Calibrating Trust in AI for Drafting-Related Professional Tasks: A Scoping Review

Cemre Ozcan Meric Ozler

Politecnico di Torino, Turin, Italy

s334701@studenti.polito.it s336819@studenti.polito.it

Abstract

AI is increasingly used to augment professional drafting, summarisation, information retrieval, and decision-support tasks, yet the conditions under which worker trust in such systems is warranted remain unclear. This scoping review maps factors that shape well-calibrated trust in AI for drafting-related professional work, following PRISMA-ScR reporting guidance [23] and a Type 3 framing focused on worker trust rather than broad adoption. Fifty sources were included across empirical, theoretical, review-oriented, legal, practice, and policy literature. The evidence spans clinical decision support, manufacturing, speech annotation, human-AI collaboration, architectural practice, and legal professional settings. Four recurring dimensions structure the findings: contextual trust conditions, intrinsic trustworthiness, trust calibration dynamics, and task stakes with human review. A cross-cutting distinction between AI types — LLMs, narrow ML systems, and XAI-equipped tools — affects calibration requirements. Cross-domain evidence from medicine and architecture supports the framework's potential transferability. Calibrated trust is conditionally warranted for low-risk, reviewable support tasks, but not as a substitute for expert judgment in high-stakes professional work.

Keywords: trust calibration, human-AI collaboration, professional drafting, generative AI, scoping review, overtrust, AI governance, PRISMA-ScR, LLM, automation bias.

1 Introduction

Artificial intelligence is increasingly embedded in knowledge work, where it supports drafting, summarisation, information retrieval, and decision support. This shift raises a practical question about judgment: workers must decide when an AI output is good enough to use, when it requires verification, and when it should not be trusted at all [1, 2, 22].

In this review, *drafting-related professional tasks* means work in which AI assists the production, structuring, revision, or evidential support of professional text or documentation. Examples include legal memoranda, contract clauses, clinical summaries, architectural design descriptions, internal reports, and policy or compliance documents. The scope excludes purely technical model development and fully automated decisions where no professional text or reviewable output is produced.

AI type matters. LLMs — including general systems such as ChatGPT or Claude and legal tools such as Harvey, Lexis+ AI, or CoCounsel — can generate fluent prose even when content is unsupported [9, 29]. Narrow ML classifiers and XAI-equipped systems raise different calibration problems: confidence scores or explanations can help users, but may also induce automation bias when they appear more reliable than actual performance [8, 5, 15, 50]. Workers can also exhibit *algorithm appreciation* — a preference for algorithmic judgment over human judgment that, when uncritical, is itself a form of miscalibration [49].

This review asks: *What factors determine whether a worker's trust in an AI system is well calibrated for drafting-related professional tasks?* Four sub-questions guide the synthesis: what contextual conditions make trust warranted; what intrinsic system properties shape trustworthiness; what factors lead to overtrust or undertrust; and how task stakes and human oversight shape appropriate reliance. The main analysis remains domain-agnostic; the worked application focuses on legal drafting.

2 Methods

2.1 Study Design and Search

This study used a scoping review approach, guided by PRISMA-ScR [23] and scoping review guidance from Peters et al. [42]. The aim was not to estimate one intervention effect, but to map a mixed body of empirical studies, theoretical trust models, reviews, legal analyses, and institutional guidance.

Searches were conducted on 10–11 April 2026, with supplementary searches on 23–24 April 2026 to expand cross-domain coverage. Sources included Google Scholar, SSRN, Springer, ACM Digital Library, PubMed, journal websites, and institutional websites. Backward citation searching identified foundational sources. No publication date restriction was applied; pre-2015 sources were retained when theoretically foundational. Boolean strings included *"calibrated trust" AND "AI" AND ("worker" OR "professional")*, *"trust calibration" AND ("human-AI collaboration" OR "automation bias")*, *"generative AI" AND ("legal drafting" OR "professional drafting") AND "trust"*, *"hallucination" AND "AI" AND ("professional" OR "legal" OR "clinical")*, *"AI" AND ("clinical decision support" OR "CDSS") AND ("trust" OR "reliance")*, *"AI" AND "architecture" AND ("professional practice" OR "trust")*, and *"XAI" AND ("trust calibration" OR "overtrust") AND "human-AI"*. The full log is in Appendix A and in the supporting ZIP.

2.2 Eligibility, Screening, and Coding

Sources were included if they addressed trust in AI systems, trust calibration or appropriate reliance, human-AI collaboration, AI-supported drafting, oversight, hallucination risk, AI output reliability, AI type distinctions, or professional responsibility in AI-supported work. Sources were excluded if they were purely technical architecture papers without relevance to work practices, highly general opinion pieces with little analytical value, or duplicates.

After duplicate removal, 86 records were screened by title and abstract; 25 were excluded. The remaining 61 full texts were assessed; 11 were excluded for limited relevance. The final corpus included 50 sources: 15 empirical studies, 7 theoretical or review-oriented sources, 8 practice-oriented or legal analysis sources, and 20 policy, institutional, or domain-specific guidance sources (Figure 1). Each source was coded for oversight,

competence, explanation, verification, governance, accountability, hallucination, AI type, and task risk; factors were grouped into four themes. Appendix D reproduces the coding table; the CSV in the supporting ZIP contains source-level data. Because this is a scoping review rather than an effectiveness review, sources were not excluded on the basis of formal quality appraisal; instead, evidence type and domain were coded to distinguish empirical studies from theoretical, practice-oriented, and policy sources.

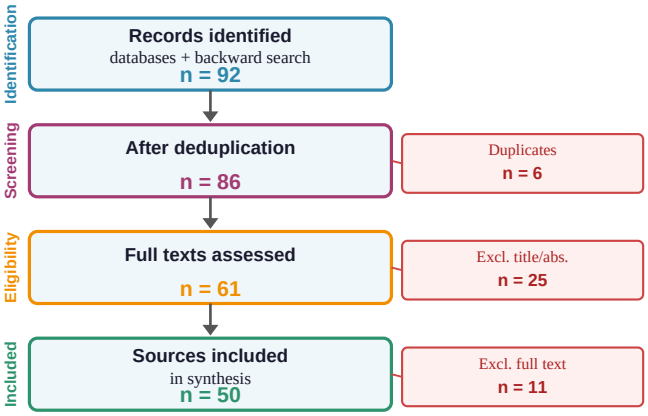

Figure 1: PRISMA-ScR screening flow.

3 Results

Across the 50 included sources, four dimensions repeatedly shaped whether trust was warranted: contextual trust conditions, intrinsic trustworthiness, trust calibration dynamics, and task stakes with human review. AI type emerged as a cross-cutting factor that changes verification and calibration demands. The four dimensions are analytically distinct: contextual conditions refer to institution-level safeguards, intrinsic trustworthiness to system-level reliability, calibration dynamics to user behaviour, and task stakes to the risk and verifiability of the specific drafting activity.

Source	Boolean string or search mode	Main yield
Google Scholar	"calibrated trust" AND "AI" AND ("worker" OR "professional")	General trust calibration
Google Scholar / ACM DL	"trust calibration" AND ("human-AI collaboration" OR "automation bias")	Human-AI collaboration evidence
Google Scholar / SSRN	"generative AI" AND ("legal drafting" OR "professional drafting") AND "trust"	Legal use-case literature
Google Scholar / SSRN	"hallucination" AND "AI" AND ("professional" OR "legal" OR "clinical")	Output-quality risks
PubMed	"AI" AND ("clinical decision support" OR "CDSS") AND ("trust" OR "reliance")	Clinical AI trust studies
Google Scholar / institutional	"AI" AND "architecture" AND ("professional practice" OR "trust")	Practice and policy guidance
ABA, CCBE, NCSC, UNESCO, EU, AIA, RIBA	Direct institutional search	Policy and institutional sources

Table 1: Main search strategy. Full log in Appendix A and CSV/readme.

Evidence category	N
Empirical studies and experiments	15
Theoretical or review-oriented sources	7
Practice-oriented or legal analyses	8
Policy, institutional, or domain guidance	20
Total included	**50**

Table 2: Composition of the final corpus.

3.1 Contextual Trust Conditions

Trust is shaped not only by the system itself but by its deployment context. Across governance and institutional literature [1, 38, 17, 20, 35], trust is more warranted when AI use is embedded in explicit rules, accountable workflows, visible human oversight, and documented deployment history. The EU AI Act treats human oversight as a core requirement for high-risk AI systems [35]. Analogous structures appear in hospitals supervising clinical AI, professional bodies regulating architectural practice, and courts or law firms governing AI-assisted legal work [30, 33, 34, 19, 21]. Context therefore provides trust-warranting conditions before users have direct experience with a system.

3.2 Intrinsic Trustworthiness

Warranted trust also depends on system competence, consistency, explanation quality, interface design, and output reliability under actual task conditions [5, 7, 8, 9]. For LLMs, the central risk is not poor prose but unsupported or fabricated content. Magesh et al. found that leading legal AI research tools hallucinated at rates high enough to undermine professional reliance [9]. For XAI-equipped systems, explanations can improve collaboration but only when they help users evaluate system limits rather than merely increasing confidence [8, 15, 26]. Interface design matters: source transparency and visible error patterns support appropriate reliance for less experienced professionals [16, 31, 29].

Evidence on explanation effects is not uniform. Naiseh et al. [5] found feature-importance explanations raised overtrust in clinical settings, while Senoner et al. [8] found XAI improved accuracy in manufacturing — a direct contradiction resolved only by attending to domain context and user expertise. Kim et al. [29] confirm this moderating role: source transparency improved appropriate LLM reliance specifically for users capable of evaluating source credibility, while Küper et al. [31] show that less experienced users benefit most from UI-level safeguards but are also most susceptible to automation bias when those safeguards are absent. Collectively, these studies suggest that explanation design cannot be treated as uniformly beneficial: its effect depends on the AI type, the user's expertise level, and the task's error-detection demands.

3.3 Trust Calibration Dynamics

The trust problem is dynamic rather than static. The goal is not to maximise trust but to calibrate reliance to actual capability [2, 22, 50]. Both overtrust and undertrust can reduce performance. Research also shows algorithm appreciation — a tendency to prefer algorithmic over human judgment that, when uncritical, is itself a form of miscalibration [49]. Self-reported scepticism does not reliably prevent behavioural reliance: Kennedy et al. found that legal professionals described themselves as cautious while

still incorporating algorithmic recommendations into decisions [4]. Ding et al. modelled confidence mismatch as a primary driver of suboptimal human-AI collaboration [6]. Evidence on cognitive forcing functions, source transparency, and deliberation pauses suggests that calibration can be improved by designs that slow automatic acceptance and make uncertainty visible [39, 24, 29]. This finding stands in partial tension with Bansal et al. [15], who showed explanations improve team performance — a contradiction resolved by noting that benefits emerge only when explanation complexity is matched to user cognitive capacity. When that match fails, the same explanations become the mechanism of automation bias rather than its remedy [26].

3.4 Task Stakes and Human Review

Trust should vary by task. Workers should not trust AI equally across routine, reversible drafting and high-stakes submissions. Stakes, reversibility, verifiability, confidentiality, and downstream consequences all matter [11, 12, 14]. Cai et al. showed that physicians relied more appropriately on imperfect AI tools when error patterns were visible and task stakes were explicit [16]. Legal and court-related guidance similarly treats human review as non-negotiable where client rights, confidentiality, or legal authorities are involved [17, 19, 20]. The more serious the consequences, the narrower the acceptable range of reliance.

3.5 AI Type and Cross-Domain Evidence

AI type is not treated as a fifth dimension; it operates as a cross-cutting modifier that changes the verification threshold within each dimension. LLMs require source and claim verification because fluent output may lack grounding [9, 29]. Narrow ML systems require scrutiny of confidence scores because overconfident scores can induce automation bias [8, 22, 50]. XAI systems require attention to explanation design because explanations may either support understanding or increase cognitive load [5, 26].

Medicine and architecture support the framework's potential transferability while showing that thresholds differ by profession. Clinical decision support studies show that trust depends on governance, diagnostic performance, automation bias, and task stakes [30, 31, 32]. Architecture reports and professional guidance similarly distinguish low-risk visualisation or concept drafting from higher-stakes compliance documentation and regulatory submissions [33, 34, 46]. These domains do not prove universal validity, but they support a domain-agnostic structure that must be adapted to each profession's accountability and verification conditions.

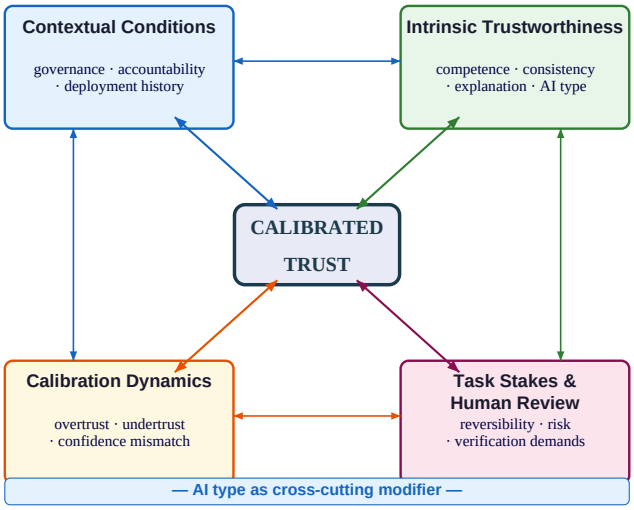

Figure 2: Relational calibrated trust framework. The dimensions interact bidirectionally rather than forming a fixed sequence.

Theme	Key conditions	Interaction effect
Contextual conditions	Governance, accountability, deployment history	Amplifies or weakens perceived trustworthiness
Intrinsic trust-worthiness	Competence, consistency, explanation quality, AI type	Sets the ceiling for appropriate reliance
Calibration dynamics	Overtrust, undertrust, confidence mismatch, cognitive load	Determines whether users rely in proportion to capability
Task stakes and review	Reversibility, risk, verification, downstream consequences	Narrows acceptable reliance as stakes rise

Table 3: Four-theme framework and interaction logic.

4 Worked Use Case: Legal Drafting

Legal drafting combines routine and high-stakes elements, making it a useful application of the framework. Tools such as Harvey, Lexis+ AI, and CoCounsel are used for drafting, legal research, clause comparison, and document review [11, 14, 41]. Because these tools are LLM-based or RAG-LLM hybrids, they require verification of citations, authorities, and factual claims [9, 13]. For RAG-LLM tools, verification should not stop at the presence of a cited source; professionals must also check whether the retrieved source is authoritative, relevant, and actually supports the generated claim.

The framework produces a differentiated judgment. Contextually, trust is more warranted when a law firm has clear AI-use policies, data-entry rules, confidentiality safeguards, and qualified human review [17, 20, 41]. Intrinsically, trust is more warranted for structural suggestions, clause paraphrasing, or preliminary summaries than for tasks requiring verified authorities or final legal reasoning. At the calibration level, junior lawyers may over-rely on polished AI drafts under time pressure, while senior lawyers may undertrust AI and reject efficiency gains [49]. This undertrust is also costly because it can block legitimate efficiency gains in low-risk tasks where AI support would remain reviewable and professionally accountable. Role-differentiated protocols are therefore needed: associates should label AI outputs as provisional, and reviewers should audit substance rather than reject AI assistance by default.

The verification-value paradox becomes practical here. AI assistance is likely to produce net value when the task is low-to-medium risk, errors are detectable through ordinary professional review, and verification time is bounded. It is likely to produce negative net value when verification effort approaches or exceeds the drafting time saved, or when the cost of an undetected error exceeds the cost of drafting without AI [14, 13]. Thus, AI is defensible as a limited support tool, not as an autonomous legal drafter.

AI type	Main trust risk	Calibration requirement
LLMs / RAG-LLMs	Fluent unsupported claims, hallucinated citations	Verify sources, claims, and authorities
Narrow ML systems	Overconfident scores, hidden error patterns	Compare confidence with observed accuracy
XAI-equipped systems	Explanations that persuade without improving understanding	Match explanation type to user expertise and task

Table 4: AI type distinctions and calibration requirements.

Domain	Lower-risk AI support	Higher-risk boundary
Law	Clause rephrasing, internal summaries	Court filings, legal advice, verified authorities
Medicine	Internal case notes, admin summaries	Diagnosis, patient-facing recommendations
Architecture	Concept descriptions, visualisation text	Code compliance, regulatory submissions

Table 5: Cross-domain illustration of the framework.

5 Discussion: Operationalising Calibrated Trust

The synthesis suggests that calibrated trust should be treated as a practical judgment rather than a general attitude toward AI. A professional does not simply decide whether an AI system is trustworthy in the abstract; the decision depends on the specific task, the system type, the institutional setting, and the available review process. This point responds directly to the strongest pattern in the reviewed literature: studies that evaluate only acceptance or perceived trust miss the central risk — whether reliance is proportionate to actual capability.

A first implication is that governance and technical reliability must be evaluated together. Policy guidance in law, courts, medicine, and architecture repeatedly requires human accountability, but accountability alone does not make an unreliable system safe. Conversely, good model performance does not justify informal use if workers lack rules for data entry, confidentiality, review, or escalation. The framework therefore rejects both technological determinism — where model quality is treated as sufficient — and purely procedural accounts — where supervision is assumed to solve all risks. For example, weak governance can turn even a technically reliable system into an overtrust risk when users lack verification norms, while strong governance can prevent fluent but unreliable outputs from being treated as authoritative.

A second implication concerns professional role. The same AI output can create different calibration problems for different workers. A junior professional may over-rely on fluent prose because it reduces time pressure and appears authoritative. A senior reviewer may underuse the same tool because professional experience makes them sceptical of automation [49]. Both responses can be miscalibrated. Training should therefore distinguish between first-draft use, verification responsibility, and final accountability rather than giving one generic instruction such as "always check AI output."

A third implication is that verification must be proportionate. For low-risk, reversible drafting support, a quick review may be sufficient. For high-stakes work, verification must be independent, documented, and performed by a qualified professional. The verification-value paradox arises when the checking required to make AI use responsible cancels the time saved by using AI. This does not mean AI has no value; it means that its value is task-dependent. Organisations should define task categories in advance and specify the minimum review standard for each category.

Finally, the cross-domain comparison shows why the framework is domain-agnostic in structure but not identical in application. A hallucinated legal citation, a misleading clinical summary, and an incorrect compliance statement are different failures, but all involve a mismatch between AI output, human reliance, and task stakes. The framework is therefore best understood as a reusable diagnostic structure rather than a universal checklist with fixed thresholds.

Task type	Examples	Warranted trust
Low-risk support	Alternative phrasings, clause reorganisation, internal brainstorming	Conditional yes, with review
Medium-risk drafting	Contract clause comparison, first-draft structures, case-file summaries	Only with independent verification
High-risk work	Court submissions, client-facing advice, regulatory filings, citation-dependent arguments	Not autonomous; full expert review required

Table 6: Task-stakes classification for generative AI in legal drafting.

Question	Interpretation
Can errors be detected with available expertise?	If no, AI use is not warranted for the task.
Is the task reversible or internal?	If yes, limited reliance is more defensible.
Is verification time bounded?	If no, AI may create negative net value.
Would an undetected error be professionally serious?	If yes, full expert review is mandatory.

Table 7: Decision criterion for the verification-value paradox.

6 Gaps and Future Work

Five gaps remain. First, task-level empirical evidence is scarce: internal brainstorming, clause comparison, and court submissions are rarely studied as distinct reliance decisions. Second, longitudinal data are limited, making it unclear whether initial calibration drifts toward routine overtrust over time [3, 7]. Third, oversight mechanisms are often assumed rather than compared: future work should test whether checklists, senior review, collaborative drafting, or forced deliberation produce better calibration. Fourth, individual differences in expertise, AI literacy, and professional role — including algorithm appreciation versus algorithm aversion [49] — require more attention. Future

work should also examine cultural, demographic, and jurisdictional differences in trust calibration, since professional reliance may vary across legal systems, organisational cultures, levels of AI literacy, and seniority. Fifth, AI type effects are currently inferred across separate literatures; matched studies comparing LLMs, ML classifiers, and XAI systems in equivalent professional tasks would substantially strengthen the evidence base. A directly testable hypothesis follows: professionals trained on AI-type-specific error signatures — LLM hallucination rates versus ML confidence miscalibration — will show lower automation bias in matched drafting tasks than those receiving generic AI literacy training, addressable through between-subject designs with professional cohorts [49, 50]. A second testable hypothesis is that structured verification checklists will reduce acceptance of hallucinated citations more than unstructured expert review in citation-dependent drafting tasks.

Source	Evidence / domain	Main factor extracted	Theme
Riegelsberger 2005; Mayer 1995	Theoretical / general	Trust warranted by institutional signals, accountability, history	Contextual conditions
ABA 2024; CCBE 2025; EU AI Act 2024	Policy / governance	Competence, confidentiality, human oversight, and override capacity required	Contextual conditions
Magesh 2025; Chang 2026	Audit / legal analysis	Legal AI outputs can contain unsupported or hallucinated authorities	Intrinsic trustworthiness
Naiseh 2023; Senoner 2024	Empirical / CDSS, manufacturing	Explanation type and confidence displays can improve or distort trust	Intrinsic trustworthiness
Bansal 2021; Stojanov 2025	CHI / review	Explanations help only when they reduce rather than increase cognitive burden	Calibration dynamics
Kennedy 2025; Ding 2025	Survey / model	Stated caution and actual reliance diverge; confidence mismatch drives errors	Calibration dynamics
Lee & See 2004; Logg 2019; Lucas 2024	Theory / review	Undertrust, overtrust, and algorithm appreciation are all miscalibration failures	Calibration dynamics
Cai 2019; Kücking 2024	Empirical / medicine	Visible error patterns and user expertise affect reliance under risk	Task stakes and review
Regalia 2024; Yuvaraj 2025	Practice / normative	Verification costs can outweigh drafting speed gains	Task stakes and review
RIBA 2025; AIA 2025	Architecture practice / policy	Trust is higher for concept drafting than compliance submissions	Cross-domain application

Table 8: Selected evidence-to-theme coding used to derive the framework.

7 Conclusion

This scoping review mapped the factors shaping well-calibrated trust in AI for drafting-related professional tasks across 50 sources. Four themes emerged: contextual trust conditions, intrinsic trustworthiness, trust calibration dynamics, and task stakes with human review. The review also highlights AI type as a cross-cutting factor: LLMs, narrow ML systems, and XAI-equipped tools require different verification and calibration practices.

The main contribution is conceptual and practical. Trust in AI-supported work cannot be reduced to model quality alone; appropriate reliance depends on the interaction between systems, users, institutions, and tasks. Applied to legal drafting, the framework shows that AI can legitimately support low-risk, reversible drafting, but cannot replace professional judgment in high-stakes work. The most defensible position is neither blind trust nor total refusal, but calibrated trust. Designing AI for professional drafting therefore requires not more trust, but better evidence about when reliance is justified.

References

[1] J. Riegelsberger, M. A. Sasse, and J. D. McCarthy. 2005. The mechanics of trust: A framework for research and design. *Int. J. Hum.-Comput. Stud.* 62, 3 (2005), 381–422.

[2] G. M. Lucas, B. Becerik-Gerber, and S. C. Roll. 2024. Calibrating workers' trust in intelligent automated systems. *Patterns* 5, 9 (2024), 101045.

[3] M. Liebherr, E. Enkel, P. Sieberg, et al. 2026. Dynamic calibration of trust and trustworthiness in AI-enabled systems. *Int. J. Softw. Tools Technol. Transfer* (2026).

[4] R. Kennedy, L. Tiede, A. Austin, and K. Ismael. 2025. Law enforcement and legal professionals' trust in algorithms. *J. Law Empir. Anal.* (2025).

[5] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *Int. J. Hum.-Comput. Stud.* 169 (2023), 102941.

[6] S. Ding et al. 2025. A new model for calculating human trust behavior during human-AI collaboration. Preprint (2025).

[7] A. A. Tutul, E. H. Nirjhar, and T. Chaspari. 2025. Investigating trust in human-AI collaboration for a speech annotation task. *Int. J. Hum.-Comput. Interact.* 41, 5 (2025).

[8] J. Senoner, S. Schallmoser, B. Kratzwald, S. Feuerriegel, and T. Netland. 2024. Explainable artificial intelligence improves human-AI collaboration. *Sci. Rep.* 14 (2024).

[9] V. Magesh, F. Surani, M. Dahl, M. Kazemi, P. Liang, and D. E. Ho. 2025. Hallucination-free? Assessing the reliability of leading AI legal research tools. Preprint. Stanford HAI

[10] G. Fragiadakis, C. Diou, G. Kousiouris, and M. Nikolaidou. 2024. Evaluating human-AI collaboration: A review and methodological framework. arXiv:2407.19098. https://arxiv.org/abs/2407.19098

[11] J. Regalia. 2024. From briefs to bytes: How generative AI is transforming legal writing and practice. *Tulsa Law Rev.* 59, 2 (2024), 193–236.

[12] J. R. Gunder. 2024. Rule 11 is no match for generative AI. Preprint (2024).

[13] C. K. Chang. 2026. Hallucinated authority: AI citations as reckless misrepresentation. SSRN preprint.

[14] J. Yuvaraj. 2025. The verification-value paradox: A normative critique of Gen AI in legal practice. Preprint (2025).

[15] G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proc. CHI 2021*. ACM. https://doi.org/10.1145/3411764.3445717

[16] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proc. CHI 2019*. ACM.

[17] American Bar Association. 2024. *Formal Opinion 512: Generative Artificial Intelligence Tools*. ABA. ABA Op. 512

[18] American Bar Association. 2024. ABA ethics opinion on generative AI offers useful framework. ABA.

[19] National Center for State Courts. 2025. *Key Considerations for the Use of Generative AI Tools in Courts*. NCSC. ncsc.org

[20] Council of Bars and Law Societies of Europe. 2025. *Guide on the Use of Generative AI for Lawyers*. CCBE. ccbe.eu

[21] UNESCO. 2025. *Guidelines for the Use of AI Systems in Courts and Tribunals*. UNESCO. unesco.org

[22] J. D. Lee and K. A. See. 2004. Trust in automation: Designing for appropriate reliance. *Hum. Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[23] A. C. Tricco, E. Lillie, W. Zarin, et al. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* 169, 7 (2018), 467–473. https://doi.org/10.7326/M18-0850

[24] T. Srinivasan and J. Thomason. 2025. Adjust for trust: Mitigating trust-induced inappropriate reliance on AI assistance. arXiv:2502.13321. https://arxiv.org/abs/2502.13321

[25] F. Kuecking, U. Huebner, M. Przysucha, et al. 2024. Automation bias in AI-decision support: Results from an empirical study. *Stud. Health Technol. Inform.* 317 (2024), 298–304. https://doi.org/10.3233/SHTI240871

[26] A. Stojanov et al. 2025. Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & Soc.* (2025). https://doi.org/10.1007/s00146-025-02422-7

[27] L. Roeder, P. Hoyte, J. van der Meer, et al. 2023. A quantum model of trust calibration in human–AI interactions. *Entropy* 25, 9 (2023), 1362. https://doi.org/10.3390/e25091362

[28] R. Visser, T. M. Peters, I. Scharlau, and B. Hammer. 2025. Trust, distrust, and appropriate reliance in (X)AI: A conceptual clarification. *Cogn. Syst. Res.* (2025), 101357.

[29] S. S. Y. Kim, J. W. Vaughan, Q. V. Liao, T. Lombrozo, and O. Russakovsky. 2025. Fostering appropriate reliance on large language models. In *Proc. CHI 2025*. ACM. https://doi.org/10.1145/3706598.3714020

[30] H. El-Haddad et al. 2025. Trust in artificial intelligence–based clinical decision support systems among health care workers: Systematic review. *J. Med. Internet Res.* 27 (2025), e69678. https://doi.org/10.2196/69678

[31] A. Kueper, G. C. Lodde, E. Livingstone, D. Schadendorf, and N. Kraemer. 2025. Psychological factors influencing appropriate reliance on AI-enabled clinical decision support systems. *J. Med. Internet Res.* 27 (2025), e58660. https://doi.org/10.2196/58660

[32] C. Onuoha et al. 2023. Artificial intelligence and clinical decision support: Clinicians' perspectives on trust, trustworthiness, and liability. *BMJ Open* (2023). https://pmc.ncbi.nlm.nih.gov/articles/PMC10681355/

[33] Royal Institute of British Architects. 2025. *RIBA AI Report 2025*. RIBA. riba.org

[34] American Institute of Architects. 2025. Artificial Intelligence Policy Resolution. AIA. aia.org

[35] European Parliament and Council of the EU. 2024. Regulation (EU) 2024/1689: Artificial Intelligence Act. Official Journal of the EU. https://artificialintelligenceact.eu/article/14/

[36] D. Warmsley, K. Choudhary, J. Rego, E. Viani, and P. K. Pilly. 2025. Self-assessment in machines boosts human trust. *Front. Robot. AI* 12 (2025), 1557075. https://doi.org/10.3389/frobt.2025.1557075

[37] Chaos Group. 2026. AI in Architecture: Trends, Hidden Risks, and What Comes Next. Industry research report. blog.chaos.com

[38] R. C. Mayer, J. H. Davis, and F. D. Schoorman. 1995. An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 3 (1995), 709–734.

[39] Z. Bucinca, M. B. Malaya, and K. Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI. In *Proc. CHI 2021*. ACM. https://doi.org/10.1145/3411764.3445780

[40] M. C. Horowitz and L. Kahn. 2023. Bending the automation bias curve: A study of human and AI-based decision making in national security contexts. arXiv:2306.16507. https://arxiv.org/abs/2306.16507

[41] Law Society of England and Wales. 2025. *Generative AI: The Essentials*. Law Society. lawsociety.org.uk

[42] M. D. J. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares. 2015. Guidance for conducting systematic scoping reviews. *Int. J. Evid. Based Healthc.* 13, 3 (2015), 141–146.

[43] I. A. Qazi, A. Ali, A. U. Khawaja, et al. 2025. Automation bias in LLM assisted diagnostic reasoning among AI-trained physicians: A randomized clinical trial. medRxiv. https://doi.org/10.1101/2025.08.23.25334280

[44] McKinsey & Company. 2024. Building Trust in AI: The Key Role of Explainability. *McKinsey Insights*. mckinsey.com

[45] National Telecommunications and Information Administration. 2024. *AI Accountability Policy Report*. NTIA. ntia.gov

[46] VirtualSpaces. 2026. AI in Architecture: Closing the 85% Exposure Gap. Industry analysis. virtualspaces.tech

[47] ABA Task Force on Law and Artificial Intelligence. 2025. Year 2 Report on the Impact of AI on the Practice of Law. ABA. americanbar.org

[48] M. Pal, H. N. Saha, and A. Chakrabarti. 2026. The Trust-Aware XAI (TAXAI) framework. *Sci. Rep.* (2026). https://doi.org/10.1038/s41598-026-44167-3

[49] J. M. Logg, J. A. Minson, and D. A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151 (2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[50] R. Parasuraman and V. Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* 39, 2 (1997), 230–253. https://doi.org/10.1518/001872097778543886

A Search Log

#	Person	Database	Boolean Search String	Notes
1	Cemre	Google Scholar	"calibrated trust" AND "AI" AND ("worker" OR "professional")	Foundational calibration papers
2	Cemre	Google Scholar	"trust in automation" AND "professional settings"	Automation bias; Lee and See
3	Cemre	ACM DL	"XAI" AND "human-AI collaboration" AND "decision making"	Explanation effects on trust
4	Meric	Google Scholar	"generative AI" AND ("legal drafting" OR "professional drafting") AND "trust"	Legal and professional drafting
5	Meric	Google Scholar	"hallucination" AND "AI" AND ("professional" OR "legal" OR "clinical")	AI reliability in professional work
6	Meric	SSRN	"professional responsibility" AND "artificial intelligence"	Practice and policy documents
7	Cemre	Springer	"human oversight" AND "generative AI" AND "governance"	Governance frameworks
8	Cemre	ACM DL	"explanation effects" AND "trust calibration"	Clinical decision support
9	Meric	ABA / CCBE	Direct institutional search	Bar association ethics guidance
10	Cemre	Google Scholar	"overtrust" AND ("automation bias" OR "AI") AND "professional work"	Backward citation search
11	Cemre	PubMed	"AI" AND ("clinical decision support" OR "CDSS") AND ("trust" OR "reliance")	Clinical AI trust supplement
12	Meric	Google Scholar	"AI" AND "architecture" AND ("professional practice" OR "trust")	Architecture domain evidence
13	Cemre	ACM DL	"appropriate reliance" AND ("large language model" OR "LLM")	LLM-specific reliance
14	Meric	Institutional websites	AIA, RIBA, EU AI Act direct search	Architecture and regulation policy

Table 9: Full search log with Boolean operators, databases, and scope notes.

B Inclusion and Exclusion Criteria

Inclusion criteria. Sources were included if they addressed at least one of: trust in AI systems; trust calibration or appropriate reliance; human-AI collaboration; AI-supported professional drafting; human oversight of AI; AI hallucination risk; reliability of AI outputs; AI type distinctions; or professional responsibility in AI-supported work in any domain. Theoretical frameworks, empirical studies, systematic reviews, and institutional or policy guidance documents were all eligible.

Exclusion criteria. Sources were excluded if they were purely technical model architecture papers with no relevance to work practices, highly general opinion pieces with little analytical value, or duplicate materials. Publication date was not used as an exclusion criterion; pre-2015 sources were retained where foundational.

C PRISMA-ScR Screening Table

Stage	Action	N
Identification	Records identified through database searches and backward citation searching	92
Screening	Records after duplicate removal	86
Screening	Records excluded after title/abstract screening	25
Eligibility	Full texts assessed for relevance to primary review question	61
Eligibility	Full texts excluded for limited relevance	11
Included	Final sample: 15 empirical studies, 7 theoretical/review, 8 practice-oriented/legal analysis, 20 policy/institutional/domain guidance	50

Table 10: PRISMA-ScR screening table corresponding to Figure 1.

D Coding Scheme and Factor Table

Authors / Year	Key factor(s)	Type	Theme
Riegelsberger et al. 2005	Governance, accountability, deployment history as trust-warranting conditions	Theoretical	Contextual
Lucas et al. 2024	Overtrust and undertrust as measurable calibration states	Review/empirical	Calibration
Liebherr et al. 2026	Dynamic recalibration over repeated interaction	Theoretical	Calibration
Kennedy et al. 2025	Gap between stated caution and enacted reliance	Survey + case	Calibration
Naiseh et al. 2023	Explanation type systematically affects overtrust rate	RCT	Intrinsic
Ding et al. 2025	Confidence mismatch as driver of suboptimal outcomes	Simulation	Calibration
Tutul et al. 2025	Trust shaped by observed error patterns over time	Longitudinal	Intrinsic
Senoner et al. 2024	XAI improves accuracy; elevated confidence raises overtrust risk	Controlled experiment	Intrinsic
Magesh et al. 2025	Hallucination in legal tools undermines professional reliance	Audit	Intrinsic
Bansal et al. 2021	Explanations aid team performance only when confidence matched	CHI experiment	Calibration

Authors / Year	Key factor(s)	Type	Theme
Cai et al. 2019	Visible error patterns and explicit stakes reduce overtrust	CHI experiment	Task stakes
Yuvaraj 2025	Verification-value paradox undermines efficiency gains	Normative critique	Task stakes
ABA 2024	Competence, confidentiality, supervision as institutional conditions	Policy	Contextual
CCBE 2025	Accountability and confidentiality frame AI delegation limits	Policy	Contextual
EU AI Act 2024	Human oversight requirements for high-risk AI systems	Regulation	Contextual
RIBA 2025	Trust drops in architecture as task stakes rise	Policy/industry	Task stakes
Lee & See 2004	Foundational model: disuse and misuse as dual failures	Theoretical	Calibration
Logg et al. 2019	Algorithm appreciation as uncritical preference for algorithmic judgment	Empirical	Calibration
Parasuraman & Riley 1997	Use, misuse, disuse, abuse taxonomy for human-automation interaction	Theoretical	Calibration

Table 11: Coding table for selected included sources. Full source-level coding is provided in coding_data.csv.

E AI Use Disclosure

ChatGPT and Claude were used during the drafting process for brainstorming, outlining, language refinement, and organisational support. The authors verified all substantive claims against the cited sources, reviewed and revised the final text themselves, and confirmed that cited sources are relevant to the arguments they support. AI tool use was governed by the same verification principles advocated in this review: AI outputs were treated as provisional drafts subject to human review, not as authoritative final text.

F Team Information and Supporting Materials

Team name: **TwoGirlsTooLate.**
Cemre Ozcan – s334701@studenti.polito.it – Politecnico di Torino.
Meric Ozler – s336819@studenti.polito.it – Politecnico di Torino.

Supporting materials are available in the ZIP archive linked here: Supporting Materials ZIP Archive. The archive contains **coding_data.csv**, **readme.txt**, **TwoGirlsTooLate_FinalPaper.pdf**, and **source_list.txt**.