Calibrating Worker Trust in AI: A Multi-Dimensional Scoping Review on Professional Delegation

TRUSTWORTHY TEAM | Deliverable 1

April 25, 2026

Abstract

This scoping review synthesizes evidence from 47 sources to define the Multi-Dimensional Trust Framework (MDTF). By mapping technical reliability against legal accountability, we identify the thresholds required for "calibrated trust." Results indicate that while LLMs can achieve near-perfect technical accuracy in data extraction, delegation remains bounded by the 'Responsibility Constraint' in professional adjudication

1. INTRODUCTION

It is widely acknowledged that the creation and introduction of incredible and powerful tools such as ChatGPT, Claude, Gemini, etcetera – in a much more expanded way, the introduction of Artificial Intelligence – is rapidly and radically changing our lives, re-drawing them for a future in which the human-AI union will be our everyday life. This reality might be much closer than we think. The latest technological advancements seem progressively faster each year, making it our task to understand and change our behaviour towards these LLMs to model them by our interests. For this aspect AI might be one of the biggest revolution and most intriguing challenge of our time, because it forces us to consider its use in various contexts, from working environments to private places. The emerging urge to grow a conscious, trained and tech-wise society – capable of using these instruments in a sustainable and efficient way – raises questions and concerns about its reliability, its ethic, its accountability. To understand deeply this necessity to model these instruments for our needs and requirements we decided to analyse profoundly how AI affects our lives and how AI re-shape our tasks; we decided to scan the involvements of direct and indirect stakeholders that are influenced by continuous use of AI; we decided to discuss about moral and ethical values, concentrating on those which enhance or worsen reliance and perception of trust towards AI. This topics and issues that we are facing nowadays can be drastically stripped down to a simple, complete and central question: Can we consider AI tools trustworthy? In this work we try to collect and summarize some paper that are linked to trust in AI algorithms or LLMs. By various research on engines such as google scholar, scopus or others we found lots of documents that address this matter; after careful analyses we collected some information and we wrote this scoping review. We are opening this work with a paragraph that explain which factors and criteria drove us to include or discard a paper, showing our entire route that lead us the creation of the general framework (which we will profoundly develop). After this we discuss about applying these values and themes collected in the framework to a specific employment. In our case, we will talk about a bankruptcy accountant and how the introduction in his workplace of AI tools can affect his job. Finally, we will reason on possible consequences and lacks of this paper.

2. METHOD

We conducted a scoping review to map and synthesize the existing evidence regarding worker trust in AI systems. The review follows a structured 8-step framework to ensure methodological transparency and the development of a generalizable framework.

2.1 Research Question

The review is centered on a Type 3 research question: "What factors determine whether a worker's trust in an AI system is well-calibrated for a given task?". This question aims to identify the specific dimensions that influence how trustors (workers) perceive and rely on AI trustees during task augmentation or automation.

2.2 Search Strategy and Eligibility

We performed a systematic search across primary databases including **Google Scholar** and the **ACM Digital Library**, supplemented by **Manual Search (Grey Literature)** to capture practice-based industry reports.

Search strings were constructed by combining terms for the phenomenon (e.g., "AI", "algorithm"), the context (e.g., "worker", "occupation"), and the outcome (e.g., "trust", "calibration"). To maintain relevance to the current technological landscape, we limited results to those published between 2015 and 2026.

Sources were screened using a three-stage funnel: title screening, abstract review, and full-text analysis. The full screening workflow and exclusion counts are summarized in Figure 1.

A source was included if it:

1. Addressed factors relevant to trust calibration.

2. Focused on AI in occupational settings.

3. Was peer-reviewed or a credible practice document.

Our final selection consists of 47 sources, meeting the target requirements for diversity:

- 21 empirical studies providing data on real workers.
- 21 theoretical frameworks proposing models of trust.
- 5 practice-based reports from professional bodies.

2.3 Data Extraction and Synthesis

For each included source, we extracted "factors" defined as any variable or condition identified by the literature as significant for trust. We employed an iterative open coding process to categorize these factors. Codes were compared across the team and grouped into the five central themes presented in the framework: **Technical Robustness**, **Transparency**, **Human Agency**, **Organizational Governance**, and **Psychological Calibration**. A detailed record of search strings, screening logs, and the full coding scheme is provided in the Appendix for verification.

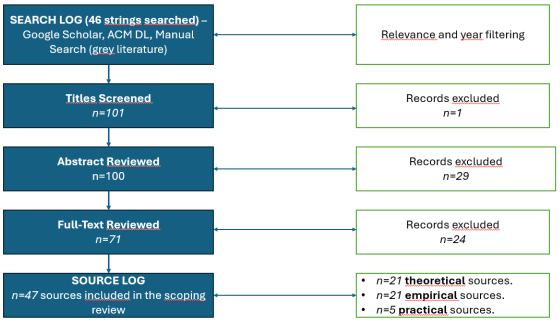

Figure 1: PRISMA-inspired screening process used in this scoping review, including exclusion steps and final source composition.

3. FRAMEWORK

The following framework synthesizes the factors identified in the scoping review to define the multi-dimensional conditions under which a worker's trust in an AI system is warranted and effectively calibrated. In accordance with the "Generality Rule", the framework identifies universal socio-technical drivers that apply across diverse occupational roles. It organizes these drivers into five hierarchical layers: technical robustness, transparency mechanisms, agency structures, psychological filters, and systemic governance.

Layer 1: Technical Robustness and Intrinsic System Integrity

Technical robustness represents the system's objective competence and serves as the primary antecedent for any trust-building process, as shown in [12]

> "Users cognitively evaluate AI capabilities through dimensions, such as reliability, accuracy, and functional efficiency [...] positioning competence as crucial for trust formation."

Trust in this layer is not a static state but develops through an iterative feedback loop of performance and consistency.

- **High-Fidelity Accuracy and Reliability:** Precision and efficiency in completing specific occupational duties are primary boosters of trust. Workers rely on the system's ability to consistently produce high-quality outputs, as any perceived inaccuracy in task completion leads to immediate "mistrust" in professional environments.
- **Temporal and Cross-Contextual Stability:** For trust to be sustainable, the AI must exhibit stability over time and maintain fairness across different cultural or demographic datasets. Systems that offer upgraded, moral-ethical, and unbiased responses have been shown to lower general distrust by proving their reliability in varying conditions.
- **Intrinsic Competence and Benevolence:** Beyond raw performance, trust is influenced by the perceived "integrity" of the system—the alignment of the system's goals with the worker's genuine interests and professional standards.
- **Four Pillars of Architecture:** Responsible AI adoption is built upon the integrity of four critical pillars: Data quality, Algorithm design, Human oversight, and Model architecture, all of which must function as strategic enablers.

Layer 2: The Cognitive Bridge: Transparency and Justification

> "The opaqueness of complex AI systems has led to widespread concerns [...] there is a demand among users for the right to know the intention, business model, and technological mechanism of AI products." [20]

Transparency acts as the mechanism through which "black-box" algorithmic logic is converted into verifiable professional judgment. It allows workers to transform AI results into defensible decisions.

- **Explainable AI (XAI) and Interpretability:** The integration of frameworks such as SHAP or LIME provides "local explanations" that allow workers to justify specific AI decisions to external stakeholders. This prevents "blind trust" and ensures that reliance is based on data that the professional can justify.

Table 1: Detailed Mapping of Trust Calibration Factors and Evidence Sources

Thematic Pillar	Core Calibration Factors	Key Sources
Technical Robustness	Accuracy, Reliability, Temporal Stability, Model Integrity, Competence, Benevolence.	[12, 31, 7, 5, 13, 8, 27, 32, 20, 14, 2, 45]
Transparency	XAI (SHAP/LIME), Local Justification, Logic Disclosure, Data Provenance, Vendor Disclosure.	[3, 8, 27, 46, 20, 47, 9, 25, 30, 24, 6, 15, 29]
Human Agency	IDA Design, Human-in-the-Loop, Verification Protocols, Accountability, Autonomy.	[35, 38, 4, 34, 8, 27, 28, 10, 26, 39, 40, 19, 9, 14, 24, 2, 42, 29]
Psychological Filter	Automation Bias, Tech Anxiety, Training, Experience, Professional Self-Confidence.	[17, 21, 33, 43, 23, 8, 22, 36, 32, 44, 26, 39, 19, 9, 6, 42, 29]
Systemic Governance	Regulation, Ethical Stewardship, Leadership Support, Supply Chain Transparency.	[37, 16, 11, 1, 8, 41, 19, 18]

- **Communication Clarity and Logic Disclosure:** The effectiveness of a system depends on the visual clarity of its explanations and the explicit disclosure of its underlying logic. High-quality communication from the system directly impacts how experts perceive its credibility and professional integrity.

- **Data Provenance and Rights:** Trust is increasingly grounded in the transparency of data sources and the verification of safety protocols. Any delegation of professional tasks must be considered responsible and defensible through technical safeguards.

- **Algorithmic Accountability:** Clear documentation regarding vendor transparency and the mitigation of inherent biases within the architecture is a fundamental principle for a trustworthy AI design.

Layer 3: The Control Layer: Human Agency and Decision Support

A critical requirement for calibrated trust is the preservation of worker autonomy and the prevention of deskilling. Trust is only warranted when the human operator remains the final authority, as shown in [4].

> "Users should be able to utilize their knowledge to improve the outcomes in situations where AI models may have limitations [...] expert knowledge guided the trust calibration, helping the users to decide when to (or not) trust the systems' suggestion."

- **Intelligent Decision Assistance (IDA) vs. Automation:** System design should favor "Intelligent Decision Assistance" over full automation. By providing explainable support while withholding direct recommendations, the system ensures workers remain actively involved in judgment and learning.

- **The "Human-in-the-Loop" Mandate:** Maintaining human control and final decision-making power in AI-augmented workplaces is essential to reduce distrust and ensure ethical validation.

- **Verification Complexity and Cognitive Workload:** Trust calibration fails if the cognitive workload required to verify AI suggestions is so high that it encourages "automation bias" or over-dependence.

- **Collaborative Design and expertise-driven double-checking:** Trust is fostered when workers can apply their specific professional expertise to "double-check" system outputs. Collaborative design with developers ensures that the tool supports the worker's ability to challenge AI results.

Layer 4: The Internal Filter: Psychological and Cultural Factors

> "This suggests that expertise and experience [...] may help reduce the potential impact of automation bias [...] domain skills and competencies safeguard clinicians from trusting false AI-enabled recommendations." [17]

Individual and cultural factors act as filters through which technical information is processed, often overriding objective performance facts.

- **Automation Bias and Over-Calibration:** Experts frequently succumb to incorrect AI prompts, highlighting a significant challenge in achieving a well-calibrated trust relationship in high-stakes environments.

- **Technology Anxiety and Emotional Resistance:** Emotional barriers, such as anxiety regarding AI's impact on job security or the complexity of the interface, can override rational assessments of the tool's usefulness.

- **Professional Self-Confidence and Skepticism:** Professional skepticism is an irreplaceable "human element" that must balance AI adoption. Specific AI training is often more influential than general professional seniority in determining how workers perceive AI utility.

- **Trust as a Cultural Anchor:** Workplace culture characterized by openness, fairness, and a "learning culture" is a prerequisite for successful human-AI

teamwork and the acceptance of AI teammates. Trust propensity, or the innate willingness to trust technology, sets the baseline for these interactions.

Layer 5: Systemic Shield: Governance and Ethical Stewardship

Trust is not merely a relationship between a worker and a tool; it is embedded within broader organizational and regulatory contexts that define the practical conditions for delegation, as it is shown in [18]:

> "Multiple general non-binding guidelines exist, either as global frameworks or guidelines [...] However, concerns are raised on how such guidelines will be implemented and whether they are enough."

- **Regulatory and Legal Mistrust:** The lack of global regulation creates uncertainty and "legal mistrust" among professionals, creating a barrier to technology adoption.
- **Ethical Stewardship and Digital Integrity:** Trust is a dynamic quality maintained through ethical decision-making. Professionals must act as ethical stewards, ensuring that any delegation to AI is grounded in integrity and the public interest.
- **Strategic Leadership and Resource Support:** Successful AI integration requires leadership that provides the necessary resources for specialized training and technical literacy.
- **Supply Chain Trust Dynamics:** Trust must be managed across the entire chain of people and organizations involved in building and using AI services, particularly when dealing with Large Language Models (LLMs).

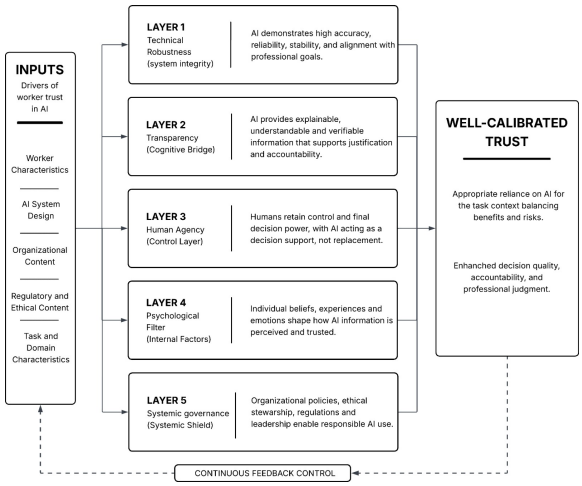

Figure 2: Framework Diagram

4. WORKED USE CASE

To demonstrate the practical utility of our general framework, we apply it to the specific professional context of a **Bankruptcy Accountant** tasked with identifying fraudulent asset transfers. This scenario is not a historical report, but a projection of the consequences derived from the application of the Multi-Dimensional Trust Framework (MDTF).

The Profession and the Task

The Bankruptcy Accountant operates under intense judicial oversight, acting as a technical consultant for the court to recover assets from insolvent estates. The core challenge involves scanning vast, complex financial datasets to detect "anomalies" specifically, assets transferred to third parties shortly before a bankruptcy filing to defraud creditors.

The AI System and its Application

The professional utilizes an AI-enabled **Forensic Auditing System** designed for automated transaction screening. The system applies Machine Learning (ML) models to historical transaction patterns to flag potential "fraudulent preferences." Unlike a simple rule-based filter, this AI identifies non-obvious correlations across thousands of data points, performing a role of **Intelligent Decision Assistance (IDA)** rather than autonomous decision-making.

Application of the Calibration Framework

- **Systemic Governance & Lifecycle Management:** Trust is not established at a single point in time but managed as a continuous process. The professional is involved in the system's calibration from the early stages. As stated by **Stefanova-Stoyanova & Danov (2025)** [40]:

  > "The AI management system should ensure that the human-machine teaming is integrated throughout the entire AI lifecycle, starting from the planning phase and continuing through to operation and monitoring."

  - **Imagined Consequence:** The accountant actively participates in the AI's planning phase to align algorithmic "red flags" with specific judicial precedents, ensuring that "Teaming" begins at the design level.

- **Technical Robustness and Temporal Stability:** Trust is initially grounded in the AI's computational consistency, which mitigates human error in large-scale data ingestion. However, the professional must verify the model's temporal stability: if financial regulations or market norms change, the system must be re-validated to ensure that flags are not generated by outdated criteria.

- **Transparency and Justification (XAI):** Since the accountant must testify before a judge, the system must provide local explanations via XAI frameworks like SHAP or LIME. This allows the professional to articulate the specific rationale behind a flagged transaction, transforming a "black box" output into legally defensible evidence.

- **Psychological Filter and Critical Engagement:** To counter **automation bias** (the risk of blindly trusting the machine), the accountant must maintain high situational awareness. According to **Frontiers in Artificial Intelligence (2024)** [39]:

  > "Effective teaming requires that the human agent maintains situational awareness and a level of critical engagement that prevents the uncritical acceptance of AI-generated outputs."

  - **Imagined Consequence:** The accountant adopts a "skeptical oversight" protocol, manually double-checking a random sample of transactions that the AI classified as "clean" (false negative check). This ensures the expert remains a cognitively engaged supervisor rather than a passive observer.

- **Human Agency and Ethical Stewardship:** The system is designed to withhold final legal interpretations. The AI "finds" the data, but the accountant remains the **Human-in-the-Loop**. The professional acts as an ethical steward: trust is warranted only if the AI operates within a clear regime of accountability where the professional's expert judgment validates every algorithmic recommendation.

5. GAPS AND FUTURE WORK

The mapping of existing evidence reveals critical areas where knowledge remains fragmented, hindering the development of a truly universal trust framework.

Empirical and Geographical Limitations

Many of the empirical studies employed in this review offer geographically limited samples, mainly from the USA [12] [14] [28] [30], Spain [26], South Korea [12] [27], China [12], India [12], Germany [12] [19], [24], the UK [12] , Austria [24] and Switzerland [24]. These findings may not fully represent the trust dynamics in different regulatory or cultural environments. Future research must prioritize cross-cultural empirical validation to ensure the framework's global applicability.

Theoretical vs. Empirical Gap

While several robust frameworks have been proposed, such as SME-TEAM or the 3D trust framework, many lack rigorous empirical validation. Many "practice" sources provide ethical foundations but focus less on the technical "how-to" of calibrating trust for specific software interfaces. Research is needed to bridge this gap by testing theoretical delegation criteria in real-world professional settings.

Long-term Impact on Professional Judgment

There is a lack of longitudinal data regarding the long-term effects of AI-augmented work on the development of professional judgment. While Intelligent Decision Assistance (IDA) is theorized to prevent deskilling, empirical evidence on how continuous AI reliance affects the "human element" of skepticism over several years is missing. Future work should investigate whether AI-driven efficiency gains eventually erode the critical thinking skills required for high-stakes professional verification.

6. CONCLUSION

This scoping review synthesized evidence from 47 sources to construct the Multi-Dimensional Trust Framework (MDTF). The framework establishes that calibrated worker trust in AI is determined by five interconnected layers: technical robustness, transparency mechanisms, human agency structures, psychological filters, and systemic governance. The application of the MDTF to the Bankruptcy Accountant use case demonstrates that raw computational accuracy is insufficient for professional delegation. Instead, AI systems must be designed as Intelligent Decision Assistance (IDA) tools. To ensure effective and responsible task delegation, the system must support local explanations (XAI) , while the human operator must strictly maintain a "Human-in-the-Loop" position. Ultimately, trust is warranted only when the professional remains an engaged ethical steward, capable of converting algorithmic outputs into legally and technically defensible decisions.

7. APPENDIX

7.1 AI TOOL USE DISCLOSURE

In this project we used artificial intelligence, this section details the collaborative use of generative tools during the scoping review process.

Mandatory Statement of Verification: As required by the project guidelines, the team explicitly declares that:

- All cited sources were read in full by at least one team member.

- No source was cited based solely on an AI-generated summary or abstract.

Table 2: AI Tool Use Disclosure and Verification Log

Tool(s) Used	Specific Application	Verification and Human Oversight Performed
ChatGPT, Gemini, Claude	Brainstorming of search strings and initial thematic grouping of factors extracted from the Source Log.	The team manually cross-referenced every factor and code against the original Scoping Review Source Log to prevent hallucinations.
Gemini, Claude	Synthesis of general framework text and formatting of the Worked Use Case into ACM-compliant LaTeX code.	Every paragraph was reviewed for technical accuracy. All profession-specific details were isolated to the Use Case to maintain the "Generality Rule".
Gemini	Automated generation of LaTeX table structures and bibliography formatting.	Manual check of every Source Reference to ensure citations accurately reflect the provided evidence.

- The final interpretation of "well-calibrated trust" and all judicial implications in the Use Case remain the sole intellectual product of the authors.

7.2 ADDITIONAL MATERIAL (ZIP ARCHIVE)

The additional material for this deliverable is available at the following link:

`https://www.dropbox.com/scl/fi/gv2rtdt4snuzoeqxei7d2/Deliverable_1_TRUSTWORTHY.zip?rlkey=qq89zwk2ozvf63n3t88xt5r68&st=asnfe16m&dl=0`

The ZIP archive contains:

- sources used for the scoping review;
- a `readme.txt` file with structure and usage notes;
- an Excel file containing:
  - full search log,
  - screening process,
  - source log.
- LaTex source files.

7.3 TEAM MEMBERS

Team name: Trustworthy

Team Member	Email
Saverio Medici	s322898@studenti.polito.it
Artemisia Lollo	s335817@studenti.polito.it
Andrea Antico	s326625@studenti.polito.it
Bruno Germanis	s343237@studenti.polito.it
Eugenio Costella	s342382@studenti.polito.it
Matteo Fissore	s340798@studenti.polito.it
Matteo Vaccaneo	s343865@studenti.polito.it
Alessio Nicotra	s325235@studenti.polito.it
Lorenzo Chiaramello	s338840@studenti.polito.it
Alessandro Denis Ciobanu	s311248@studenti.polito.it

References

[1] ACCA. Professional accountants – the future: Ethics and trust in a digital age, 2025. Professional guidance document.

[2] Abhishek Adhikari. Human factors in designing trustworthy agentic ai for industrial systems. *none*, 2025.

[3] Rehan Akhtar, Yahaya Sanusi, Abbas Nuhuyau, Thinghi Zaw, Abiodun Okunola, Juma Alvi, Tariq Khan, and Sayem Hossain. Explainable ai (xai) in the audit room: Evaluating the interpretability of shap and lime frameworks for justifying machine learning-based fraud flags to human auditors, 2025. Preprint / working paper.

[4] Saara Ala-Luopa, Thomas Olsson, Kaisa Väänänen, Maria Hartikainen, and Jouko Makkonen. Trusting intelligent automation in expert work: Accounting practitioners' experiences and perceptions. *Computer Supported Cooperative Work (CSCW)*, 33(4):1343–1371, 2024.

[5] Abdellatif Aziki, Abdelaziz Ourrani, and Moulay Hachem Fadili. Trust or rust: The crucial role of trust in ai acceptance by professional accountants. *Procedia Computer Science*, 251:186–191, 2024.

[6] Tita A Bach, Jenny K Kristiansen, Aleksandar Babic, and Alon Jacovi. Unpacking human-ai interaction in safety-critical industries: A systematic literature review. *IEEE Access*, 12:106385–106414, 2024.

[7] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. Unpacking trust dynamics in the llm supply chain: An empirical exploration to foster trustworthy llm production & use [supplementary material], 2025. Supplementary material.

[8] Amir Behzadan and Armita Dabiri. Factors influencing human trust in intelligent built environment systems. *AI and Ethics*, 5(6):5841–5855, 2025.

[9] Bryan Cassady. Human expertise as the multiplier in ai-assisted decision making evidence on decision quality, error detection, and when ai narrows or widens

expertise gaps. *Error Detection, and When AI Narrows or Widens Expertise Gaps (January 06, 2026)*, 2026.

[10] Zichen Chen, Yunhao Luo, and Misha Sra. Engaging with ai: How interface design shapes human-ai collaboration in high-stakes decision-making. *arXiv preprint arXiv:2501.16627*, 2025.

[11] CPA. Ai solution due diligence guide for accounting firms. a practical evaluation framework for ai-enabled technology solutions, 2025. Professional guidance document.

[12] Qinpu Dang and Guiquan Li. Unveiling trust in ai: The interplay of antecedents, consequences, and cultural dynamics. *AI & SOCIETY*, 41(1):669–692, 2026.

[13] Md Meftahul Ferdaus, Mahdi Abdelguerfi, Elias Loup, Kendall N. Niles, Ken Pathak, and Steven Sloan. Towards trustworthy ai: a review of ethical and robust large language models. *ACM Computing Surveys*, 58(7):1–43, 2026.

[14] Chenjun Guo, Antoni Borghini, Kosa Goucher-Lambert, and Gaëlle Baudoux. Human-gen ai co-design: Exploring factors impacting trust calibration. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 89244, page V004T06A040. American Society of Mechanical Engineers, 2025.

[15] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[16] KPMG. Ai in financial reporting and audit: Navigating the new era. kpmg international. *KPMG*, 2025.

[17] Florian Kücking, Ursula Hübner, Mareike Przysucha, Niels Hannemann, Jan-Oliver Kutza, Maurice Moelleken, Cornelia Erfurt-Berge, Joachim Dissemond, Birgit Babitsch, and Dorothee Busch. Automation bias in ai-decision support: Results from an empirical study. In *German Medical Data Sciences 2024*, pages 298–304. IOS Press, 2024.

[18] Tina B Lassiter. The role of algorithmic audits and other soft law approaches in informing users' calibrated trust in artificial intelligence tools. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 54–56, 2024.

[19] Johann Laux and Hannah Ruschemeier. Automation bias in the ai act: On the legal implications of attempting to de-bias human oversight of ai. *European Journal of Risk Regulation*, pages 1–16, 2025.

[20] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[21] Yugang Li, Baizhou Wu, Yuqi Huang, and Shenghua Luan. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-ai trust. *Frontiers in psychology*, 15:1382693, 2024.

[22] Magnus Liebherr, Ellen Enkel, Effie L-C Law, Mohammad Reza Mousavi, Matteo Sammartino, and Philipp Sieberg. Dynamic calibration of trust and trustworthiness in ai-enabled systems. *International Journal on Software Tools for Technology Transfer*, pages 1–17, 2026.

[23] Gale M Lucas, Burcin Becerik-Gerber, and Shawn C Roll. Calibrating workers trust in intelligent automated systems. *Patterns*, 5(9), 2024.

[24] Simon Mahler. *Building trust in workplace AI: Why governance outweighs employee co-creation in building trust*. PhD thesis, FH Vorarlberg (Fachhochschule Vorarlberg), 2025.

[25] Guido Marchi. Decoding the "black-box": explainable artificial intelligence towards trustworthy advancement in respiratory medicine. *Breathe*, 22(1):250318, 2026.

[26] Frederic Marimon, Marta Mas-Machuca, and Anna Akhmedova. Trusting in generative ai: Catalyst for employee performance and engagement in the workplace. *International Journal of Human–Computer Interaction*, 41(11):7076–7091, 2025.

[27] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4):1–45, 2024.

[28] Saumya Pareek, Niels Van Berkel, Eduardo Velloso, and Jorge Goncalves. Effect of explanation conceptualisations on trust in ai-assisted credibility assessment. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–31, 2024.

[29] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Human-ai interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15, 2021.

[30] Keonyoung Park and Ho Young Yoon. Beyond the code: The impact of ai algorithm transparency signaling on user trust and relational satisfaction. *Public Relations Review*, 50(5):102507, 2024.

[31] Sanket Ramchandra Patole, Getrude Hewapathirana, and Smita Padmanabhan. Artificial intelligence and organizational culture: A research agenda for trust, equity, and learning in hrd, 2026. Manuscript / research agenda.

[32] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. Investigations of performance and bias in human-ai teamwork in hiring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12089–12097, 2022.

[33] Katarzyna Prędkiewicz and Krzysztof Biegun. Factors that influence accountants' acceptance of artificial intelligence: An extended technology acceptance model that incorporates technology anxiety and experience. *Zeszyty Teoretyczne Rachunkowości*, 49(49 (4)):147–168, 2025.

[34] Gopalan Puthukulam, Anitha Ravikumar, Ravi Vinod Kumar Sharma, and Krishna Murthy Meesaala. Auditors' perception on the impact of artificial intelligence on professional skepticism and judgment in oman. *Universal Journal of Accounting and Finance*, 9(5):1184–1190, 2021.

[35] Massimo Regona, Tan Yigitcanlar, Carol Hon, and Melissa Teo. Building trust in artificial intelligence: A systematic review through the lens of trust theory. *ACM Computing Surveys*, 2026.

[36] Giuseppe Romeo and Daniela Conti. Exploring automation bias in human–ai collaboration: a review and implications for explainable ai. *Ai & Society*, 41(1):259–278, 2026.

[37] Iqbal H Sarker, Helge Janicke, Ahmad Mohsin, and Leandros Maglaras. Sme-team: leveraging trust and ethics for secure and responsible use of ai and llms in smes. *npj Artificial Intelligence*, 2(1):12, 2026.

[38] Max Schemmer, Niklas Kühl, and Gerhard Satzger. Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence. *arXiv preprint arXiv:2109.13827*, 2021.

[39] Malika Soulami, Saad Benchekroun, and Asiya Galiulina. Exploring how ai adoption in the workplace affects employees: a bibliometric and systematic review. *Frontiers in artificial intelligence*, 7:1473872, 2024.

[40] Varbinka Stefanova-Stoyanova and Petko Danov. Compliance with iso/iec 42001, 22989, 27001 standards and integration of human-machine teaming in ai lifecycle management. In *2025 XXXIV International Scientific Conference Electronics (ET)*, pages 1–4, 2025.

[41] Aurelia Tamò-Larrieux, Clement Guitton, Simon Mayer, and Christoph Lutz. Regulating for trust: Can law establish trust in artificial intelligence? *Regulation & Governance*, 18(3):780–801, 2024.

[42] Carmen Wendy Ulizio, Devika Dua, Naya Meenkashi Mukul, Santosh Areti, Kristin Kostick-Quenet, and Vasiliki Nataly Rahimzadeh. Exploring the ethical and practical considerations of artificial intelligence in real-world health care settings: Stakeholder focus group study. *JMIR AI*, 5(1):e85163, 2026.

[43] Yinying Wang. Algorithmic decision-making in organizations: a systematic review toward an integrated tension alignment framework. *Organization Management Journal*, 23(1):115–131, 2026.

[44] Yanjun Wen, Jiale Wang, and Xiaoxi Chen. Trust and ai weight: human-ai collaboration in organizational management decision-making. *Frontiers in Organizational Psychology*, 3:1419403, 2025.

[45] Magdalena Wischnewski, Alisa Scharmann, Annika Ridder, and Nicole Krämer. Certified but imperfect: Investigating the role of ai certifications and system performance on trust in and reliance on ai systems. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2026.

[46] Gerui Xu, Shruthi Venkatesha Murthy, and Bochen Jia. Enhancing intuitive decision-making and reliance through human–ai collaboration: A review. In *Informatics*, volume 12, page 135. MDPI, 2025.

[47] Chen Zhong and Sunita Goel. Transparent ai in auditing through explainable ai. *Current Issues in Auditing*, 18(2):A1–A14, 2024.

[35] [43] [21] [16] [37] [31] [7] [12] [5] [4] [38] [13] [3] [34] [17] [33] [11] [1] [18] [15] [45] [25] [24] [2] [23] [22] [46] [10] [6] [36] [40] [19] [32] [9] [14] [41] [28] [30] [20] [26] [39] [47] [44] [8] [27] [42] [29]