AI-supported and what skills require renewal. Fifth, they should authorise reliance according to risk tiers. Sixth, they should learn and recalibrate through feedback, incident review, training and policy adjustment. Figure 2 illustrates this loop.

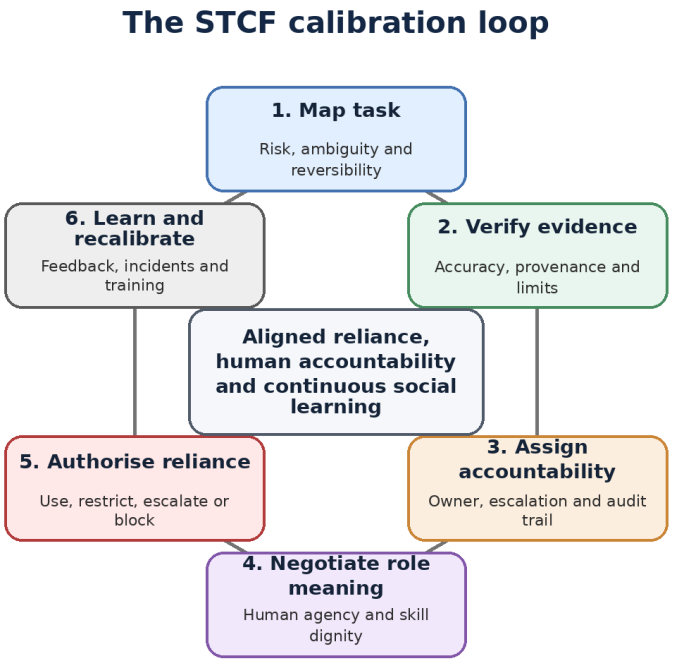

Figure 2: Six-step calibration loop for organisational AI adoption

Maturity model

The fourth analytical result is a maturity model. Many organisations begin at a tool-use level where employees use AI informally. They then move to procedural reliance, where verification routines and acceptable-use rules are introduced. The next stage is governed reliance, where risk tiering, audit evidence and accountable approval are institutionalised. The highest stage is calibrated symbiosis, where AI, human judgement, institutional assurance and professional development operate as a coherent system. The model suggests that trust maturity is not achieved by deploying more AI; it is achieved by aligning social practices around AI.

Table 3: STCF maturity model

Level	Dominant pattern	Trust risk	Development priority
Level 1: Informal tool use	Employees use AI individually	Uneven practice and hidden errors	Basic policy and disclosure norms
Level 2: Procedural reliance	Teams use verification routines	Rules may be inconsistent	Role-specific checklists and escalation
Level 3: Governed reliance	Risk tiering and audit trails are institutionalised	Governance may become bureaucratic	Proportional controls and incident learning
Level 4: Calibrated symbiosis	AI, human judgement and governance reinforce one another	Strategic dependency and identity risk	Continuous learning, role renewal and assurance

DISCUSSION

Theoretical implications

The STCF extends technology acceptance theory by shifting the focus from willingness to use to warranted reliance. Perceived usefulness and ease of use remain relevant, but they are insufficient in contexts where AI systems influence professional judgement, accountability and social legitimacy. The framework also extends trust in automation theory by placing calibration within organisations and institutions rather than only inside the user-system relationship. Finally, it extends organisational sensemaking theory by showing how AI adoption changes the meaning of competence and authority.

The framework has implications for stakeholder theory and AI governance. Tan's Agentic Stakeholder Ecosystem theory argues that advanced AI systems must be recognised as significant actors within value creation systems, while preserving human agency (Tan, 2025a). The STCF translates this insight into a trust mechanism. It does not require treating AI as a moral person. Instead, it requires mapping AI's functional agency and its consequences for stakeholders, then designing accountable processes around that agency. Similarly, Tan's work on distributed responsibility and dynamic equilibrium ethics supports the claim that responsibility must remain proportionate, transparent and adjustable under uncertainty (Tan, 2025b, 2026b).

Practical implications for organisations

For organisations, the framework suggests that AI adoption programmes should not begin with tools. They should begin with task classification. Low-risk drafting, ideation or summarisation may require light verification. High-impact decisions affecting employment, education, finance, healthcare, public services or legal rights require stronger controls, human approval and evidence trails. This risk-tiered approach is consistent with contemporary AI governance guidance, but STCF adds a social layer: employees must understand why controls exist and how they preserve rather than undermine professional judgement.

Second, organisations should design verification routines as work practices, not as vague warnings. Users need to know which outputs require source checking, peer review, domain expert approval or refusal. Third, organisations should establish escalation channels that are psychologically safe. Employees should be rewarded, not penalised, for identifying AI errors, reporting boundary cases or challenging outputs. Fourth, managers should address identity trust. Workers need to see how AI changes their roles, what skills remain valued and how accountability is fairly distributed. Ignoring identity threat may produce resistance even when the technical system is sound.

Implications for education and professional development

Universities, professional bodies and workplace educators should teach AI literacy as trust calibration rather than prompt proficiency. Learners should be trained to classify task risk, evaluate evidence, identify hallucination risk, document AI assistance, challenge outputs, explain human judgement and understand accountability. This is especially important for adult learners and professionals whose identity is tied to expertise. AI literacy should therefore combine technical, ethical and reflective dimensions.

Policy implications

For policymakers, the STCF suggests that trustworthy AI policy should evaluate whether organisations possess the social capacity to use AI responsibly. Compliance documents are not enough. Regulators and accrediting bodies should ask whether affected users can understand AI limitations, contest decisions, identify accountable humans and access remedies. Public trust in AI will depend not only on model quality but also on whether institutions demonstrate visible fairness, auditability and respect for human agency.

Limitations and future research

The main limitation of this paper is that the STCF is a conceptual framework and has not yet been empirically validated. Future research should develop survey scales for each trust layer, test the relationship between trust calibration and AI adoption outcomes, and conduct qualitative case studies of AI implementation in education, healthcare, finance, public administration and professional services. Longitudinal research is particularly important because trust changes over time as users encounter errors, successes, organisational messaging and regulatory events. Experimental studies can also examine whether calibration training reduces overtrust and distrust compared with standard AI literacy training.

CONCLUSION

This paper developed the Social Trust Calibration Framework for responsible human-AI adoption in agentic AI work systems. The central argument is that AI trust should not be maximised; it should be calibrated. Appropriate reliance arises when users and organisations align AI capability, process visibility, institutional accountability and human identity. Accuracy is necessary but insufficient. Without process trust, users cannot verify or challenge AI. Without institutional trust, reliance lacks legitimacy. Without identity trust, adoption may damage professional agency and produce resistance or performative compliance.

The STCF contributes a four-layer framework, a trust-state typology, a six-step calibration loop, a maturity model and seven propositions for future testing. It reframes responsible AI adoption as a social science problem involving vulnerability, meaning, authority and legitimacy. For organisations, the key lesson is that trust cannot be outsourced to the model. Trust must be designed into work practices, governance routines, professional learning and institutional accountability. The future of responsible AI will depend not only on more capable systems but on more calibrated societies of use.

RESEARCH PROPOSITIONS

The STCF generates propositions that can be tested through surveys, experiments, longitudinal case studies or mixed-method research. The propositions translate the conceptual model into empirically tractable relationships.

Table 4: Testable propositions derived from the STCF

Proposition	Expected relationship	Possible empirical indicators
P1	Capability evidence will predict appropriate reliance more strongly when task risk is explicitly disclosed.	Risk salience, output verification rate, reliance accuracy
P2	Process trust will reduce overtrust by increasing verification behaviour.	Checklist use, challenge frequency, error detection
P3	Institutional trust will moderate the relationship between AI adoption and perceived legitimacy.	Appeal availability, audit confidence, fairness perception
P4	Identity trust will predict employee willingness to use AI beyond perceived usefulness.	Role security, professional dignity, adoption intention
P5	Psychological safety will increase reporting of AI errors and boundary cases.	Incident reporting, escalation use, learning climate
P6	Clear accountability will reduce responsibility displacement in AI-supported decisions.	Decision-owner clarity, blame attribution, approval records
P7	Mature trust calibration will be associated with higher AI value realisation and lower AI incident rates.	Productivity gains, quality outcomes, incident frequency

Ethical Considerations

This study did not involve human participants, animal subjects, private organisational data or intervention with identifiable individuals. Formal institutional ethics approval was therefore not required. The manuscript relies on conceptual synthesis of published literature, public guidance documents and publicly accessible research outputs. The paper avoids fabricated empirical findings and presents the STCF as a theory-building contribution requiring future empirical validation.

Data Availability

No primary dataset was generated or analysed. The conceptual synthesis, framework components, tables and figures are contained within this manuscript. All cited sources are listed in the references.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

ACKNOWLEDGEMENTS

The author acknowledges the interdisciplinary scholarship on trust, automation, organisational sensemaking and responsible AI that informed the conceptual development of this paper.

REFERENCES

1. Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T
2. Chae, H. S., Kim, S., & Lee, J. (2025). Factors affecting human-generative AI collaboration: The roles of trust and perceived usefulness. Information, 16(10), 856. https://doi.org/10.3390/info16100856
3. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340. https://doi.org/10.2307/249008
4. 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
5. Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer.
6. Edmondson, A. (1999). Psychological safety and learning behavior in work teams. Administrative Science Quarterly, 44(2), 350-383. https://doi.org/10.2307/2666999
7. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union.
8. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1
9. Giddens, A. (1991). Modernity and self-identity: Self and society in the late modern age. Stanford University Press.
10. 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
11. 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
12. Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organisational decision making. Business Horizons, 61(4), 577-586. https://doi.org/10.1016/j.bushor.2018.03.007

13. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2
14. Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366-410. https://doi.org/10.5465/annals.2018.0174
15. 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
16. 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
17. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organisational trust. Academy of Management Review, 20(3), 709-734. https://doi.org/10.5465/amr.1995.9508080335
18. Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211022780
19. Milanez, A. (2025). Algorithmic management in the workplace. OECD Publishing.
20. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1
21. National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). https://doi.org/10.6028/NIST.AI.600-1
22. Orlikowski, W. J., & Scott, S. V. (2008). Sociomateriality: Challenging the separation of technology, work and organization. Academy of Management Annals, 2(1), 433-474. https://doi.org/10.5465/19416520802211644
23. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse and abuse. Human Factors, 39(2), 230-253. https://doi.org/10.1518/001872097778543886
24. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. Academy of Management Review, 46(1), 192-210. https://doi.org/10.5465/amr.2018.0072
25. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873
26. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118
27. Suchman, L. (1995). Making work visible. Communications of the ACM, 38(9), 56-64. https://doi.org/10.1145/223248.223263
28. Sun, H., Liu, W., Wu, D., Yu, G., & Yao, M. (2025). Revisiting trust in the era of generative AI: Factorial structure and latent profiles. arXiv. https://arxiv.org/abs/2510.10199
29. Tan, K. H. (2025a). Artificial intelligence as stakeholder: A novel framework for ethical recognition in value-creation ecosystems. ResearchGate.
30. Tan, K. H. (2025b). Dynamic equilibrium theory for ethical AI: Balancing epistemic uncertainty, human autonomy and social equity in high-stakes fluctuational decision systems. ResearchGate.
31. Tan, K. H. (2025c). The digital productivity paradox: A novel framework for understanding the asymmetric distribution of technological benefits. ResearchGate.
32. Tan, K. H. (2025d). The paradox of technological displacement: A novel framework for understanding AI and automation's impact on human capital development and labor market transformation. ResearchGate.
33. Tan, K. H. (2026a). Beyond automation: Sensemaking, ontological insecurity, and the emergence of the AI-form organisation in the digital era. British Research Review, 1(1). https://doi.org/10.67177/gywqbn07
34. Tan, K. H. (2026b). Agentic AI and the dissolution of managerial authority: A critical qualitative analysis of organisational sensegiving in the era of autonomous decision-making systems. ResearchGate.
35. UNESCO. (2023). Guidance for generative AI in education and research. UNESCO.

36. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425-478. https://doi.org/10.2307/30036540
37. Weick, K. E. (1995). Sensemaking in organizations. Sage Publications.
38. Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.