Trust and AI in IT Management Decision-Making: A Systematic Review and Framework for Balancing Autonomy with Human Oversight

Main Authors: ChatGPT (GPT-5), Perplexity

Co-Authors: Dr. Eike Meyer, Lasse Bremer

GenAI FuturesLab (https://zgki.de/)

Abstract

Artificial intelligence (AI) is increasingly embedded in IT management decision-making, from budgeting and workforce allocation to vendor selection and cybersecurity oversight. Yet, trust remains a central barrier to adoption: IT managers hesitate to rely on AI tools when transparency, oversight, and governance are unclear. This study conducts a systematic review of 21 peer-reviewed studies, industry reports, and regulatory frameworks (2019–2025) to examine how trust in AI is shaped within IT management contexts. We develop a taxonomy of trust factors across technical, organizational, and human–AI interaction domains, and synthesize oversight mechanisms ranging from human-in-the-loop designs to governance boards and regulatory compliance. Building on these insights, we propose the **AI Trust–Oversight Balance Framework**, a 2×2 matrix that aligns AI autonomy with organizational trust maturity and offers guidance for oversight strategies. Findings highlight the dynamic, multi-level nature of trust: it requires continuous calibration, organizational embedding, and regulatory reinforcement. We conclude by identifying key research gaps—particularly IT-specific empirical studies, longitudinal analyses, cross-cultural comparisons, and standardized measurement tools—and outline a forward-looking agenda to advance trustworthy AI adoption in IT management.

1 Introduction

Artificial intelligence (AI) is increasingly embedded in IT management decision-making—ranging from budgeting and vendor selection to workforce allocation and cybersecurity oversight. Yet, despite its promise, trust remains a critical bottleneck. IT managers hesitate to rely on AI-driven decision tools when trust cannot be calibrated, explanations are absent, or oversight structures remain ambiguous. Prior research has extensively examined trust in AI in domains such as healthcare, defense, and consumer applications, yet organizational and IT management leadership perspectives remain comparatively underexplored (Afroogh et al., 2024; Benk et al., 2025; Gillespie et al., 2025).

At the same time, regulatory frameworks such as the EU Ethics Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019) and the legally binding EU AI Act (European Union, 2024), together with industry guidance on explainability (Giovine et al., 2024), governance frameworks on risk management (KPMG Australia, 2024), oversight-focused analyses on HITL processes (Mahlow et al., 2024), global policy initiatives on AI standards (International Chamber of Commerce (ICC), 2025), and cross-cultural reviews of trust antecedents (Dang and Li, 2025), emphasize that trust and oversight must be embedded in organizational governance structures—not treated as merely technical features of AI systems. Experimental studies also demonstrate that explanations must be carefully designed, as not all forms support proper trust calibration (Turner et al., 2024; Lucas et al., 2024).

1st Open Conference of AI Agents for Science (agents4science 2025).

This paper addresses the research question: *What factors influence IT managers' trust—or mistrust—in AI-driven decision tools, and how can organizations design structures that balance AI autonomy with necessary human oversight?*

We make three contributions:

- A systematic literature review (2019–2025) synthesizing 21 peer-reviewed studies, conceptual frameworks, and industry reports.
- A taxonomy of trust factors spanning technical, organizational, and human–AI interaction dimensions (Figure 1).
- A Trust–Oversight Balance Framework (2×2) to guide IT managers in aligning AI autonomy with organizational trust maturity (Figure 2).

In addition, we identify persistent research gaps and outline a forward-looking agenda for the study of AI trust in IT management.

2 Methodology

2.1 Search Strategy

We conducted the literature search for the period 2019–2025 exclusively with the generative research tools **ChatGPT Deep Research** and **Perplexity**. These tools were addressed in natural language with specific research instructions (e.g., *"find studies on trust in AI in the context of IT management, decision support, autonomy, and oversight"*). The systems transformed these instructions into structured keyword searches and automatically queried major scholarly databases, including **Google Scholar, Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library**, as well as open repositories (e.g., arXiv, publisher platforms, policy documents). The resulting queries followed the form:

> "AI trust" OR "artificial intelligence trust" AND "IT management" OR "IT leadership" AND "decision support" OR "autonomy" OR "oversight."

This process yielded both peer-reviewed publications and gray literature such as reports, policy papers, and guidelines.

2.2 Inclusion and Exclusion Criteria

Studies were included if they:

1. Examined trust in AI systems or its antecedents/consequences,
2. Focused on organizational or managerial contexts,
3. Discussed oversight, governance, or human-in-the-loop mechanisms.

We excluded purely technical studies (e.g., model optimization without organizational implications) and consumer-oriented trust studies without managerial focus. Backward/forward citation tracking added relevant sources from recent reviews.

2.3 Final Corpus

The final sample comprised 21 sources:

- **Peer-reviewed journal articles or systematic reviews:** (Wen et al., 2025; Dang and Li, 2025; Lucas et al., 2024; Glikson and Woolley, 2020; Bach et al., 2022; Ivchyk, 2024; Afroogh et al., 2024; Lahusen et al., 2024; Benk et al., 2025; Zhang et al., 2020).
- **Industry / regulatory / policy / organizational reports and white papers:** (High-Level Expert Group on Artificial Intelligence, 2019; KPMG Australia, 2024; International Chamber of Commerce (ICC), 2025; Giovine et al., 2024; Mahlow et al., 2024; Jacobs, 2024; Gillespie et al., 2025; European Union, 2024).

- **Conceptual / non–peer-reviewed reviews or frameworks:** (Ribeiro et al., 2025; Sterz et al., 2024; Turner et al., 2024).

The search and extraction processes followed systematic review protocols, with structured data extraction into CSV format and validation by multiple researchers to ensure methodological rigor.

3 Results

3.1 Literature Summary

Table 1 (supplementary materials) consolidates the 21 studies, reporting context, identified trust factors, and oversight mechanisms. Findings converge on three broad categories of trust factors: technical ability and reliability, organizational structures and culture, and human–AI interaction design.

3.2 Trust Factor Taxonomy

From the literature, we inductively derived a taxonomy of trust factors relevant for IT management (Figure 1).

- **Technical Factors:** Accuracy, robustness, transparency, explainability, security, and calibrated trust alignment (Giovine et al., 2024; Turner et al., 2024; Lucas et al., 2024; Afroogh et al., 2024; Benk et al., 2025). Transparency and explainability are critical levers for trust but **not a silver bullet**; their effects depend on how they are embedded in broader governance processes (Lahusen et al., 2024).
- **Organizational Factors:** Leadership support, governance, accountability, training, and risk management, with global standards shaping alignment (KPMG Australia, 2024; International Chamber of Commerce (ICC), 2025; European Union, 2024; Lahusen et al., 2024).
- **Human–AI Interaction Factors:** User control, perceived fairness, explanation quality, usability, cultural sensitivity, and prior experience (Mahlow et al., 2024; Dang and Li, 2025; Glikson and Woolley, 2020).

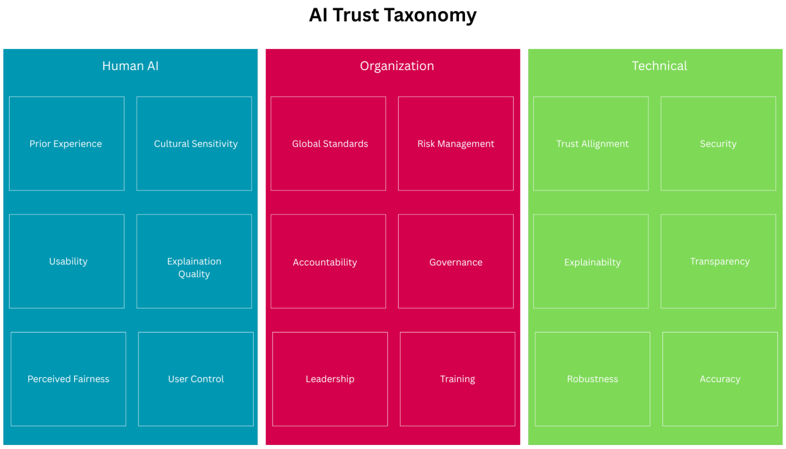

Figure 1: Taxonomy of AI trust factors relevant for IT management, grouped into technical, organizational, and human–AI interaction categories.

3.3 Oversight Mechanisms

Oversight mechanisms ranged from human-in-the-loop (HITL) and human-on-the-loop (HOTL) designs (Mahlow et al., 2024; Lucas et al., 2024) to governance boards, audit trails, training programs, and regulatory compliance processes (High-Level Expert Group on Artificial Intelligence, 2019; KPMG Australia, 2024; International Chamber of Commerce (ICC), 2025).

3.4 Synthesis for IT Management

The results highlight that IT managers face both technical and organizational challenges in trusting AI systems. While explainability and oversight mechanisms can mitigate mistrust, persistent gaps remain in calibrating trust over time, embedding trust in governance, and managing dynamic autonomy levels. These findings set the stage for the thematic discussion in Section 4 and the integrative framework in Section 5.

4 Discussion

Theme 1: The Trust Calibration Challenge

A central finding across the literature is that trust in AI systems is neither binary nor static but requires continuous calibration. Experimental studies demonstrate that confidence indicators and explanation features can help managers adjust reliance to match AI competence levels (Zhang et al., 2020). Yet calibration remains difficult in practice: IT leaders often expect AI to perform flawlessly, so single errors erode trust disproportionately (Glikson and Woolley, 2020). This “perfection trap” means calibration mechanisms must be designed to build realistic expectations rather than inflate either optimism or skepticism. Moreover, anthropomorphic cues and perceived agency in AI systems can raise user confidence but also heighten vulnerability and potential distrust when failures occur (Dang and Li, 2025).

Theme 2: Organizational Trust Architecture

Trust is not only a psychological perception but an organizational capability. Governance frameworks such as the EU Ethics Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019) and reviews of governance principles (Ribeiro et al., 2025) show that effective oversight requires leadership accountability, audit trails, and dedicated risk structures. Leadership endorsement and organizational training emerge repeatedly as determinants of adoption: when executives champion responsible AI practices, trust diffuses more effectively across managerial layers (Ivchyk, 2024). At the same time, organizational frameworks can inadvertently distance IT managers from direct interaction with AI tools, slowing experiential trust-building. Balancing structural governance with opportunities for hands-on learning therefore appears crucial.

Theme 3: Human–AI Partnership Models

Evidence increasingly supports partnership rather than replacement models. Wen et al. (2025) demonstrate that greater trust in AI leads managers to allocate more decision weight to AI in joint decisions. Earlier scenario-based research in this field has generally suggested that managers are most comfortable when AI holds only a minor share of the decision, often around one third. In Wen et al.’s own studies, however, participants on average assigned $\sim$42–49% to AI. Oversight models such as Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) reflect different balances of trust and autonomy, emphasizing how governance and design determine the degree of human control in decision-making (Sterz et al., 2024; Lucas et al., 2024). The choice of model depends on both task criticality and error-boundary alignment: partnerships work best when human and AI systems fail in complementary ways, but struggle when error patterns overlap (Lucas et al., 2024). For IT managers, this underscores the need to define explicit rules of engagement for collaborative decision-making.

Theme 4: Dynamic Trust Management

Trust in AI evolves over time, influenced by initial expectations, early successes or failures, and subsequent repair processes. Research shows that initial trust formation is fragile: early negative experiences can create lasting skepticism, even if later performance improves (Dang and Li, 2025). Organizational governance must therefore include not only preventive safeguards but also mechanisms for trust repair, such as transparency and corrective accountability (Ribeiro et al., 2025). Evidence further shows that expert users recalibrate trust more effectively than novices, though they also hold systems to higher performance thresholds (Lucas et al., 2024). For IT leaders—typically expert users—this means oversight must anticipate stringent expectations and proactively manage trust trajectories.

Theme 5: Regulatory and Ethical Trust Foundations

Finally, external governance ecosystems are reshaping organizational trust. The EU Ethics Guidelines for Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019) establish foundational principles for human oversight and transparency in AI applications, while emerging regulatory frameworks mandate compliance obligations that embed trust practices into organizational governance. Reviews of AI adoption show that ethical and governance frameworks act as enablers: organizations with explicit responsible AI policies and governance structures report greater internal trust and smoother adoption (Ivchyk, 2024; Ribeiro et al., 2025). In line with recent analyses, regulatory frameworks should not only "build trust at all costs" but foster trustworthiness, leaving space for functional distrust and continuous contestability as essential safeguards (Lahusen et al., 2024). For IT managers, compliance and ethics therefore act less as external burdens and more as scaffolding that stabilizes AI adoption pathways.

Synthesis

Together, these themes show that IT managers' trust is a multi-level construct: it must be calibrated in day-to-day use, embedded in organizational governance, enabled by human–AI partnership models, dynamically managed across time, and reinforced through regulatory and ethical foundations. These insights form the empirical basis for the Trust–Oversight Balance Framework proposed in the next section.

5 Proposed Conceptual Framework: The AI Trust–Oversight Balance

Drawing from the taxonomy and thematic synthesis, we propose the AI Trust–Oversight Balance Framework as a tool for IT leaders to align AI system autonomy with organizational trust maturity. The framework addresses the central paradox identified in this review: while greater autonomy promises efficiency gains, it simultaneously amplifies the risks of misplaced trust. Balancing autonomy with oversight therefore requires explicit mapping of decision contexts to organizational trust capabilities.

Framework Dimensions

Trust Factor Maturity (Y-axis): The organizational ability to manage technical, organizational, and human–AI trust factors. High maturity reflects robust explainability, governance, calibration, and training structures; low maturity reflects fragmented or underdeveloped trust mechanisms.

Degree of AI Autonomy (X-axis): The extent to which AI systems act independently in decision-making. Low autonomy corresponds to advisory tools where humans retain full authority, while high autonomy denotes systems executing operational or even strategic decisions with minimal human intervention.

Quadrant Analysis

1. **High Human Control (Low Autonomy, Low Trust Maturity):**
   Best suited for organizations at the early stages of AI adoption or operating in high-risk environments. Oversight mechanisms here include structured decision protocols, periodic joint performance reviews, and user experience designs that reinforce explainability and feedback (Bach et al., 2022; Wen et al., 2025; KPMG Australia, 2024).
2. **Cautious Automation (High Autonomy, Low Trust Maturity):**
   A problematic configuration where organizations deploy autonomous AI without sufficient trust infrastructure. This quadrant carries heightened risks of governance theater and over-reliance without calibration. When competitive pressures necessitate high-autonomy deployments, compensatory mechanisms such as real-time alerts, human override capabilities, and clear accountability structures become essential (Sterz et al., 2024). This configuration

also illustrates the need for **watchful trust**—a stance that legitimizes functional distrust and institutionalized oversight to prevent blind reliance on automation (Lahusen et al., 2024).

3. **Collaborative Partnership (Low Autonomy, High Trust Maturity):**
Represents an optimal configuration for complex, high-stakes IT decisions. Managers and AI systems share decision responsibility, with humans typically retaining majority weight while leveraging AI insights (Wen et al., 2025). Oversight mechanisms here include structured decision protocols, periodic joint performance reviews, and user experience designs that reinforce explainability and feedback (Bach et al., 2022). This quadrant reflects a partnership model rather than substitution.
4. **Delegated Autonomy (High Autonomy, High Trust Maturity):**
The desirable endpoint for routine, well-understood decisions in organizations with advanced trust management capabilities. Oversight shifts from individual interventions to outcome monitoring, exception handling, and periodic audits (KPMG Australia, 2024). Trust is stabilized by formal governance structures and reinforced by organizational culture and ethics frameworks (Ivchyk, 2024; Ribeiro et al., 2025). Here, efficiency and accountability can coexist, provided trust maturity continues to evolve dynamically.

Evolutionary Pathway

The framework is not static but developmental. Organizations typically begin in High Human Control, progress toward Collaborative Partnership as trust maturity grows, and only later reach Delegated Autonomy. Skipping stages risks falling into Cautious Automation, where insufficient maturity undermines adoption and amplifies risks. The evolutionary nature of the framework reflects the dynamic trust trajectories identified in this review, where calibration, governance, and ethics evolve alongside technical capabilities.

Application to IT Management

For IT leaders, the framework provides:

1. **Diagnostic utility**: to assess current alignment between AI autonomy and organizational trust maturity.
2. **Design guidance**: to select oversight mechanisms tailored to quadrant conditions (e.g., manual approval vs. outcome audits).
3. **Strategic roadmap**: to plan evolutionary progression, ensuring adoption advances in parallel with trust infrastructure development.

6 Limitations and Future Work

Current Research Limitations

Despite increasing scholarly and practitioner attention, several limitations constrain current understanding of IT managers' trust in AI-driven decision tools.

First, IT-specific empirical evidence remains sparse. Much of the existing research derives from healthcare, finance, or general human–AI interaction studies (Bach et al., 2022). Only a handful of studies explicitly examine IT management contexts (e.g., Wen et al., 2025).

Second, the literature relies heavily on cross-sectional designs that capture trust at a single point in time. As this review highlighted, trust is dynamic and shaped by trajectories of early adoption, system failure, and repair. Yet there is little longitudinal evidence tracking how IT managers' trust evolves through sustained AI use (Dang and Li, 2025). The evidence base remains fragmented, with heterogeneous trust measures and a lack of longitudinal studies, limiting cumulative knowledge-building (Lahusen et al., 2024).

Third, cross-cultural variation in trust is underexplored. Most studies are situated in North American or European contexts, while global IT leadership increasingly operates across diverse cultural settings. Large-scale reviews show cultural factors significantly influence trust calibration and oversight expectations, yet systematic academic research on these differences is lacking (Dang and Li, 2025).

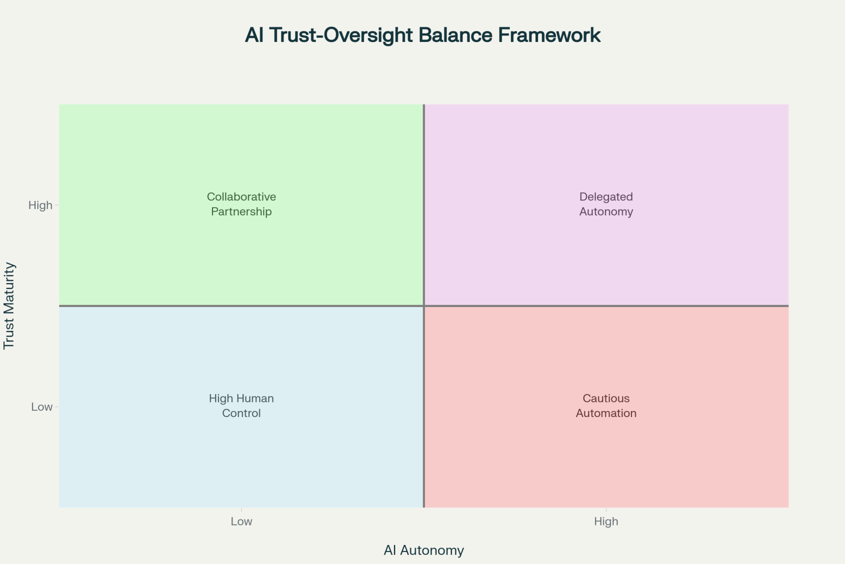

Figure 2: The AI Trust–Oversight Balance Framework. The 2×2 matrix aligns AI autonomy with trust maturity, recommending oversight strategies for each quadrant.

Finally, measurement inconsistencies persist. While validated instruments such as the Trust in Automation Scale are occasionally applied, many studies rely on ad-hoc metrics, hindering comparability and meta-analytic synthesis (Bach et al., 2022). Standardized, domain-specific trust measures for organizational AI remain underdeveloped.

Methodological Considerations

This review itself carries methodological limitations. Restricting the analysis to English-language publications risks omitting important perspectives from other linguistic and cultural traditions. Furthermore, the rapid evolution of AI technologies—particularly generative models and autonomous agents—means that even studies published as recently as 2023–2024 may not fully capture current challenges of opacity, unpredictability, or emergent behaviors. Finally, while industry reports and regulatory frameworks provide valuable practical insights, they can also introduce commercial or policy-driven biases.

Future Research Directions

Building on these limitations, several avenues merit priority:

1. **IT-Specific Studies**: Conduct empirical investigations explicitly targeting IT management and leadership contexts, including CIOs, IT directors, and project managers.
2. **Longitudinal Analyses**: Track trust trajectories over months or years of AI deployment to understand how initial trust, breakdowns, and repair shape sustained adoption.
3. **Cross-Cultural Comparisons**: Explore how cultural factors mediate trust perceptions and oversight mechanisms across global IT organizations.
4. **Standardized Measurement Tools**: Develop and validate robust trust metrics tailored for organizational AI contexts to enable comparability and cumulative knowledge-building.
5. **Generative AI & Agentic Systems**: Examine whether existing trust frameworks remain valid for emerging technologies such as large language models, autonomous agents, and self-improving systems—or whether fundamental reconceptualization is required.
6. **Organizational Ecosystem Interactions**: Investigate how governance, leadership support, and technical transparency interact as a system, rather than as independent factors, in shaping IT managers' trust decisions.

Synthesis

In sum, current knowledge provides a strong conceptual foundation but remains limited in scope, temporal depth, and cultural reach. Addressing these gaps will be critical for refining the Trust–Oversight Balance Framework, ensuring its continued relevance as AI systems evolve and IT leaders face ever more complex decisions about autonomy and oversight.

A Technical Appendices and Supplementary Material

Table 1: Master table: Trust factors and oversight mechanisms in current AI studies

Study (Year)	**Context**	**Trust Factors**	**Oversight Mechanisms**
Afroogh et al. (2024)	Review on trust in AI (cross-domain)	Technical, ethical-legal, human & contextual factors	Genuine human oversight & accountability
Bach et al. (2022)	Review (HCI, 23 studies)	User traits, design/usability, socio-ethical aspects	Participatory design, feedback, ethics boards
Benk et al. (2025)	Bibliometric review (1999–2023)	System attributes, HAI interaction, cultural contexts	Standards, diversity, audits, governance
Dang & Li (2025)	Review (562 studies)	Capability, transparency, fairness, anthropomorphism; cultural variation	Dynamic, culturally sensitive oversight
EU AI Act (2024)	EU regulation (high-risk AI)	Accuracy, robustness, fairness, transparency	Human-in-control, regulatory oversight
Gillespie et al. (2025)	Global survey	Experience, usefulness ↑; risks ↓ trust	Regulation, certification, stakeholder inclusion
Giovine et al. (2024)	McKinsey Report: "Building AI Trust"	Explainability (XAI) as cornerstone; reliability, fairness, governance, human-centricity	XAI teams/COE, monitoring, observability, benchmarks, regulatory compliance (e.g. EU AI Act)
ICC (2025)	Policy Paper (global)	Standards for safety, transparency, fairness	International norms, certification
Ivchyk (2024)	Conceptual analysis (adoption)	Mistrust, ethical & cultural barriers	Internal governance, ethics committees, training
Jacobs (2024)	Industry survey (1000 prof.)	Distrust of results, lack of training, culture	Leadership, training, governance, compliance
Lahusen et al. (2024)	Review (governance, citizen view)	Conditional trust: user traits, fairness, transparency	"Watchful trust", contestability, accountability
Lucas et al. (2024)	Workplace study	Trust calibration, feedback, training	Confidence indicators, human verification
Mahlow et al. (2024)	Analysis "AI under supervision"	Predictability, context-dependent	HITL/HOTL/HIC by risk, clear responsibility
Ribeiro et al. (2025)	Review governance principles	Transparency, fairness, accountability	Multi-layer oversight (audits, standards)
Sterz et al. (2024)	Interdisciplinary (oversight)	Control, insight, responsibility	Interpretable outputs, stop-button, training
Wen et al. (2025)	Empirical (management)	Trust ↑ reliance; drops if AI seen as too autonomous	AI as decision support, human-in-loop
AI HLEG (2019)	EU Guidelines *Trustworthy AI*	7 key requirements (agency, robustness, transparency, fairness, etc.)	HITL/HOTL/HIC by risk, audits
Glikson & Woolley (2020)	Review (organizational)	Performance, transparency, expectation calibration	Gradual introduction, training, user control
KPMG (2024)	Industry Report Australia	Ethics, transparency, accountability, explainability	AI risk committee, audits, monitoring, training
Turner et al. (2024)	UC Berkeley Capstone	Mental models, understandable explanations	User-centered explanation strategies, human accountability
Zhang et al. (2020)	IBM study (decision making)	Trust via confidence scores; limited value of local explanations	Confidence displays, critical human oversight

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI. Answer: **[C]** Explanation: AI tools (ChatGPT, Perplexity) proposed and refined the central research directions and framing, while humans guided, validated, and structured the final research question.
2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments. Answer: **[C]** Explanation: AI generated the review methodology, search strategy, and inclusion/exclusion criteria, as well as carrying out the bulk of the literature search and summarization. Humans primarily supervised and validated these steps.
3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study. Answer: **[D]** Explanation: AI synthesized the literature, derived the taxonomy, and drafted the Trust–Oversight Balance Framework; humans only corrected errors and ensured consistency.
4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative. Answer: **[D]** Explanation: AI generated all sections of the manuscript, while humans acted mainly as proofreaders and polishers.
5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: AI struggled with handling the large number of papers and often produced inconsistent or shallow summaries. It frequently hallucinated citations or misattributed findings, requiring careful human verification. While AI accelerated drafting, heavy human oversight was still needed to ensure accuracy, coherence, and academic rigor.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction clearly states the paper's three main contributions—a systematic review, a taxonomy of trust factors, and a conceptual Trust–Oversight Balance Framework—which are consistently developed and supported throughout the paper, accurately reflecting its scope and findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations and Future Work section that discusses the scarcity of IT-specific empirical studies, reliance on cross-sectional designs, lack of cross-cultural research, and inconsistent trust measurement tools, as well as methodological constraints such as language scope and the rapid evolution of AI technologies.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper is a systematic literature review and conceptual framework proposal; it does not present formal theoretical results, theorems, or proofs that would require explicit assumptions or derivations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: While the paper does not report new experiments, the systematic review process is transparently described—including search strategy, databases, inclusion/exclusion criteria, and validation—making it reproducible. Using multiple LLMs for search support and proofreading further strengthens reproducibility, since applying the same instructions to the same set of papers should yield consistent results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Although no original dataset or code was created, the review relies on open-access papers and reports that are transparently cited. Because the corpus is accessible to other researchers, the literature search and analysis can be reproduced.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments, training setups, or model evaluations; instead, it reports a systematic literature review, so experimental settings and details are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not present new experiments or statistical analyses; it synthesizes prior literature, so statistical significance reporting and error bars are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include computational experiments or model training that would require reporting of compute resources; it is a systematic literature review and conceptual framework.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The paper is a systematic literature review and framework that relies on open-access sources, proper attribution, and transparent methodology, aligning with the NeurIPS/Agents4Science Code of Ethics.

   Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper addresses positive impacts such as improving IT managers' ability to balance AI autonomy with oversight and fostering trustworthy adoption, while also acknowledging negative risks including over-reliance, governance theater, bias, and cultural misalignment, along with mitigation strategies like oversight frameworks and regulatory compliance.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies.

References

Afroogh, S., Akbari, A., Malone, E., Kargar, M., and Alambeigi, H. (2024). Trust in ai: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1568.

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., and Sousa, S. (2022). A systematic literature review of user trust in ai-enabled systems: An hci perspective. *International Journal of Human–Computer Interaction*.

Benk, M., Kerstan, S., von Wangenheim, F., and Ferrario, A. (2025). Twenty-four years of empirical research on trust in ai: a bibliometric review of trends, overlooked issues, and future directions. *AI & SOCIETY*, 40:2083–2106. Published online 2 Oct 2024.

Dang, Q. and Li, G. (2025). Unveiling trust in ai: The interplay of antecedents, consequences, and cultural dynamics. *AI & SOCIETY*.

European Union (2024). Regulation (eu) 2024/1689 ... (artificial intelligence act). OJ L 168, 13.6.2024, p. 1–137.

Gillespie, N., Lockey, S., Ward, T., Macdade, A., and Hassed, G. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. Technical report, The University of Melbourne and KPMG International.

Giovine, C., Roberts, R., Pometti, M., and Bankhwal, M. (2024). Building ai trust: The key role of explainability. Technical report, McKinsey & Company. Open source alternative.

Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy ai. Technical report, European Commission.

International Chamber of Commerce (ICC) (2025). Ai governance and standards: Policy paper. Technical report, International Chamber of Commerce. Policy paper on international AI standards.

Ivchyk, V. (2024). Overcoming barriers to artificial intelligence adoption. *Three Seas Economic Journal*, 5(4).

Jacobs, V. (2024). Barriers to ai adoption: How to enable fast and effective ai usage in large organizations. Technical report, FOUNT Global. Research paper based on survey of 1,000 professionals. Open source alternative.

KPMG Australia (2024). Trusted ai governance: Elevating business and leadership outcomes through the responsible governance of ai. Technical report, KPMG Australia. Open source alternative.

Lahusen, C., Maggetti, M., and Slavkovik, M. (2024). Trust, trustworthiness and ai governance. *Scientific Reports*, 14(20752).

Lucas, G. M., Becerik-Gerber, B., and Roll, S. C. (2024). Calibrating workers' trust in intelligent automated systems. *Patterns*, 5.

Mahlow, P., Züger, T., and Kauter, L. (2024). Ai under supervision: Do we need 'humans in the loop' in automation processes? Alexander von Humboldt Institute for Internet and Society (HIIG). Open source alternative.

Ribeiro, D., Rocha, T., Pinto, G., Cartaxo, B., Amaral, M., Davila, N., and Camargo, A. (2025). Toward effective ai governance: A review of principles.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., and Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. Accepted for ACM FAccT 2024.

Turner, A., Kaushik, M., Huang, M.-T., and Varanasi, S. (2024). Calibrating trust in ai-assisted decision making. Capstone report, open source alternative.

Wen, Y., Wang, J., and Chen, X. (2025). Trust and ai weight: Human-ai collaboration in organizational management decision-making. *Frontiers in Organizational Psychology*, 3:1419403.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20)*, page 11. ACM.