The Human Side of Service Engineering, Vol. 221, 2026, 235–244
https://doi.org/10.54941/ahfe1007728

The Dynamics of Trust in AI Systems: Human–Machine Collaboration Within the MLS Exploitation Process

Satoshi Okuda and Naoshi Uchihira

School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

ABSTRACT

This study examines how trust in human–machine collaboration evolves within the exploitation process of machine learning systems (MLS), spanning from non-ML environments to fully autonomous operation. Building on prior research on the exploitation process of machine learning systems, the study positions trust as a foundational mechanism enabling sustainable human-machine collaboration. Drawing on the trust framework of Siau and Wang (2018), five trust-related dimensions—usability and reliability, collaboration and communication, sociability and bonding, security and privacy protection, and interpretability—were operationalized as survey items measured on a five-point Likert scale. Differences across exploitation phases were examined using ANOVA, and principal component analysis (PCA) was conducted to identify latent user archetypes underlying trust orientations. The results suggest distinct trust configurations across phases of the MLS exploitation process: Visualization, Human-centered ML Assistance, ML-centered Human Assistance, and Autonomy. Early phases emphasize collaboration and communication, highlighting the role of relational trust in initial adoption. As system autonomy increases, interpretability and privacy-related concerns gain relative importance, while in the autonomy phase, security and privacy protection tend to become more salient trust requirements. Interpretability gradually declines as MLS becomes operationally embedded, consistent with a shift from transparency-based trust to stability-based trust. Furthermore, the analysis identifies three latent user archetypes shaping trust orientations in human–machine collaboration: the General Users, the AI Practitioners, and the AI Translators. These findings clarify process-dependent trust dynamics and user-level differentiation within MLS exploitation, offering practical guidance for designing trustworthy AI systems aligned with system maturity and organizational roles.

Keywords: Human-machine collaboration, Software engineering for machine learning systems, Project management, Technology and innovation management

INTRODUCTION

Machine learning systems are increasingly embedded in software products and business processes in terms of digital transformation (Okuda and Uchihira, 2023) across many industries. The development of machine learning systems differs from the development of conventional IT systems, particularly the

Received February 21, 2026; Revised April 13, 2026; Accepted April 28, 2026; Available online July 20, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

common waterfall software development model (Balaji and Murugaiyan, 2012), and many new problems have emerged. Major technology companies such as Microsoft and Google not only provide data infrastructures but also propose development workflows for machine learning systems. For example, Amershi et al. (2019) observed the development process of machine learning systems at Microsoft and identified the nine stages of the machine learning workflow and challenges unique to AI engineering, such as data management and model customization. These workflows define steps for building machine learning systems; however, they do not explicitly describe how such systems are incrementally updated and refined through real-world use, nor how human–machine collaboration evolves during this process. Okuda and Uchihira (2024) proposed an exploitation process model that explains how MLS mature through real-world use and how human roles co-evolve across phases. The purpose of this research is to clarify the human-machine collaboration factors in each phase of the exploitation process and describe the actor characteristics. The formal research questions examined in this study are as follows.

RQ1: What trust-related dimensions shape human–machine collaboration within the exploitation process of machine learning systems?
RQ2: What actor characteristics emerge within the exploitation process of machine learning systems?

LITERATURE REVIEW

Research on data-driven system development has established structured process models such as KDD and CRISP-DM (Fayyad et al., 1996; Wirth and Hipp, 2000), later extended to machine-learning–specific workflows that include data labeling and continuous monitoring (Amershi et al., 2019). While these models provide development guidance, they do not explain how systems are incrementally exploited in organizational settings. Engineering studies highlight the technical vulnerabilities of machine learning systems. Technical debt and design uncertainty have been documented through anti-patterns (Sculley et al., 2015) and mitigated through design patterns (Washizaki et al., 2022). Yet these contributions treat system development as largely static and do not characterize how system use evolves or how human roles transform once deployment begins.

In industry, machine learning has demonstrated business value across healthcare, manufacturing, and services (Jordan and Mitchell, 2015; Agrawal et al., 2018). Success increasingly depends on collaboration among diverse actors and organizational adaptation (Nahar et al., 2022), indicating that outcomes are shaped not only by model accuracy but by how humans interpret, supervise, and integrate models into decision making.

Human–machine collaboration research has identified that the relationship between people and AI systems is dynamic. Studies describe hybrid intelligence formation (Pan, 2016), reciprocal adaptation supported by explainability (Bosch and Bronkhorst, 2018), hybrid labor between human-only and machine-only tasks (Daugherty and Wilson, 2018), and staged transitions

from AI assistants to teammates (Babic et al., 2020). Although guidelines for interaction exist (Amershi et al., 2019; Yang et al., 2020), they do not formalize how collaboration changes over time through system exploitation. Addressing this gap, Okuda and Uchihira (2024) propose a four-phase Exploitation Process for Machine Learning Systems (Figure 1) that explicitly links system maturity with evolving human roles. The model identifies the following stepwise modes of work:

1. Visualization – Data is collected, clarified, and shared to support hypothesis formation and organizational alignment.
2. Human-centered ML Assistance – MLS performs prediction and detection, but humans retain interpretive control and judgment.
3. ML-centered Human Assistance – MLS becomes the primary decision engine, while humans intervene mainly for exception handling or trust-building.
4. Autonomy – Workflows become highly standardized and irregular handling approaches zero, based on validated knowledge from prior phases.

This paper adopts the Exploitation Process for Machine Learning Systems (Okuda and Uchihira, 2024) as its central theoretical framework. By empirically examining how human–machine collaboration unfolds across the four phases—and how organizations shift from visualization to autonomy—this study integrates previously fragmented research areas in process modeling, engineering patterns, and human–AI collaboration. The findings offer a unified perspective on the sustainable exploitation of machine learning systems in real business contexts.

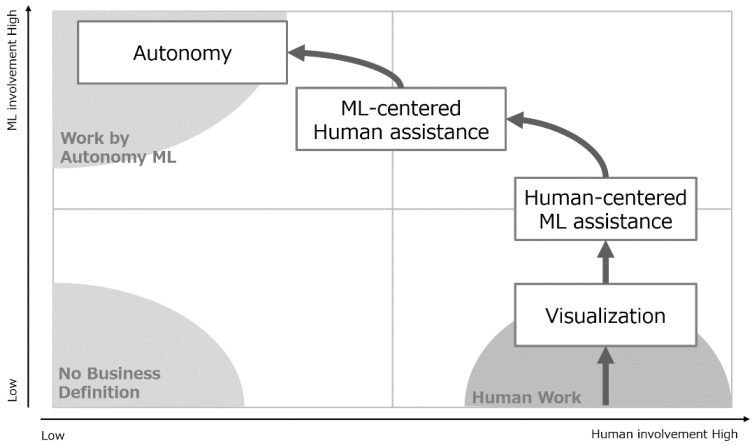


Figure 1: Exploitation process for machine learning systems and human and machine modes (Okuda and Uchihira, 2024).

RESEARCH METHODS

Data were collected through a questionnaire survey administered to 200 participants, from which 171 valid responses were obtained. The survey was conducted using a paid online survey service. As part of a preliminary screening, respondents were asked whether they were in a position responsible for promoting digital transformation. Only those who answered "yes" were included in the main survey. Siau and Wang (2018) identified foundational elements required for building trust in AI systems. Drawing on their framework and aligning it with the exploitation process of machine learning systems, this study operationalized trust in terms of five performance-related characteristics: usability and reliability, collaboration and communication, sociability and bonding, security and privacy protection, and interpretability. ANOVA and principal component analysis (PCA) were performed using SPSS Ver. 29.0.2.0, and multivariate analysis was conducted using Microsoft Excel Ver. 2403.

PCA with varimax rotation and Kaiser normalization was conducted to explore the underlying structure of trust dimensions. Sampling adequacy was confirmed (KMO = .840; Bartlett's test of sphericity: $\chi^2(10) = 264.125$, $p < .001$), indicating that the correlation matrix was suitable for component extraction. A one-way analysis of variance (ANOVA) was conducted to test for differences across the four phases of the exploitation process model for machine learning systems. Given the practitioner-based and exploratory nature of this study, both the conventional 5% significance level ($\alpha = .05$) and a 10% threshold ($\alpha = .10$) were adopted. Under these criteria, three of the five dependent variables demonstrated marginal significance at the 10% level: Usability & Reliability ($p = .089$), Collaboration & Communication ($p = .090$), and Security & Privacy Protection ($p = .067$). By contrast, Sociability and Bonding ($p = .135$) and Interpretability ($p = .172$) did not show differences across exploitation phases. These results suggest that perceptions related to trust, coordinated work, and responsible deployment are more sensitive to phase progression than sociability or interpretability alone. Accordingly, marginal effects are interpreted as indicative trends within an exploratory research context.

The analysis was conducted in two stages. In the first stage, changes in human–machine collaboration factors across different phases of the machine learning systems exploitation process were examined. In the second stage, PCA was applied to the Human-machine collaboration factors (HMC factors) to explore characteristic user profiles associated with human–machine collaboration. PCA of the five HMC variables retained three components. Together they explained 84.4% of variance, met the eigenvalue-greater-than-one rule, and yielded interpretable loading patterns. The third component also reflected a distinct bridging segment absent under a two-component solution.

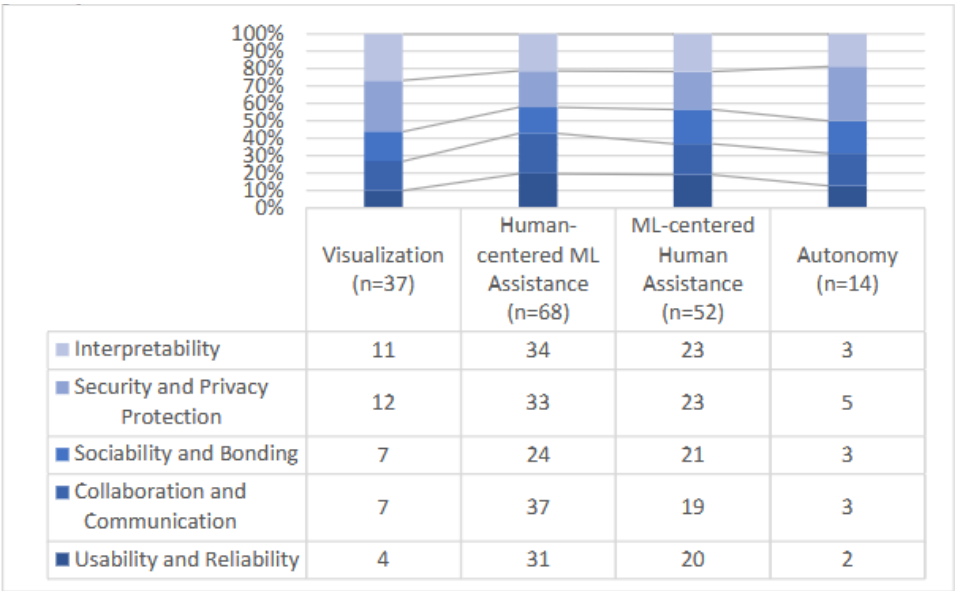
HUMAN–MACHINE COLLABORATION TRUST FORMATION FACTORS

This section describes changes in HMC factors across phases within the exploitation process of machine learning systems. Figure 2 presents a 100% stacked bar chart that summarizes HMC factors across phases of machine

learning systems exploitation process. For each phase, the number of respondents who selected values of 4 or 5 on a five-point Likert scale for each HMC factor was aggregated. The HMC factors consist of the following five dimensions:

- Usability and reliability: AI technologies are convenient and reliable
- Collaboration and communication: Ability to collaborate and communicate with humans
- Sociability and bonding: Ability to exhibit sociability and form bonds
- Security and privacy protection: Security and privacy are protected
- Interpretability: Ability to explain how AI outputs are derived

In the visualization phase, machine learning systems are not yet applied; therefore, responses are based on respondents' expectations rather than actual experience. In this phase, high values are observed for security and privacy protection and interpretability. Among users in the human-centered ML assistance phase, collaboration and communication shows the highest values, followed by interpretability. In the ML-centered human assistance phase, interpretability and security and privacy protection exhibit similarly high values. In the autonomous phase, security and privacy protection shows the highest values. Overall, interpretability tends to decline as the exploitation process advances. This trend can be interpreted as reflecting users' increased understanding of the behavior of machine learning systems through continued operation. In phases where either humans or machine learning systems provide assistance, security and privacy protection shows relatively low values, whereas usability and reliability remains high. This suggests that, during the exploitation process of machine learning systems, usability and reliability are prioritized in human–machine collaboration, as it is essential for stakeholders to recognize that the system can be effectively utilized. During these assistance phases, security and privacy protection may be perceived as less critical, as a certain level of control is maintained by human oversight. In the human-centered ML assistance phase, where machine learning systems are first introduced, collaboration and communication is particularly emphasized. This phase represents the initial step in establishing a cooperative relationship between humans and machines, during which trust is fostered while actively involving users. Finally, in the autonomous phase, machine learning systems operate independently with minimal human intervention, resulting in heightened requirements for security and privacy protection. It should be noted that the number of respondents classified in the Autonomy phase was relatively small (n = 14), reflecting the still-limited diffusion of fully autonomous MLS in organizational settings. Therefore, findings related to this phase should be interpreted as exploratory indications rather than definitive patterns. Although empirical representation of the Autonomy phase remains limited, its inclusion is theoretically essential, as it represents the culmination of the exploitation process where trust shifts from relational and interpretive dimensions toward institutionalized safeguards such as security and privacy.



	Visualization (n=37)	Human-centered ML Assistance (n=68)	ML-centered Human Assistance (n=52)	Autonomy (n=14)
Interpretability	11	34	23	3
Security and Privacy Protection	12	33	23	5
Sociability and Bonding	7	24	21	3
Collaboration and Communication	7	37	19	3
Usability and Reliability	4	31	20	2

Figure 2: Trust-related HMC factors across phases of the MLS exploitation process.

The second stage of the analysis focuses on identifying user characteristics in human–machine collaboration. Human–machine collaboration factors were subjected to PCA. The results are presented in Table 1. Components up to the third component, which together account for a cumulative contribution rate of 84.4%, were selected for analysis. Based on the two variables with the highest loadings for each component, interpretations were derived for Components 1 through 3.

Table 1: PCA results for trust dimensions in human–machine collaboration.

	Factor		
	1	2	3
Usability and reliability: AI technologies are convenient and reliable	0.786	–0.093	–0.518 #3
Collaboration and communication: Ability to collaborate and communicate with humans	0.829 #1	–0.041	–0.179
Sociability and bonding: Ability to exhibit sociability and form bonds	0.723	0.658 #2	0.092
Security and privacy protection: Security and privacy are protected	0.795	–0.393 #2	0.228
Interpretability: Ability to explain how AI outputs are derived	0.810 #1	–0.070	0.380 #3

Component 1: General Users
Characteristics: Overall high expectations
Key factors: Collaboration and communication (0.829) and Interpretability (0.810) #1

The first component represents users who expect effective communication with AI while placing strong importance on interpretability throughout the interaction process. These users value collaboration with machine learning

systems and simultaneously require the ability to interpret system outputs. In addition, relatively high values are observed for security and privacy protection as well as usability and reliability, indicating generally elevated expectations across multiple dimensions. The model suggests a pathway in which General Users may develop into AI Practitioners, supported by AI Translators.

Component 2: AI practitioners
Characteristics: AI as a work partner
Key factors: Sociability and bonding (0.658) and Security and privacy protection (–0.393) #2

The second component reflects a tendency to expect machine learning systems to exhibit social behaviors and to form bonds similar to those of humans. This component emphasizes the importance of AI not merely as a tool but as a collaborative team member capable of mutual communication, embodying the concept of human–machine teaming. In contrast, concern for security and privacy protection is relatively low within this group.

Component 3: AI Translators
Characteristics: Interest in deep logical interpretation
Key factors: Usability and reliability (–0.518) and Interpretability (0.380) #3

The third component indicates a preference for interpretability even at the expense of usability and reliability. This component can be interpreted as reflecting the mindset of "AI translators" who are highly interested in deeply understanding the underlying logic of machine learning systems. These users prioritize the interpretation of machine learning outputs and consider reduced usability and reliability acceptable in exchange for greater transparency and explainability.

The author identified that user groups in human-machine collaboration can be classified into three characteristic roles (Figure 3). The diagram illustrates a structure in which general users evolve into AI practitioners, with AI translators mediating between the two groups. This relationship provides an important perspective for understanding the maturation process of human–machine collaboration.

Figure 3: Actor relationship diagram in human–machine collaboration.

These results reveal key factors for understanding user expectations toward machine learning systems and for establishing trust in the systems. By combining these components with phase-specific human–machine collaboration factors, the analysis provides insights into the elements necessary to facilitate effective human–machine collaboration. A summary of the PCA is presented in Table 2.

Table 2: User profiles of human–machine collaboration.

Component	Name	Key Factors	Descriptions
1	General Users	Collaboration and communication (0.829), Interpretability (0.810)	Effective communication with AI and interpretation of its outputs are considered important. Expectations for security and privacy protection, as well as usability and reliability, are generally high.
2	AI Practitioners	Sociability and bonding (0.658), Security and privacy protection (–0.393)	Emphasis is placed on AI functioning as a member of the team and engaging in mutual communication. Concern for security and privacy protection is relatively low.
3	AI Translators	Usability and reliability (-0.518), Interpretability (0.380)	Interpretability is prioritized, reflecting strong interest in deeply understanding AI logic. Reduced usability and reliability are considered acceptable in exchange for greater interpretability.

DISCUSSION AND CONCLUSION

This study conducted a PCA to clarify the psychological structures underlying trust and adoption of machine learning systems in organizational settings. The findings revealed three latent factors, which can be conceptualized as distinct user roles: General Users, AI Practitioners, and AI Translators. These factors correspond to differing orientations toward MLS—ranging from reliance on the technology as a supporting tool, to active utilization in daily tasks, and finally to mediating and explaining system behavior to others. This result suggests that AI adoption in organizations does not proceed uniformly; rather, it develops through differentiated user positions that coexist within the same environment.

The three user roles identified through this research provide important implications for theory and practice. First, the concept of AI adoption must be reframed as a socio-technical phenomenon involving complementary human roles, rather than a monolithic user group. General Users represent the broadest base and employ machine learning systems primarily as an assistive tool. AI Practitioners, in contrast, integrate machine learning systems more actively into their work processes and seek to enhance collaboration, efficiency, and decision making. Finally, AI Translators occupy a unique mediating position—interpreting model outputs, articulating underlying rationales, addressing uncertainty, and reconciling machine-generated insights with human expectations. Second, the three-layer structure aligns strongly with existing theoretical frameworks, providing academic validation of our interpretation. Babic et al. (2020) conceptualize the progression of AI acceptance from Assistant to Monitor, Coach, and ultimately Teammate. Under this lens, General Users align with the assistant phase, AI Practitioners correspond to mid-level coaching roles, and AI Translators serve key functions associated with monitoring and coaching in organizational

contexts. Similarly, the "Missing Middle" proposed by Daugherty and Wilson (2018)—the hybrid space where humans and AI jointly contribute value—maps directly onto the Translator role identified in this study. Translators are positioned at the boundary between technical and human spheres, ensuring that AI is neither blindly relied upon nor prematurely rejected. Furthermore, the user role differentiation observed here provides empirical grounding for Human–AI Interaction frameworks (Amershi et al., 2019; Yang et al., 2020), which argue that effective AI adoption depends on the co-evolution of human capabilities, organizational structures, and system maturity. Third, the results can be connected to the Exploitation Process for Machine Learning Systems proposed by Okuda and Uchihira (2024). Their model identifies four phases—Visualization, Human-centered ML assistance, ML-centered human assistance, and Autonomy—as markers of progressive adoption. The roles identified in this study can be understood as enablers of transitions between these phases. General Users provide foundational exposure and contribute to the Visualization phase by generating organizational awareness and trust. AI Practitioners facilitate movement toward Assistance by integrating machine learning systems outputs into operational workflows and optimizing decision making. AI Translators play a pivotal role in enabling delegation, by establishing explainability, and building consensus across diverse stakeholders. In this sense, the differentiation of user roles is not merely descriptive but functionally necessary for sustained machine learning systems exploitation. Technical maturity alone cannot move organizations beyond early adoption without corresponding advancement in user competencies.

Finally, the findings carry actionable implications for human resource development, organizational structures, and AI deployment strategies. Traditional efforts in AI introduction have often concentrated on technical training or expanding user familiarity. However, this study suggests that successful adoption hinges critically on supporting and cultivating AI translator roles. In environments lacking such mediators, machine learning systems may remain underutilized or may generate mistrust, even when technically sound. Conversely, organizations with well-developed mediator capabilities may progress rapidly through exploitation phases and unlock higher-order benefits from machine learning systems integration. Despite these contributions, several limitations should be acknowledged. This study is based on cross-sectional self-reported survey data and does not track longitudinal transitions across exploitation phases. The relatively small number of respondents in the Autonomy phase limits statistical power and suggests cautious interpretation of those findings. In addition, the exploratory use of PCA and the focus on Japanese practitioners may constrain generalizability and call for confirmatory and cross-cultural validation in future research. Ultimately, the findings suggest that sustainable machine learning systems exploitation requires the deliberate design of social systems that amplify the strengths of each user type and harness their collective contribution toward organizational transformation.

REFERENCES

Agrawal, P., and Narain, R., (2018). Digital supply chain management: An Overview. In IOP conference series: materials science and engineering. IOP Publishing. 455(1): 012074.

Amershi, S. et al. (2019). Software Engineering for Machine Learning: A Case Study. Proc. - 2019 IEEE/ACM 41st Int. Conf. Softw. Eng. Pract. ICSE-SEIP, 2019: 291–300.

Babic, B. et al. (2020). A better way to onboard AI. Harvard Business Review, 2020 July– August.

Balaji, S., and Murugaiyan, 2012, "Waterfall vs. V-Model vs. Agile: A comparative study on SDLC", International Journal of Information Technology and Business Management, 2(1):26–30.

Bosch, K. V. D. and Bronkhorst, A. (2018). Human-AI cooperation to benefit military decision making. STO June: 1–12.

Daugherty, P. R. and Wilson, H. J. (2018). Human+ machine: Reimagining work in the age of AI. Harvard Business Press.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Commun. ACM, 39–11: 27–34.

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245): 255–260.

Nahar, N., Zhou, S., Lewis, G., and Kästner, C. (2022). Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In Proceedings of the 44th International Conference on Software Engineering, 413–425.

Okuda, S., Uchihira, N. (2023). Digital Transformation Classification Types and Evolution Process for Established Companies. The Human Side of Service Engineering. AHFE (2023) International Conference. AHFE Open Access, vol. 108. AHFE International, USA.

Okuda, S., & Uchihira, N. (2024). Exploitation Process for Machine Learning Systems. International Journal of Innovation and Technology Management, 21(06), 2450048.

Pan, Y. (2016). Heading toward artificial intelligence 2.0. Engineering 2, 2016: 409–413.

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. Adv. Neural Inf. Process. Syst, 2015-January: 2503–2511.

Siau, K., & Wang, W., 2018, "Building trust in artificial intelligence, machine learning, and robotics", Cutter business technology journal, 31(2):47–53.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Washizaki, H. et al. (2022). Software-Engineering Design Patterns for Machine Learning Applications. In Computer, 55(3): 30–39.

Yang, Q., Steinfeld, A., Rosé, C. and Zimmerman, J.. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. Conf. Hum. Factors Comput. Syst. - Proc: 1–13.