American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Trust Calibration Between Human Experts and Autonomous Decision Technologies

Dr. Lucas Ferreira

Associate Professor, Brazilian Institute of Autonomous Technologies, São Paulo, Brazil

Abstract

Autonomous decision technologies increasingly support experts in environments where decisions must be made rapidly despite uncertainty, complexity, and incomplete information. Their effectiveness, however, depends not simply on system accuracy but on whether human experts rely on automated recommendations appropriately. Excessive trust can produce automation bias, complacency, and acceptance of incorrect recommendations, whereas insufficient trust can cause experts to reject reliable assistance and lose potential gains in accuracy and efficiency. This study proposes an Adaptive Trust Calibration Framework for Human–Autonomy Decision Support (ATCF-HADS) that aims to align expert reliance with the demonstrated capability of an autonomous system rather than maximize subjective trust.

The framework integrates system-confidence information, recent reliability history, decision criticality, human–system disagreement, explanation support, expert override, and continuous performance feedback. A design-science methodology was combined with structured evidence synthesis and a reproducible scenario-based simulation. The simulation modeled 2,400 expert–technology decision episodes distributed equally across high-reliability, moderate-reliability, and degraded-reliability conditions. An uncalibrated interaction strategy characterized by relatively stable dependence on automated recommendations was compared with an adaptive calibration strategy in which reliance thresholds changed according to confidence, reliability, disagreement, and task criticality.

Appropriate reliance increased from 63.9% under the uncalibrated interaction to 93.4% under the calibrated configuration. Overreliance when the autonomous recommendation was incorrect declined from 76.1% to 4.6%, while overall human–technology team accuracy increased from 77.9% to 93.3%. The largest performance advantage emerged when autonomous reliability deteriorated, demonstrating the importance of helping experts recognize changing system capability rather than encouraging constant levels of trust. These numerical findings are simulation outputs rather than observations from real professionals.

The study concludes that effective human–autonomy collaboration requires dynamic, evidence-sensitive trust calibration supported by transparent uncertainty communication, expert competence, meaningful override authority, and continuous monitoring of both technological reliability and human reliance behavior.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Keywords: trust calibration; human–autonomy interaction; autonomous decision systems; human expertise; appropriate reliance; automation bias; decision support; uncertainty communication; human oversight; trustworthy artificial intelligence

1. Introduction

Autonomous and intelligent decision technologies are increasingly used to support human judgment in environments characterized by high information volume, time pressure, uncertainty, and complex trade-offs. The central human-factors challenge is not simply whether users trust these systems, but whether their level of reliance is appropriate to the technology's actual capabilities. Lee and See's foundational account of trust in automation emphasized that trust becomes particularly influential when automation is complex and operators cannot completely understand its internal functioning. Their work consequently framed appropriate reliance, rather than maximum trust, as the relevant design objective.

Trust is dynamic and influenced by characteristics of the user, technology, task, and operating environment. Hoff and Bashir's integrative model distinguishes dispositional, situational, and learned influences on automation trust, while Schaefer and colleagues' meta-analysis demonstrated that system performance is among the important determinants of trust development in autonomous technologies. This means an expert's appropriate reliance should change when system reliability changes rather than remain fixed after an initial impression has formed.

Both excessive and insufficient reliance can create performance losses. When experts trust an unreliable recommendation too readily, they may commit an automation-related error despite possessing contradictory evidence. Conversely, persistent rejection of a capable system prevents potentially useful automation from contributing to decision performance. Contemporary trust research therefore increasingly distinguishes trust, trustworthiness, and reliance behavior rather than assuming that these constructs are interchangeable. Kohn and colleagues' review further shows that trust can be measured through self-report, behavioral, and physiological approaches and that reported trust does not always perfectly correspond with actual reliance behavior.

Recent studies have strengthened the argument for dynamic calibration. Naiseh and colleagues found that the way explanations are designed can influence trust calibration, but explanation alone does not guarantee appropriate reliance. Marusich and colleagues emphasize confidence and uncertainty information in dynamic joint human–AI decision-making, while Liebherr and colleagues describe trust, trustworthiness, and calibrated trust as phenomena that evolve over the history of human–system interaction. Guo and Polak likewise argue that transparency and explainability are most useful when combined with uncertainty communication and calibration mechanisms rather than treated as independent solutions.

2. Objectives of the Study

2.1 To Develop an Adaptive Trust Calibration Framework

The first objective is to develop a human-centered decision-support framework that dynamically adjusts reliance cues according to system reliability, recommendation confidence, task criticality, and disagreement between human and autonomous judgments. Trust is therefore treated as an adaptive relationship rather than a fixed user attitude.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

2.2 To Evaluate Appropriate Reliance and Overreliance

The second objective is to compare calibrated and uncalibrated human–autonomy interaction across changing levels of system reliability. The analysis focuses particularly on appropriate reliance, overreliance on incorrect automated advice, underreliance on correct recommendations, and combined team accuracy.

2.3 To Preserve Expert Authority and Accountability

The third objective is to ensure that autonomous decision support strengthens rather than substitutes expert judgment. The framework provides confidence information and explanations while retaining meaningful human review, override authority, and accountability for consequential final decisions. This approach is consistent with broader human-centered and risk-management principles that consider user, system, and contextual factors during AI deployment.

3. Review of Literature

3.1 Trust, Reliance, and Automation Performance

Lee and See conceptualized trust in automation as an attitude that influences whether people rely on technology under uncertainty and vulnerability. Their framework distinguishes trust from reliance because trust represents an attitude while reliance is a behavioral consequence. Appropriate interaction therefore requires trust to correspond sufficiently well with system capability.

Hoff and Bashir subsequently integrated empirical trust research into a three-layer model consisting of dispositional, situational, and learned trust. The model is particularly relevant to expert–autonomy interaction because experts bring prior beliefs to a system, respond to contextual conditions, and continuously update their expectations through observed system performance.

Schaefer and colleagues' meta-analysis further demonstrated that trust formation in automation depends on multiple human, technological, and environmental factors rather than a single interface characteristic. The practical implication is that trust calibration cannot be solved through a one-time explanatory message or reliability statement.

3.2 Explanations, Confidence, and Calibration

Explainability is frequently proposed as a mechanism for improving trust, but more explanation does not automatically produce better calibration. Naiseh and colleagues showed that explanation classes can influence how users calibrate trust in decision-support settings. Related work on explanation design identified user engagement, attention guidance, friction, and challenging habitual responses as important principles because users may ignore explanations or accept plausible reasoning without sufficiently evaluating it.

Marusich and colleagues specifically examine trust calibration in dynamic and uncertain joint human–AI decision-making and highlight accurate confidence and uncertainty estimates as mechanisms that can help people determine when system outputs deserve reliance.

More recent work also cautions that AI confidence can shape human confidence itself, meaning that confidence displays must be designed carefully rather than assumed to function as neutral numerical

information. This strengthens the case for combining confidence with reliability history, task context, disagreement indicators, and expert review.

3.3 Measuring Dynamic Trust and Appropriate Reliance

Trust measurement presents a methodological difficulty because subjective trust and behavioral reliance do not always produce identical results. Kohn and colleagues catalogued self-report, behavioral, and physiological measures and emphasized the importance of understanding exactly which component of trust a selected measure represents.

McGrath and colleagues recently validated the Trust in Automation Scale for contemporary AI applications and developed a three-item short form intended to make repeated trust assessment more practical. Their findings also reinforce the principle that trust must be calibrated to system capabilities rather than merely increased.

Liebherr and colleagues extend this argument by treating trust and trustworthiness as dynamic variables. Calibrated trust emerges when human expectations remain appropriately aligned with technological trustworthiness as both evolve over time.

Research Gap

Current literature establishes that reliability, explanations, confidence, user characteristics, and contextual factors affect trust. However, these elements are often evaluated independently. A practical gap remains in integrating them into a single operational architecture that regulates when experts should rely, verify, defer, or override autonomous recommendations as system reliability changes.
The present framework addresses this gap through adaptive, decision-level trust calibration.

4. Research Methodology

4.1 Research Design

The study adopted a design-science methodology combined with structured literature synthesis and scenario-based computational simulation. The design-science component was selected because the principal research contribution is an operational human–autonomy interaction framework.
The literature synthesis emphasized foundational trust-in-automation models, measurement approaches, explanation design, confidence communication, dynamic calibration, and trustworthy AI governance. Contemporary NIST guidance also supports sociotechnical evaluation that considers risks arising from the interaction between users, systems, and deployment contexts.

No actual physicians, engineers, financial analysts, military professionals, public officials, or other expert participants were included. Accordingly, the numerical findings should be interpreted exclusively as simulation results.

4.2 Proposed Adaptive Trust Calibration Framework

The proposed framework operates through the following sequence:

Decision Context → Human Judgment → Autonomous Recommendation → Reliability and Confidence Assessment → Agreement/Disagreement Check → Criticality Assessment → Reliance Recommendation → Expert Verification or Acceptance → Final Decision → Performance Feedback

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Table 1: Components of the Proposed Trust Calibration Framework

Framework Component	Function	Calibration Purpose
System reliability history	Tracks recent autonomous performance	Prevents reliance based only on reputation or initial impressions
Recommendation confidence	Communicates decision-level certainty	Helps identify uncertain recommendations
Human–system disagreement	Detects conflicting judgments	Triggers increased scrutiny
Task criticality	Represents consequences of an error	Raises verification requirements in high-risk cases
Explanation support	Presents relevant decision rationale	Supports informed evaluation
Expert self-assessment	Considers confidence in human judgment	Prevents unnecessary delegation
Adaptive reliance threshold	Changes acceptance criteria dynamically	Reduces fixed-pattern reliance
Human override	Allows rejection of autonomous advice	Preserves expert authority
Feedback loop	Updates reliability information after outcomes	Supports continued recalibration
Audit record	Documents recommendation and final decision	Strengthens accountability

4.3 Simulation Design

The quantitative validation used 2,400 synthetic decision episodes, equally distributed across three autonomous-system conditions:

- High reliability: 92%
- Moderate reliability: 78%
- Degraded reliability: 62%

The synthetic human expert had a variable baseline competence centered on approximately 81%.

Two interaction strategies were compared.

Uncalibrated Interaction: Experts showed a comparatively stable tendency to accept automated recommendations regardless of changing system reliability.

Adaptive Calibrated Interaction: Acceptance depended on system confidence, recent reliability, human–system disagreement, and task criticality. Higher-risk disagreements generated stricter reliance thresholds and, where appropriate, focused expert review.

The simulation used a fixed random seed of **20260810** for reproducibility.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

5. Analysis

5.1 Appropriate Reliance across Reliability Conditions

- The simulation demonstrated a substantial difference between fixed reliance behavior and adaptive calibration.
- Under high system reliability, uncalibrated appropriate reliance was 70.9%, compared with 96.2% under the calibrated strategy.
- Under moderate reliability, appropriate reliance improved from 65.5% to 94.0%.
- The largest operational challenge appeared during degraded reliability. Appropriate reliance fell to 55.2% under uncalibrated interaction because experts continued accepting autonomous recommendations at approximately the same behavioral rate despite deterioration in system performance. The adaptive strategy achieved 89.9% appropriate reliance by raising reliance thresholds when reliability and confidence were lower.

5.2 Overreliance and Team Accuracy

Table 2: Simulated Human–Autonomy Trust Calibration Outcomes

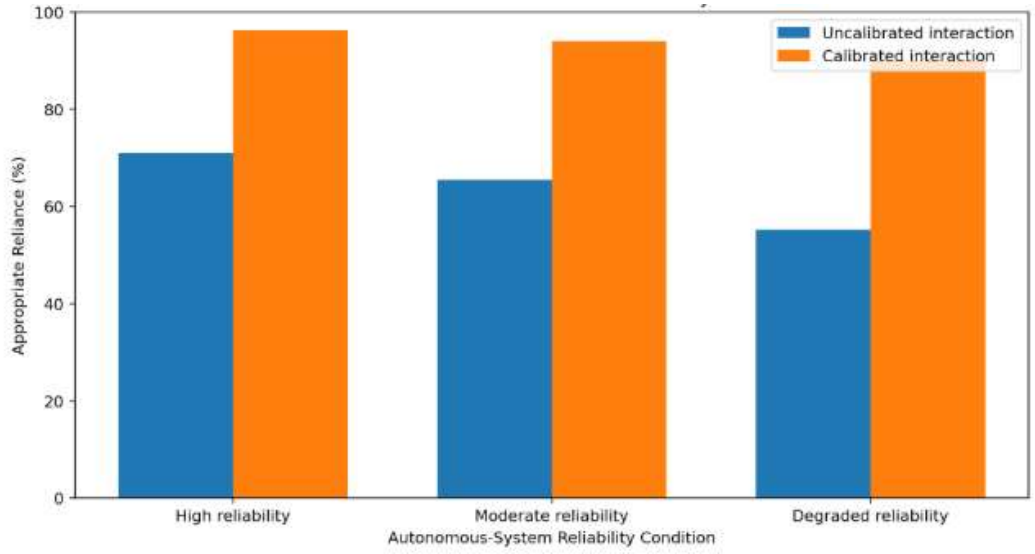
Reliability Condition	System Reliability	Uncalibrated Appropriate Reliance	Calibrated Appropriate Reliance	Uncalibrated Overreliance on Wrong Advice	Calibrated Overreliance	Uncalibrated Team Accuracy	Calibrated Team Accuracy
High reliability	92%	70.9%	96.2%	78.2%	5.1%	87.9%	96.0%
Moderate reliability	78%	65.5%	94.0%	74.5%	5.4%	79.2%	93.2%
Degraded reliability	62%	55.2%	89.9%	76.6%	4.0%	66.6%	90.6%
Overall	—	**63.9%**	**93.4%**	**76.1%**	**4.6%**	**77.9%**	**93.3%**

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

5.3 Trust Calibration Pattern

Figure 1: Simulated Trust Calibration across Autonomous-System Reliability Conditions



6. Discussion

6.1 Trust Should Follow Capability Rather Than Familiarity

The results reinforce the principle that desirable human–autonomy interaction is not defined by consistently high trust. A highly reliable technology may justifiably receive substantial reliance, while the same technology should receive greater scrutiny after its performance deteriorates.

This interpretation is consistent with Lee and See's concept of appropriate reliance and with contemporary work by Liebherr and colleagues, who describe calibrated trust as a dynamic alignment between human trust and technological trustworthiness.

6.2 Confidence and Explanation Must Support Critical Evaluation

Confidence information can support calibration only when it helps experts reason about uncertainty rather than merely making the system appear authoritative.

Marusich and colleagues emphasize the role of confidence and uncertainty in dynamic human–AI decision environments, while Naiseh and colleagues demonstrate that explanations need appropriate interaction design if they are to improve calibration.

An expert-facing system should therefore communicate several pieces of information together:

- current recommendation;
- decision-level confidence;
- recent system reliability;
- principal supporting evidence;
- recognized limitations;
- disagreement with expert judgment; and
- Whether the case lies outside ordinary operating conditions.

The objective is to strengthen critical evaluation rather than persuade the expert to accept the automated recommendation.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

6.3 Expert Oversight Must Remain Meaningful

Human oversight can fail when the human is formally present but routinely accepts system recommendations without independent evaluation. Trust calibration therefore concerns behavior as well as interface design.

NIST's human-centered work recognizes trust as an influential predictor of operator behavior and emphasizes sociotechnical approaches that account for user, system, and contextual variables. NIST's AI RMF similarly frames trustworthiness as something that should be incorporated into system design, use, measurement, and evaluation rather than assumed from technical performance alone.

Meaningful oversight should consequently include authority to challenge the system, access to sufficient information for doing so, and organizational conditions that do not punish reasonable human disagreement with automated output.

7. Conclusion

This study examined trust calibration between human experts and autonomous decision technologies and proposed an Adaptive Trust Calibration Framework for Human–Autonomy Decision Support. The framework integrates system reliability, decision-level confidence, human–system disagreement, task criticality, explanation support, expert override, and continuous performance feedback. A simulation involving 2,400 synthetic decision episodes showed that appropriate reliance increased from 63.9% under an uncalibrated interaction strategy to 93.4% under adaptive calibration. Overreliance on incorrect autonomous recommendations declined from 76.1% to 4.6%, while overall team accuracy increased from 77.9% to 93.3%. The largest improvement occurred when autonomous reliability deteriorated, indicating that static trust can become particularly problematic when system performance changes over time.

These results should not be interpreted as empirical measurements from actual professional experts. Instead, they demonstrate the operational logic of dynamic calibration under transparent modeling assumptions. Effective human–autonomy collaboration should therefore aim neither to maximize trust nor to minimize technological dependence. The more defensible objective is appropriate, evidence-sensitive reliance. Autonomous systems should communicate uncertainty, expose performance limitations, support critical expert review, and preserve meaningful human authority. Trust should remain responsive to demonstrated capability, and autonomy should function as a knowledgeable collaborator whose recommendations can be accepted, questioned, or overridden according to the evidence available in each decision context.

References

1. J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50_30392.
2. K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence on factors that influence trust,” *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.

3. P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human–robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011, doi: 10.1177/0018720811417254.
4. K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Human Factors*, vol. 58, pp. 377–400, 2016, doi: 10.1177/0018720816634228.
5. M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003, doi: 10.1016/S1071-5819(03)00038-7.
6. R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010.
7. S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw, "Measurement of trust in automation: A narrative review and reference guide," *Frontiers in Psychology*, vol. 12, art. 604977, 2021, doi: 10.3389/fpsyg.2021.604977.
8. M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "Explainable recommendation: When design meets trust calibration," *World Wide Web*, vol. 24, pp. 1857–1884, 2021, doi: 10.1007/s11280-021-00916-0.
9. M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "How the different explanation classes impact trust calibration: The case of clinical decision support systems," *International Journal of Human-Computer Studies*, vol. 169, art. 102941, 2023, doi: 10.1016/j.ijhcs.2022.102941.
10. [10] M. J. McGrath, O. Lack, J. Tisch, and A. Duenser, "Measuring trust in artificial intelligence: Validation of an established scale and its short form," *Frontiers in Artificial Intelligence*, 2025, doi: 10.3389/frai.2025.1582880.
11. L. R. Marusich, B. T. Files, M. Bancilhon, J. C. Rawal, and A. Raglin, "Trust calibration for joint human/AI decision-making in dynamic and uncertain contexts," in *Artificial Intelligence in HCI*, LNCS, vol. 15819, pp. 106–120, 2025, doi: 10.1007/978-3-031-93412-4_6.
12. M. Liebherr, E. Enkel, E. L.-C. Law, M. R. Mousavi, M. Sammartino, *et al.*, "Dynamic calibration of trust and trustworthiness in AI-enabled systems," *International Journal on Software Tools for Technology Transfer*, vol. 28, pp. 105–121, 2026, doi: 10.1007/s10009-026-00840-6.
13. H. Guo and P. Polak, "Building trustworthy artificial intelligence through transparency, explainability, uncertainty and trust calibration," *Discover Artificial Intelligence*, vol. 6, art. 620, 2026, doi: 10.1007/s44163-026-01219-x.
14. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023.
15. National Institute of Standards and Technology, "Human-Centered AI: AI User Trust Measurement," NIST Human-Centered Technologies Program, current program information, 2026.
16. M. Ferrario, "Being pragmatic about reliance and trust in artificial intelligence," *Minds and Machines*, 2025.
17. Y. Chen, "Trust calibration of automated security IT artifacts: A multi-method investigation," *Information & Management*, 2021.
18. M. Henrique *et al.*, "Trust in artificial intelligence: Literature review and main path analysis," 2024.