

Human-In-The-Loop Vs. Fully Autonomous Enterprise Ai Agents: A Trust Calibration Framework For Governance, Control, And Decision-Making

Aiswarya Gurram

Independent Researcher, USA.

Mail Id: gurramaiswarya98@gmail.com

Article Received 29-07-2026, Revised 22-08-2026, Accepted 14-09-2026

Author(s) Retains the Copyrights of This Article

Abstract

This study develops an Enterprise AI Agent Trust Calibration Framework (EAATCF) that can be used to establish proper human control and autonomy of enterprise AI agents. The framework incorporates reliability, decision risk, explainability, reversibility and accountability, and is tested on synthetic data, machine-learning classification, dynamic recalibration, and robustness. Findings identify reliability and risk as determinants of dominant autonomy. While the intermediate levels of autonomy occur more often compared to the extremes of control. Conditional robustness to weights, distributions and thresholds are shown by sensitivity analyses, though there is a restriction to extreme-level prediction in terms of imbalance of minority classes. EAATCF brings in a computationally consistent, context sensitive framework of matching trust, governance, control and responsible enterprise decision-making.

Keywords: Enterprise AI Agent Trust Calibration Framework (EAATCF), Enterprise AI Agents, Trust Calibration, Human Oversight, Autonomy, Reliability, Decision Risk, Explainability, Reversibility, Accountability, Dominant Autonomy, AI Governance.

I. Introduction

Enterprise AI Agents and the Shift Toward Autonomy

Enterprise AI agents are transforming the organization from a traditional AI-supported decision support system to an autonomous system that can reason, plan, and execute tasks [1]. It also interacts and communicates over an enterprise platform. Such transition allows the increased efficiency of its operations, scalability, and timeliness of decisions. Nevertheless, the rise of autonomy promotes issues too concerning trust, accountability, transparency, as well as human accountability, and thus, governance is a key organizational need.

Research Problem

AI agents in the enterprise acquiring the power of autonomous decision-making, organizations are confronted by the challenge of deciding on the appropriate levels of human intervention [2]. Overarching of the human factor can decrease its efficiency, and extreme autonomy can escalate the operational, ethical, and compliance risks [3]. The lack of a systematized trust calibration system poses an ambiguity of the balancing of governance, control and credible decision-making in varied enterprise situations and setups.

Research Aim

This study aims to develop a trust calibration framework for appropriate human oversight and autonomy levels, determining enterprise AI agents, to enhance governance, control, accountability, and decision-making in organizations.

Research Objectives

- To examine the factors of trust influencing human dependence on enterprise AI agents.
- To compare governance and control needs of Human-in-the-Loop and fully autonomous AI agents.
- To assess the impact of risk, reliability, explainability and accountability on proper AI autonomy.
- To develop a dynamic trust calibration framework that promotes responsible enterprise AI agent decision-making and governance practices.

Novel Contribution of the Study

This study introduces the Enterprise AI Agent Trust Calibration Framework (EAATCF) to combine trust calibration, governance, risk, explainability and accountability into a cohesive approach to enterprise AI autonomy [4]. Such a framework contrasts with the current research, where human control or autonomous agents are analyzed separately [5]. Whereas a dynamic approach is proposed to decide the context-specific

amount of human control and autonomous decision-making in enterprise environments.

II. Related Work

Trust and Human Reliance on Enterprise AI Agents

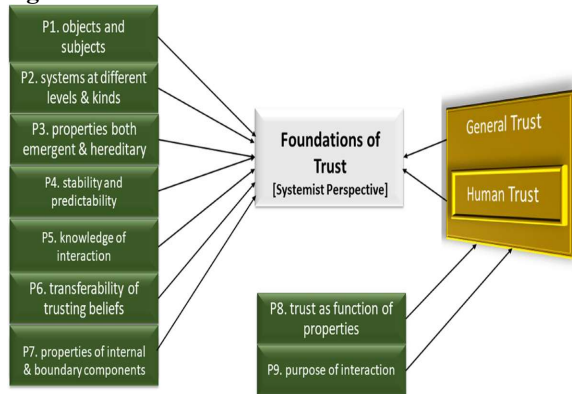


Fig 1. Foundational Trust Framework

Literature defines trust as a condition to be applied to AI dependence however usually takes it as a static concept and does not take into account the contextual difference [6]. Trust is affected by accuracy, transparency, predictability and competence [7]. However there has been little research on how these attributes interact with agents of autonomous enterprises [8]. Automation bias is a threat of over-trust and beneficial delegation is a limitation of under-trust [9]. The literature has focused primarily on traditional types of decision-support systems, which have generated little knowledge on how organizations are supposed to tune human dependence based on evidence of demonstrated agent capability, situational uncertainty, outcome of the decision and an emerging situation of operational changes.

Governance and Control in Human-in-the-Loop versus Fully Autonomous AI Agents

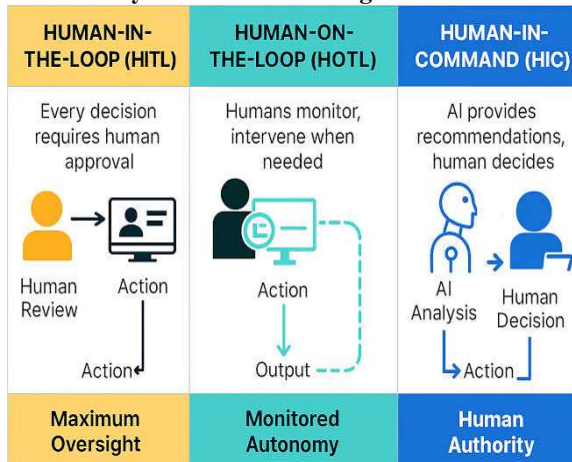


Fig 2. Human-in-the-Loop vs Fully Automated AI Systems

AI governance focuses on oversight, accountability, transparency and control; however, the aspect of Human-in-the-Loop code can prove ineffective in cases where the intervention can be untimely or cognitively challenging [10]. Fully autonomous agents enhance scalability and efficiency whilst making the independent actions harder to be held responsible [11]. The current literature often considers these models individually and does not take into consideration supervisory or conditional autonomy [12]. This means that comparative research is still needed to identify the context-sensitive governance arrangements that balance between significant human control, autonomous ability, efficiency of the operation, and careful decision making by the enterprise in various organizational settings [13].

Risk, Reliability, Explainability, and Accountability in Determining AI Autonomy

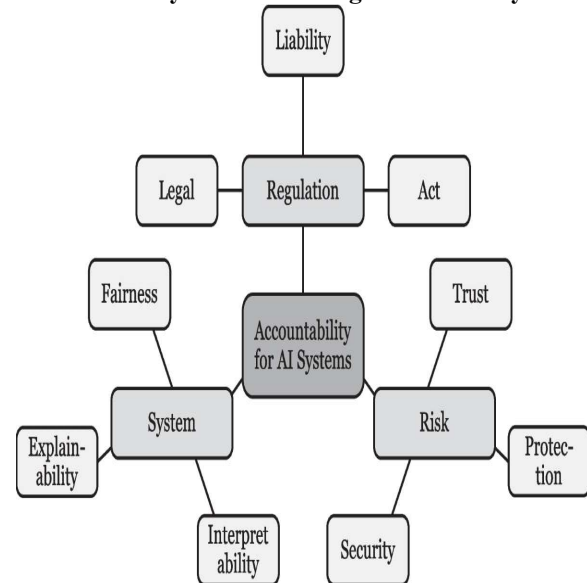


Fig 3. The Present and Future of Accountability for AI Systems

Literature identifies Reliability, risk, explainability, and accountability are known as critical factors to AI adoption, yet they are often investigated separately in literature [14]. Delegation can be justified by reliability, and cannot be used solely to determine autonomy in cases of large organizational impacts [15]. Explainability enhances scrutiny however it does not ensure accountability and proportionate oversight is demanded by risk-based governance [16]. There is not much literature synthesizing these dimensions and decision reversibility [17]. As a result, a unified system is needed to decide on contextually proper levels of autonomy of enterprise AI agents.

Dynamic Trust Calibration for Enterprise AI Agent Decision-Making

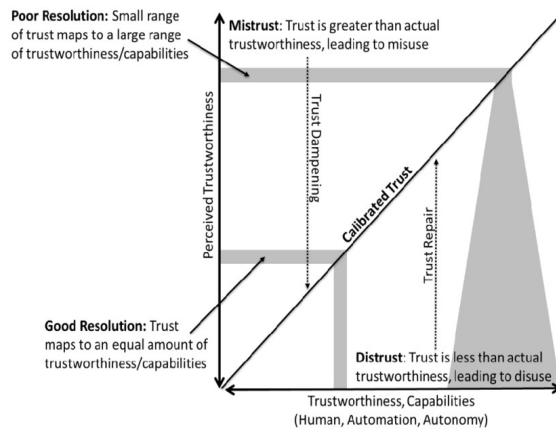


Fig 4. Model of trust calibration and resolution between trust and trustworthiness Adaptive

Calibration of trust entails congruence between human trust and shown AI ability [18]. The current studies focus mainly on personal relations with automation without taking into account perpetually running enterprise agents that become more reliable, more complex, more uncertain and riskier. This can render the implementation of static autonomy assignments inappropriate [20]. Even though the field of governance studies has focused on monitoring and oversight, there is a lack of integration between performance feedback and autonomy adjustment, leaving a gap in theoretical foundations of dynamic frameworks involving trust as well as risk regulation, governance and control [21].

Research Gap

The current literature lacks the sufficient integration of trust, governance, risk, reliability, explainability and accountability in the determination of the autonomy of enterprise AI agents [22]. Studies tend to compare Human-in-the-Loop systems and purely autonomous systems following a dichotomy that does not necessarily consider intermediate control systems, nor adaptive conditions of their operation [23]. Therefore, there is still a great space to bridge a dynamic trust calibration model that can connect the array of circumstantial risk, agent ability, human control, and corporate determination-making needs successfully.

III. Enterprise Ai Agent Trust Calibration Framework (EAATCF)

Conceptual Foundations of the EAATCF

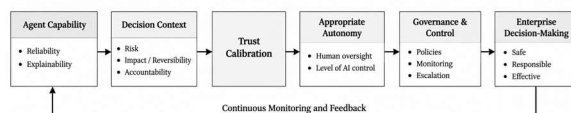


Fig 5. Conceptual Architecture of the Enterprise AI Agent Trust Calibration Framework (EAATCF)

EBATCF is a combination of trust-in-automation, human-AI trust, AI governance, and organizational decision making to distribute human-AI decision authority [24]. It operationalizes trust as measured trust, facilitated by governance, accountability and situational risk evaluation. Relevant enterprise AI autonomy is thus contingent on shown agent competence, uncertainty, outcomes, along with risk attributes of a decision-context.

Dimensions of Trust Calibration

Table 1. Core Dimensions Of Enterprise Ai Agent Trust Calibration

Dimension	Meaning	Role in Autonomy
Agent Reliability	Consistency and accuracy of agent performance	Higher demonstrated reliability may support greater autonomy
Task/Decision Risk	Potential consequences of failure	Higher risk generally requires stronger oversight
Explainability /Transparency	Ability to understand and evaluate agent actions	Supports informed reliance and scrutiny
Decision Impact/Reversibility	Severity and recoverability of erroneous decisions	Irreversible consequences justify stronger control
Accountability Requirements	Responsibility for decisions and outcomes	Strong accountability obligations can constrain autonomy

The framework incorporates five aspects: agent reliability denotes regular performance; decision risk denotes organizational effects; explainability allows assessing agency actions; decision effect and reversibility identify how to recover on the errors; using accountability brings about responsibility [25]. These dimensions have a combined influence in not only the proper autonomy but also the human surveillance needed.

Risk-Reliability Trust Calibration Matrix

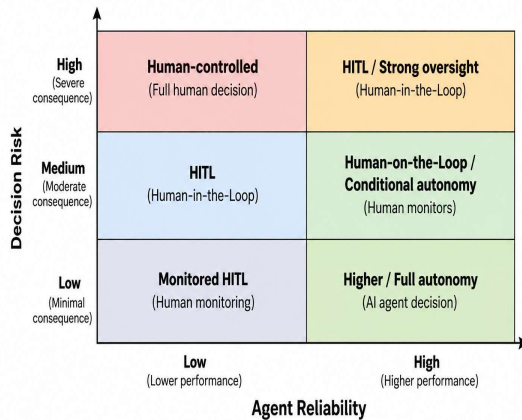


Fig 6. EAATCF Risk-Reliability Trust Calibration Matrix

The autonomy is calibrated in the calibration matrix based on decision risk and agent reliability. Low-Risk, high-risk decisions can only be managed by humans whereas greater reliability means that HITL can be involved. High-reliability tasks, such as medium-level risk settings or conditional autonomy, can be supported by HITL in contexts with low risk, whereas larger autonomy is feasible [26]. These theoretical groupings need to be empirically proven within the framework of enterprises.

Levels of Human-AI Control

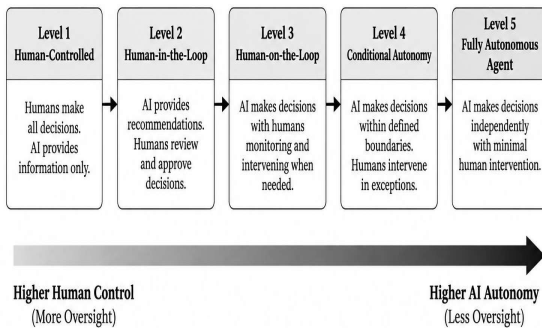


Fig 7. Enterprise AI Agent Autonomy and Human-Control Continuum

The EAATCF has five levels of autonomy: Human-Controlled, Human-in-the-Loop, Human-on-the-Loop, Conditional Autonomy, and Fully Autonomous Agent [27]. It uses this continuum to replace binary human-AI control; this allows enterprises in a more nimble way to proportionately delegate autonomy based on capability, risk, explainability, reversibility, and accountability and thus facilitates context-sensitive governance without necessarily incentivizing more autonomy.

Dynamic Governance, Monitoring and Escalation

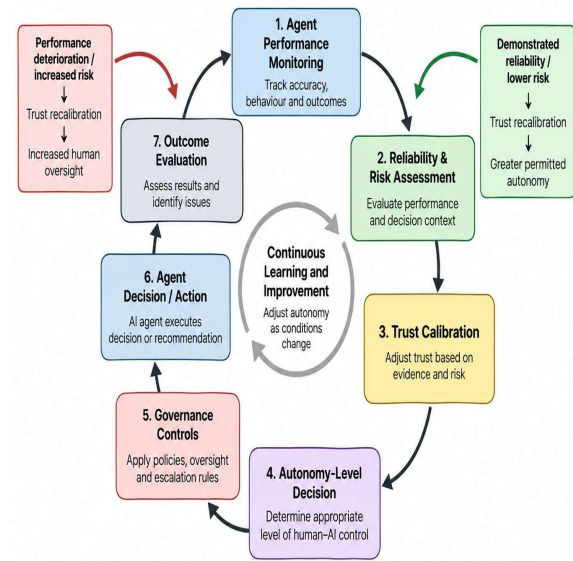


Fig 8. Dynamic Trust Calibration and Governance Feedback Cycle

The approach to autonomy remains adaptive by continuously monitoring, recalibration of trust and feedback of governance. When performance is becoming worse or riskier, it is time to escalate and have more human control over it. However, when proven high reliability occurs in a low-risk environment, more freedom may be allowed [28]. This dynamic process aligns decision authority, accountability, monitoring and control with changing conditions in the enterprise.

Development of Research Propositions

There are five relationships which are put forward by EAATCF: P1, reliability enhances delegated authority; P2, decision risk increases human oversight; P3, explainability enhances calibrated reliance; P4, reversibility moderates trust and autonomous action; and P5, accountability moderates' reliability and allowed autonomy [29]. These propositions are collectively operationalization's of the empirical testing framework.

IV. Research Methodology

Research Philosophy and Approach

This research is pragmatic and based on theory in computing and how to develop EAATCF with a combination of the theoretical expertise in human -AI trust, governance, automation, and organizational decision-making coupled with synthetic experimentation [30]. It analyses the presence of appropriate systematically determined enterprise AI autonomy under controlled computational

circumstances without primary human-subject data on the basis of theoretically defined trust dimensions.

Research Design

The research has a conceptual-computational design which entails the development of theoretical frameworks, generation of synthetic data, machine-learning modelling and comparison analysis. The EAATCF has five key variables of input, which include agent reliability (R), decision risk (D), explainability (E), reversibility (V) and accountability requirements (C). The target variable, autonomy level (A) has five categories:

$$A = \{1, 2, 3, 4, 5\}$$

Where Human-Controlled, Human-in-the-Loop, Human-on-the-Loop, Conditional Autonomy and Fully Autonomous Agent are denoted as 1, 2, 3, 4, and 5 respectively.

Synthetic Data Generation

Experimental scenario decision making of an enterprise is computationally derived by varying the five EAATCF dimensions, in the absence of recognizable human or organizational information. All the input variables are scaled to a range of zero to one:

$$0 \leq R, D, E, V, C \leq 1$$

The latent trust-calibration score is initially specified as.

Trust Calibration Score:

$$T_c = w_R R + w_E E + w_V V - w_D D - w_C C$$

Where T_c can be defined as calibration of trust; R, reliability; E, explainability; V, reversibility; D, decision risk and C, accountability. w denotes sensitivity tested weights that finds effects. w_R , w_E , w_V , w_D and w_C are theoretically assigned and these weights are sensitivity-tested. Autonomy is more justified with reliability, explainability and reversibility and risk and accountability inhibit autonomy consequently increased T_c increases autonomy of the agent.

Operationalization of Constructs

Reliability is the performance of agents, decision risk is the severity of failure, explainability is interpretability and reversibility are recoverability, and accountability is responsibility in organizations. Autonomy is represented as:

$$A = f(R, D, E, V, C)$$

T_c thresholds levels have the benefit of assigning levels of autonomy, whereas alternate distributions and weights tested under assumptions have their stability under T_c tests.

Pseudocode

```

Algorithm 1
Program for EAATCF Trust Calibration and
Autonomy Assignment

Initialize
    Input R, D, E, V, C
    Set theoretical weights
    Set autonomy thresholds
    Generate synthetic scenarios
Start Loop
    Normalize R, D, E, V, C

    Calculate Trust Calibration Score
     $T_c = w_R(R) + w_E(E) + w_V(V)$ 
     $- w_D(D) - w_C(C)$ 

    IF  $T_c \leq T_1$  THEN
        Assign Level 1: Human-Controlled
    IF  $T_1 < T_c \leq T_2$  THEN
        Assign Level 2: Human-in-the-
Loop
    IF  $T_2 < T_c \leq T_3$  THEN
        Assign Level 3: Human-on-the-
Loop
    IF  $T_3 < T_c \leq T_4$  THEN
        Assign Level 4: Conditional
Autonomy
    IF  $T_c > T_4$  THEN
        Assign Level 5: Fully Autonomous
Agent

    Train MLR, Random Forest and
XGBoost
    Evaluate Accuracy, Precision, Recall and
Macro-F1
    Analyze Feature Importance and SHAP

    Test Weight, Distribution and Threshold
Sensitivity

    IF Reliability decreases OR Risk
increases THEN
        Recalibrate Trust
        Increase Human Oversight
    ELSE IF Reliability increases AND Risk
remains low THEN
        Recalibrate Trust
        Permit Greater Autonomy
    END IF

End Loop
Output Appropriate EAATCF Autonomy
Level
    
```

Fig 9. Pseudocode

The EAATCF is operationalized by the pseudocode as it calculates trust calibration using 5 dimensions, and allocates the correct level of autonomy, evaluating machine-learning performance, and checks robustness, and provides machine-learned human oversight when reliability or risk varies.

Machine-Learning Models and Analysis

The comparison is conducted on three supervised machine-learning models: Multinomial Logistic Regression as a baseline, which can be interpreted, and Random Forest and XGBoost as nonlinear relationships and variable interactions, respectively. The artificial dataset will consist of training and testing data sets. Accuracy, precision, recall and F1-score is used to measure performance.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Where TP = True Positive; FN = False Negative; FP = False Positive; $F1_1$ – $F1_5$ = F1 score of autonomy Level 1-5. To evaluate overall balanced performance in case of the five classes of autonomy, it is additionally be evaluated using:

$$\text{Macro } F1 = (F1_1 + F1_2 + F1_3 + F1_4 + F1_5) / 5$$

Feature importance and SHAP (SHapley Additive exPlanations) which evaluates the strengths of reliability, risk, explainability, reversibility, and accountability in model prediction and whether they are supported by the propositions in EAATCF.

Robustness, Validity and Ethical Considerations

The robustness is tested by synthetic simulations, cross-validation, a sensitivity test, and alternative distributions. The construct validity is based on grounded EAATCF variables. Synthetic data minimize confidentiality risks. Results do not constitute empirical validation, however computer-based behavior instead. Assumptions, weights, thresholds, model limitations and simulation biases are reported.

V. Analysis And Discussion

Trust Calibration and Autonomy Structure

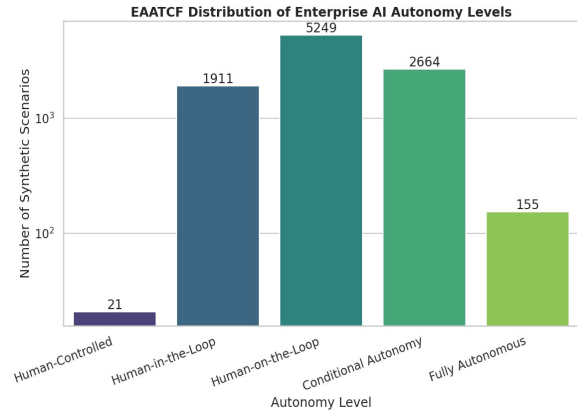


Fig 10. EAATCF Distribution of Enterprise AI Autonomy Levels

There are 5,249 Human-on-the-Loop, 2,664 Conditional Autonomy and 1,911 HITL, 155 Fully Autonomous and lastly only 21 Human-Controlled scenarios in the distribution. This bending towards intermediate levels is a clear pointer to the notion that in most cases, calibrated enterprise governance is inclined to lean towards supervised or conditional autonomy in relation to the opposite extremes of control.

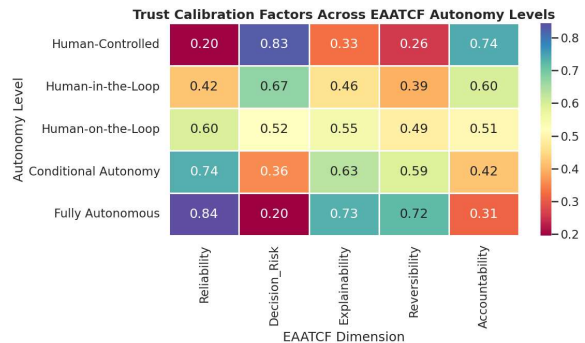


Fig 11. Trust Calibration Factors Across EAATCF Autonomy Levels

The heatmap indicates that reliability increases with increased autonomy between 0.20 and 0.84 and the decision risk decreases between 0.83 and 0.20. The explainability increases 0.33-0.73, reversibility 0.26-0.72, and accountability decreases 0.74-0.31, which broadly supports levels of autonomous decision authority increasing statistically.

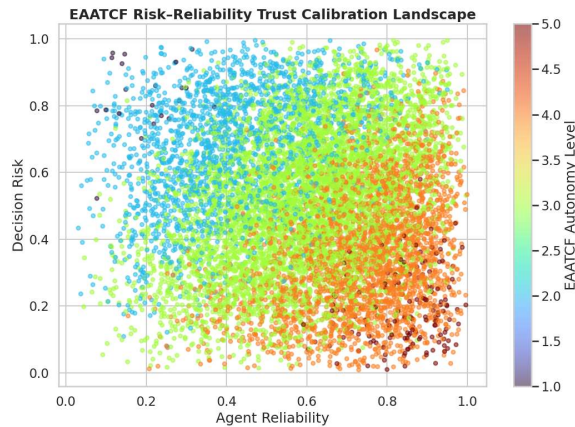


Fig 12. EAATCF Risk-Reliability Trust Calibration Landscape

The topography shows increased autonomy is concentrated where reliability is great and the risk of decision is minimal. Levels 4-5 are mostly found in cases where the reliability approaches 0.7-1.0 with less risk but Level 1-2 is found in more cases with a lower degree of reliability or high levels of risk.

Predictive Performance and Class Imbalance

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
0 Multinomial Logistic Regression	0.9905	0.7936	0.7456	0.7658
1 Random Forest	0.9250	0.7451	0.7032	0.7213
2 XGBoost	0.9410	0.7585	0.6965	0.7225

Fig 13. Test-Set Machine-Learning Performance

The Multinomial Logistic Regression has a precision of 0.9905, macro-precision of 0.7936, macro-recall of 0.7456 and macro-F1 of 0.7658. Random Forest records 0.9250, 0.7451, 0.7032, and 0.7213; XGBoost records 0.9410, 0.7585, 0.6965, and 0.7225, showing MLR provides the strongest overall test-set performance among models evaluated here.

Comparative Performance of EAATCF Machine-Learning Models



Fig 14. Comparative Performance of EAATCF Machine-Learning Models

MLR leads with 0.9905 accuracy and 0.7658 macro-F1, compared with Random Forest at 0.9250 and 0.7213, and XGBoost at 0.9410 and 0.7225. Its accuracy, precision, recall and F1 are all simultaneous which aids in restoring the synthetic EAATCF classification structure.

Model	Accuracy mean	std	Macro Precision mean	std
Multinomial Logistic Regression	0.9891	0.0024	0.7924	0.0015
Random Forest	0.9181	0.0053	0.8984	0.0549
XGBoost	0.9386	0.0078	0.8375	0.1130

Model	Macro Recall mean	std	Macro F1 mean	std
Multinomial Logistic Regression	0.7460	0.0172	0.7651	0.0115
Random Forest	0.7528	0.0287	0.7950	0.0262
XGBoost	0.6950	0.0406	0.7362	0.0519

Fig 15. Five-Fold Cross-Validation Performance Summary

Across five folds, MLR averages 0.9891 accuracy and 0.7651 macro-F1, Random Forest 0.9181 and 0.7950, and XGBoost 0.9386 and 0.7362. The standard deviation of the accuracy of MLR is substantially low (0.0024), whereas other macro measurements significantly indicate that the sensitivity to minority autonomy classes in general is greater.

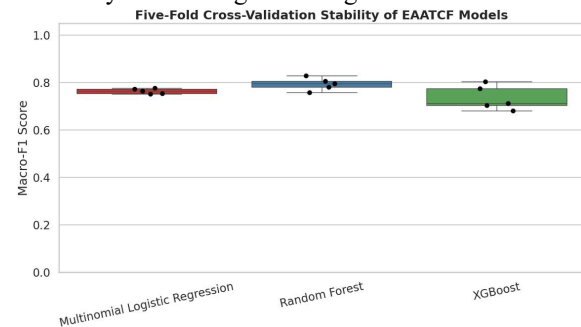


Fig 16. Five-Fold Cross-Validation Stability of EAATCF Models

Random Forest has the largest mean macro-F1 of 0.7950 and; this is in contrast to MLR and XGBoost of 0.7651 and 0.7362 respectively. Nevertheless, MLR has the least macro-F1 standard deviation 0.0115, which is a clear indicator of a higher fold-to-fold stability; although overall, the average performance of Random Forest is better in all cases, and is more associated with class balances.

Table 2. Per-Class Imbalance Performance

Class	Support	Precision	Recall	F1
Human-Controlled	4	0.000	0.000	0.000
HITL	382	0.990	0.997	0.993
Human-on-Loop	1,050	0.993	1.000	0.997
Conditional	533	0.985	0.989	0.987

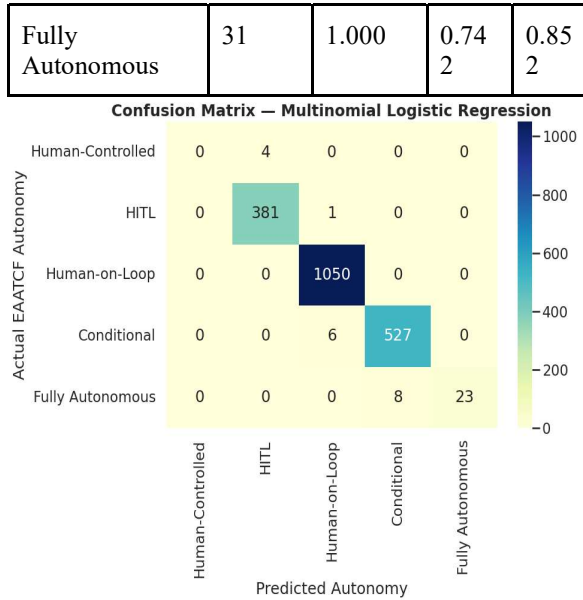


Fig 17. Confusion Matrix for Multinomial Logistic Regression

The model precisely forecasts 381 HITL, 1,050 Human-on-Loop, 527 Conditional and 23 Fully Autonomous predictions. However, all the Human-Controlled cases are categorized as HITL and the Eight Fully Autonomous cases would be Conditional, and show less strong discrimination of the few extreme autonomy classes.

Drivers, Dynamics and Robustness

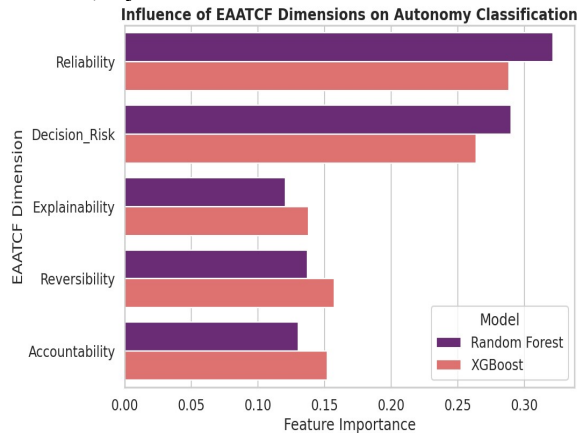


Fig 18. Influence of EAATCF Dimensions on Autonomy Classification

Random Forest places the highest values in reliability, 0.32 and decision risk, 0.29; XGBoost also places relatively the same values, 0.29 and 0.26. The degrees of explainability, reversibility, and accountability are approximately around 0.12-0.16, which can be viewed as strong evidence on the motivations to reliability and risk that the autonomy classification decisions are overall.

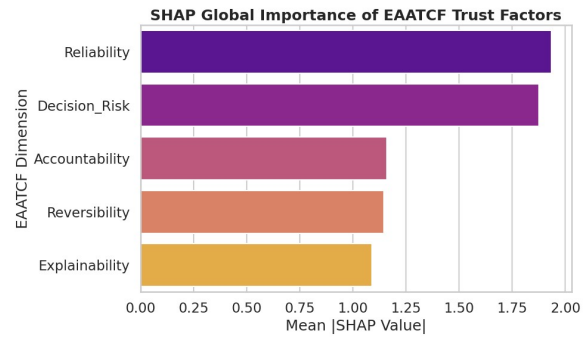


Fig 19. SHAP Global Importance of EAATCF Trust Factors

The consideration of reliability has the greatest mean absolute SHAP value of 1.93 with the next consideration of decision risk 1.87, accountability at 1.16, reversibility at 1.14 and explainability at 1.09. The independent confirmation of reliability and risk as the overall best predictive drivers of the framework is established in this ordering.

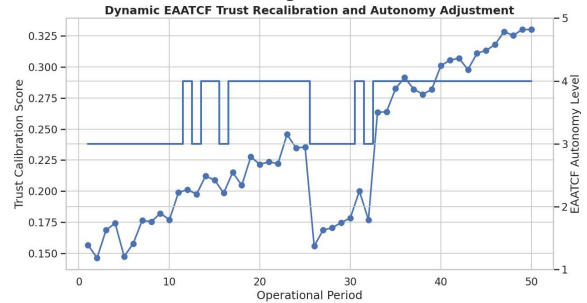


Fig 20. Dynamic EAATCF Trust Recalibration and Autonomy Adjustment

The Trust calibration is increased by a rate between 0.15 and 0.33 across 50 periods of operation but it decreases drastically between period 26–32 before recovering. Autonomy is in its turn an alternate in between levels 3 and 4, which proves that the performance decline may activate the weaker autonomy and stronger oversight.

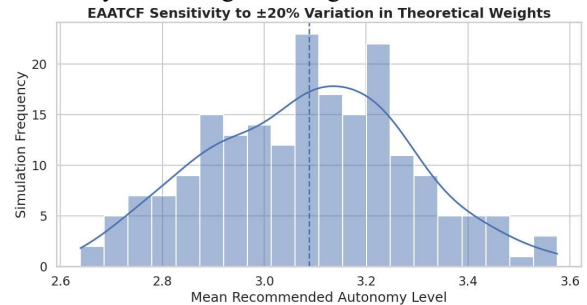


Fig 21. EAATCF Sensitivity to $\pm 20\%$ Variation in Theoretical Weights

Through 200 simulations with theoretical weights with $\pm 20\%$ variation in each direction the mean recommended autonomy is near 3.1 and ranges between 2.65-3.57. The focus of Level 3 suggests

moderate sensitivity, rather than instability and therefore supporting strong autonomy suggestions despite realistic uncertainty with regard to the weighting of theoretical weighting elements.

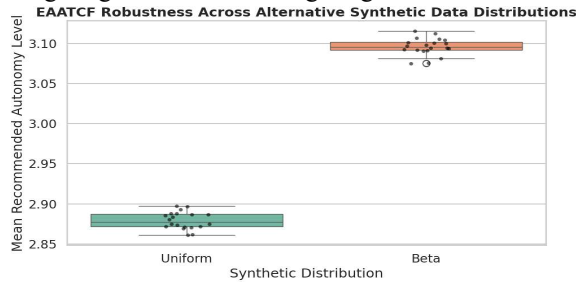


Fig 22. EAATCF Robustness Across Alternative Synthetic Data Distributions

Mean recommended autonomy of uniform simulations are 2.87-2.90, and those of Beta simulations clusters are all around 3.08-3.12. Repeatability is shown by narrow spreads per distribution, however the difference at the 0.21-approximate level between distributions establishes EAATCF results to be sensitive to the synthetic data-generation assumptions and distributions.

Table 3. Proposition-Testing Summary Table

Proposition	Computational evidence	Interpretation
P1	Reliability: SHAP 1.93	Consistent
P2	Risk: SHAP 1.87	Consistent
P3	Explainability: SHAP 1.09	Consistent
P4	Reversibility: SHAP 1.14	Consistent
P5	Accountability: SHAP 1.16	Consistent

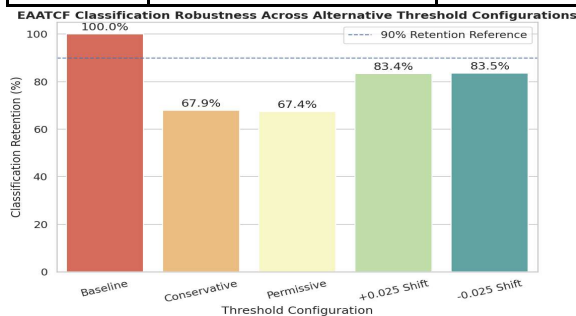


Fig 23. EAATCF Classification Robustness Across Alternative Threshold Configurations

This figure shows that there is moderate threshold sensitivity. Base retention is 100%, while conservative and permissive configurations retain 67.9% and 67.4% of classifications. Smaller shifts of ± 0.025 maintain 83.4-83.5% showing more stability at more modest changes and larger (materially, significant) threshold adjustments have significant effects on EAATCF autonomy assignments and enterprise governance results.

Discussion

The results indicate the computational consistency of EAATCF and show crucial limitations. The core autonomy (79.13%) prevails with reliability and risk being key drivers. MLR has a 99.05% accuracy however is less effective on the extremes of minorities. The interpretations of dynamic recalibration and weight, distribution and threshold sensitivity justify the adaptive oversight and the conditional robustness, respectively. Together, it can be seen that P1-P5 produce profiles of computational consistency that contribute to risk-sensitive governance as opposed to free autonomy of the enterprise.

VI. Conclusion And Future Research

Conclusion

This study develops EAATCF as an adaptable model of human control and enterprise AI autonomy and how to calibrate it. Computational evidence points to reliability and risk being the main determinants of autonomy with explainability, reversibility and accountability being the other sources of governance restrictions. Conditional stability is demonstrated through sensitivity analysis and robustness analysis which advocates adaptive, context-specific and responsible enterprise AI decision-making.

Future Research

Further research needs to be carried out on empirical evaluation of EAATCF by the use of actual deploys of enterprise agents in industries, settings and regulatory requirements. The longitudinal studies need to be conducted of trust and autonomy changing over time and other calibration thresholds and governance mechanisms have to be validated. Future studies might include behavioral, ethical, cybersecurity, and economic considerations to enhance the applicability of the frameworks.

References

- [1]. Xu, G., Murthy, S.V. and Jia, B., 2025, December. Enhancing intuitive decision-making and reliance through human-AI collaboration: A review. In *Informatics* (Vol. 12, No. 4, p. 135). MDPI.
- [2]. Merchán-Cruz, E.A., Gabelaia, I., Savrasovs, M., Hansen, M.F., Soe, S., Rodriguez-Cañizo,

- R.G. and Aragón-Camarasa, G., 2025. Trust by design: An ethical framework for collaborative intelligence systems in Industry 5.0. *Electronics*, 14(10), p.1952.
- [3]. Fernández-Miguel, A., Ortiz-Marcos, S., Jiménez-Calzado, M., Fernández del Hoyo, A.P., García-Muiña, F.E. and Settembre-Blundo, D., 2025. Agentic AI in smart manufacturing: Enabling human-centric predictive maintenance ecosystems. *Applied Sciences*, 15(21), p.11414.
- [4]. Srinivas, S., Kirk, B., Zendejas, J., Espino, M., Boskovich, M., Bari, A., Dajani, K. and Alzahrani, N., 2025. AI-augmented SOC: A survey of LLMs and agents for security automation. *Journal of Cybersecurity and Privacy*, 5(4), p.95.
- [5]. Desai, S., 2025. Agentic AI frameworks: Building autonomous, self-healing systems for financial infrastructure. *Journal of Computer Science and Technology Studies*, 7(12), pp.364-383.
- [6]. Brohi, S., Mastoi, Q.U.A., Jhanjhi, N.Z. and Pillai, T.R., 2025. A research landscape of agentic AI and large language models: applications, challenges and future directions. *Algorithms*, 18(8), p.499.
- [7]. Arumilli, S.V.T. (2025) 'Maintaining Human Oversight Within Agentic Artificial Intelligence-Driven Workflows for Enterprise Operational Decision-Making and Automated Business Process Optimization Systems', *Journal of Computational Analysis and Applications*, 34(10), pp. 530–538.
- [8]. Romero-Obon, M., Rouaz-El-Hajoui, K., Sancho-Ochoa, V., Vargas, R., Pérez-Lozano, P., Suñé-Pou, M. and García-Montoya, E., 2025. Human-in-the-loop AI use in ongoing process verification in the pharmaceutical industry. *Information*, 16(12), p.1082.
- [9]. Holowaychuk, T.J., 2025. Building Trusted Enterprise Ecosystems through Explainable Artificial Intelligence Secure Automation and Intelligent Decision Support. *International Journal of Engineering & Extended Technologies Research (IJEETR)*, 7(6), pp.11279-11289.
- [10]. Patricio, L., Varela, L., Silveira, Z., Felgueiras, C. and Pereira, F., 2025. A framework for integrating robotic process automation with artificial intelligence applied to industry 5.0. *Applied Sciences*, 15(13), p.7402.
- [11]. Hokmabadi, H., Rezvani, S.M., Hokmabadi, H. and de Almeida, N.M., 2025. Business resilience through AI-agent automation for SMEs and startups: a review on agile marketing and CRM. *Information*, 16(11), p.1000.
- [12]. Kumar, R., 2023. Cloud-integrated AI framework for transaction-aware decision optimization in agile healthcare project management. *International Journal of Computer Technology and Electronics Communication*, 6(1), pp.6347-6355.
- [13]. Frenkenberg, A. and Hochman, G., 2025. It's scary to use it, it's scary to refuse it: the psychological dimensions of AI adoption— anxiety, motives, and dependency. *Systems*, 13(2), p.82.
- [14]. Pathak, A. and Bansal, V., 2024. AI as decision aid or delegated agent: The effects of trust dimensions on the adoption of AI digital agents. *Computers in Human Behavior: Artificial Humans*, 2(2), p.100094.
- [15]. Marocco, S., Barbieri, B. and Talamo, A., 2024. Exploring facilitators and barriers to managers' adoption of AI-based systems in decision making: a systematic review. *AI*, 5(4), pp.2538-2567.
- [16]. Jiang, P., Niu, W., Wang, Q., Yuan, R. and Chen, K., 2024. Understanding users' acceptance of artificial intelligence applications: A literature review. *Behavioral Sciences*, 14(8), p.671.
- [17]. Yu, L. and Li, Y., 2022. Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. *Behavioral Sciences*, 12(5), p.127.
- [18]. Gurram, A. (2025) 'Engineering Trustworthy Autonomous AI Agents: Architectural Patterns for Self-Planning, Tool-Oriented Reasoning, and Enterprise Task Completion', [Journal name], 5(6), pp. 2486–2494.
- [19]. Bensch, L., Bensch, O., Nilsson, T., Paradiso, J.A., Davidoff, S., Blanchard, L., Pataranutaporn, P., Weber, M., Kutz, J.N. and Flemisch, F., 2025. Human-in/on-the-Loop AI: Enabling Human Controllability and Decision-Making in Spaceflight. In *Handbook of Human-Centered Artificial Intelligence* (pp. 1-46). Singapore: Springer Nature Singapore.
- [20]. Bandi, A., Kongari, B., Naguru, R., Pasnoor, S. and Vilipala, S.V., 2025. The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9), p.404.
- [21]. Chamola, V., Hassija, V., Sulthana, A.R., Ghosh, D., Dhingra, D. and Sikdar, B., 2023. A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*, 11, pp.78994-79015.

- [22]. Arumilli, S.V.T. (2024) 'Evaluating Reliability Risks and Failure Propagation in Autonomous Multi-Step Agentic AI Frameworks', *International Journal of Communication Networks and Information Security*, 16(2), pp. 279–289.
- [23]. Alelyani, T., 2025. A validated framework for responsible AI in healthcare autonomous systems. *Scientific Reports*, 15(1), p.44432.
- [24]. Fernández Llorca, D., Hamon, R., Junklewitz, H., Grosse, K., Kunze, L., Seiniger, P., Swaim, R., Reed, N., Alahi, A., Gómez, E. and Sánchez, I., 2025. Testing autonomous vehicles and AI: perspectives and challenges from cybersecurity, transparency, robustness and fairness. *European Transport Research Review*, 17(1), p.38.
- [25]. Jedličková, A., 2024. Ethical considerations in risk management of autonomous and intelligent systems. *Ethics & Bioethics*, 14(1-2), pp.80-95.
- [26]. Malik, F., ul Haq, M.R. and Shahid, Z., 2025. From competence to calibration: modeling cognitive trust in human-AI collaborative systems. *Journal of Computer Science and Technology Studies*, 7(12), pp.173-182.
- [27]. Gundwal, M. and Sharma, N.K., 2025, October. Building Trust Through Responsible AI: Leadership Strategies for Business Intelligence Governance and Collaborative Innovation. In *International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 359-373). Cham: Springer Nature Switzerland.
- [28]. Arumilli, S.V.T. (2024) 'Human-Guided Agentic AI Workflows for Enterprise Operational Decision-Making', *International Journal of Intelligent Systems and Applications in Engineering*, 12(22s), pp. 2513–2523.
- [29]. Mylrea, M. and Robinson, N., 2023. Artificial Intelligence (AI) trust framework and maturity model: applying an entropy lens to improve security, privacy, and ethical AI. *Entropy*, 25(10), p.1429.
- [30]. Kovari, A., 2024. AI for decision support: Balancing accuracy, transparency, and trust across sectors. *Information*, 15(11), p.725.
- [31]. S. K. Nalluri, V. T. Bathini and V. K. B. Parasaram, "AI-Augmented Manufacturing Execution Systems for Precision Healthcare: A Scalable Framework for Smart Genomic and Proteomic Production," *2026 IEEE 2nd International Conference on Secure IoT, Assured and Trusted Computing (SATC)*, Houston, TX, USA, 2026, pp. 1-6, doi: 10.1109/SATC69565.2026.11542384.