IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4637425/
Vol. 2025 (34), Article ID 4637425

Antecedents of Trust in Artificial Intelligence: A Literature Review*

Peter RITTGEN

University of Borås, Borås, Sweden

Correspondence should be addressed to: Peter RITTGEN, peter.rittgen@hb.se

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

The rise of Artificial Intelligence (AI) in information systems offers a host of new opportunities but also substantial challenges. In 2019 the EU commission released the "Ethical Guidelines for Trustworthy AI". So far, these guidelines are only recommendations and not enforced by law. Six years after their release it is still unclear how these guidelines can be put into practice. Maturity with respect to the ethical use of AI is still limited in organizations so a deeper look into the status quo of research in this area seems imperative. We started by doing an exploratory literature study to find the most relevant antecedents determining trust in AI. We then proceeded by performing an in-depth literature review on how these antecedents, and thereby ultimately trust, can be achieved. Our results indicate that trust is an inherently complex issue that is still poorly understood.

Keywords: Ethics in Artificial Intelligence, trust, antecedents

Introduction

The importance of ethical issues in Artificial Intelligence (AI) has recently been recognized by many regulatory bodies and, consequently, the EU commission has released the "Ethical Guidelines for Trustworthy AI" in 2019 (European Commission, 2019). This is a notable progress because the EU commission has a regulatory status for the 26 member states and therefore for most of the European countries. Despite this fact the guidelines are only recommendations so far.

Studies on the compliance of companies with these guidelines show that awareness exists but adoption is slow (Arbelaez Ossa et al., 2024) and enforcement is lacking (Kijewski et al., 2024). Especially SMEs lag considerably (Soudi & Bauters, 2024). It is therefore necessary to look at the relevant literature to identify requirements for the implementation of ethical AI.

Methodology

As the purpose of this study is to investigate the status quo of research concerning ethical AI, we conducted a literature review (Webster & Watson, 2002). The study is concept-centric with the purpose of developing a concept matrix showing the relevant concepts, i.e. ethical criteria. For each criterion, we take the number of publications dedicated to this criterion as an indicator for its relative relevance.

Cite this Article as: Peter RITTGEN, Vol. 2025 (34) "Antecedents of Trust in Artificial Intelligence: A Literature Review " Communications of International Proceedings, Vol. 2025 (34), Article ID 4637425, https://doi.org/10.5171/2025.4637425

The survey is based on the Scopus database, which is commonly used in IS. To ensure currency of the literature we restricted the survey to papers published between 2020 and 2025. To ensure the quality of papers we restricted our search to peer-reviewed journal papers. The field of the considered research papers was restricted to those that deal with machine learning, which is the primary method for establishing AI leading to a clear scope for the considered papers.

Relevant Criteria for Establishing Ethical AI

A study (Ahmad et al., 2022) investigated more than 300 research papers arriving at the following list of relevant criteria (key values) for ethical AI: confidentiality, privacy, accountability, fairness, justice, transparency, and trust. Table 1 shows the results with percentages relative to the number of "ethical" papers, except the ethical percentage that is relative to the total number of publications in the field.

Table 1. Relevant criteria for ethical AI (own table)

Key values	Publications	Percentage of ethics
Confidentiality	1439	15 %
Privacy	**7321**	**78 %**
Accountability	1227	13 %
Fairness	1693	18 %
Justice	1677	18 %
Transparency	3164	34 %
Trust	**5759**	**62 %**

Among the papers addressing ethical issues, privacy and trust seem to be the ones considered to be most relevant. The issue of privacy has already been identified as highly relevant by legislation so that the EU introduced a respective law, the General Data Protection Regulation (GDPR) that was adopted on 14 April 2016 and became enforceable 25 May 2018.

Concerning trust, no such regulation has as yet been established. We therefore concluded that it is necessary to consider the state-of-art in current research concerning this issue in the following.

Antecedents of Trust

If trust is the most relevant ethical issue not yet covered by legislation, we need to investigate how trust in AI can be established, i.e. what are the antecedents or prerequisites creating trust in AI systems. It is obvious that systems that are not trustworthy will not be used at length by their users. Trust should therefore be an issue that is relevant for developers of such systems as well as for legislators.

According to (Mayer, Davis & Schoorman, 1995) we define trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." A recent large scoping study (Ueno et al. 2022) found the factors shown in figure 1.

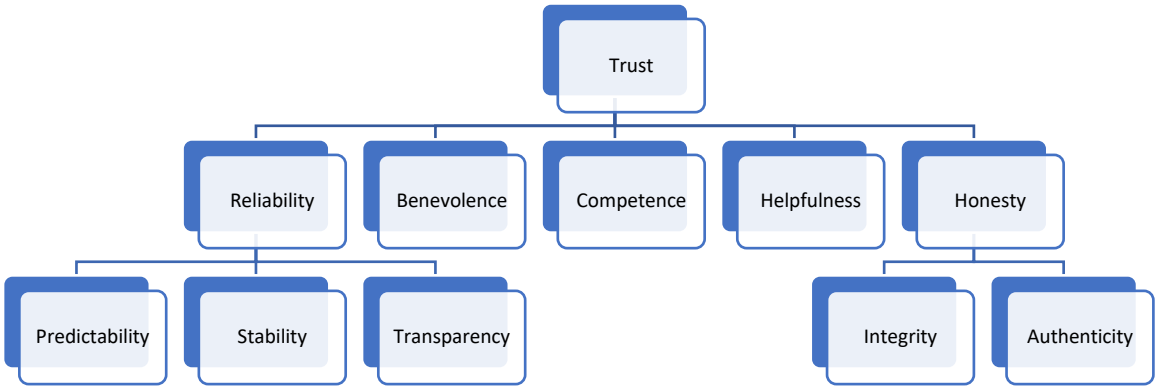

Fig. 1. Antecedents of trust (own figure)

Analysis of the Antecedents of Trust

Our analysis focuses on literature that addresses the antecedents mentioned in Fig. 1. The goal is to address the core issues that need to be addressed to achieve each antecedent and thereby ultimately achieve trust in Artificial Intelligence. Our analysis will take up the leaves of the taxonomy as important precursors of trust.

Predictability

Predictability of AI systems is not a simple concept but consists of several prerequisites that must be met, such as the size of the model (scaling), the quality of the training data and the estimation of confidence in the predictions. It also involves robustness towards changes in the environment.

There is empirical evidence showing that predictability in certain areas is quite stable, such as in machine translation (Gordon et al., 2021). A survey by (Wang et al., 2023) shows that predictive reliability is recognized by many researchers as a central issue and that, consequentially, research in this area is increasing.

Nevertheless, several important issues remain. Firstly, qualitative changes in behaviour present a major challenge to predictability. A study by (Liu et al., 2024), for example, shows that predictive quality depends heavily on the level of quantisation of large language models, and thereby on scale. But (Minderer et al., 2021) finds that quality in prediction depends more on architecture than on scale. Calibration of AI models alone cannot ensure high predictability. More advanced architectures even tend towards lower degrees of predictability. There seems to be no silver bullet for it.

As already mentioned, robustness is a closely related issue. While scaling rules help in predicting common behaviour, they usually perform badly in changing environments. There is a huge gap between a system's performance in average and worst cases.

Stability

Stability refers to the consistency and reliability of an AI system's performance when subject to variations in input, training, or the environment. A stable AI model will deliver similar results when given similar inputs. It will not drastically change with respect to its internal behaviour and will maintain its performance under reasonable changes in the environment.

The concept of stability is increasingly recognised as critical for trustworthiness, monitoring, reproducibility, and operational deployment of AI systems. Empirical studies show that instability even arises when average accuracy is sufficient. For example, a study of model robustness to changes in dataset (Subbaswamy, Adams & Saria, 2021) shows that stability degrades under distribution shifts, rendering it useless in changing environments. Similar

results were obtained by (Lei & Ying, 2020), showing that even small changes in training data can substantially deteriorate system performance.

Some research suggests that ensemble-based models might offer better stability. E.g. (Soloff, Foygel Barber & Willett, 2024) have found that the introduction of bagging methods can reduce variance and improve stability across datasets under varying distributions. Stability in highly uncertain environments such as stock markets has not been addressed, though.

Recent research has therefore turned attention to explicitly training models for stability. (Wang et al., 2023) introduces a so-called attention-based architecture where stability is introduced as a specific training parameter. This kind of research is still immature so it cannot be considered the ultimate solution yet.

In summary, stability is still not fully understood. Many studies focus only on average performance without considering systems' behaviour under realistic conditions of high variability. The factors behind instability are often very complex and not well-studied. Even under relatively controlled circumstances, stability is not trivial. More research is therefore needed on stability, especially under realistic conditions in changing environments.

Transparency

Transparency means the degree to which the internal reasoning and decision-making of an AI system can be made visible and communicated to users in an understandable way. A transparent system allows users to see how inputs lead to outputs, and which assumptions underly the model, and how limitations or risks can be assessed (Liu, 2021).

Transparency is considered as a vital part of trustworthy AI, allowing for accountability and governance, essential components of ethical AI. It reduces uncertainty and increases trust in AI systems. The study by Liu (2021) found that providing meaningful explanations of a a decision reduces the perceived uncertainty of users and improves their trust. But transparency extends beyond mere visibility and explainability of results. According to Haresamudram, Larsson and Heintz (2022) it can be studied on three levels: algorithmic transparency (how the system operates), interaction transparency (how the system communicates its results), and social transparency (how the system deals with the societal context).

Larsson et al. (2023) extended this model by adding four facets of AI transparency, which they call: explainability, mediation, literacy, and governance. They argue that true transparency requires not only technical clarity but also cultural and institutional mechanisms for interpretation. In the medical domain, they found significant gaps in transparency, where most studies did not fully disclose details about training data, validation methods, and model parameters. The same holds for many other application domains.

Transparency is also compromised by the fact that full disclosure would violate privacy, intellectual property or organizational secrets. It also needs to be adapted to the recipient of the AI output: what is meaningful to a software developer might be irrelevant or incomprehensible to a businessman, and vice versa. Perceived transparency, and thereby trust and acceptance, are also determined by users' perceptions of a system's competence and reliability (Al-Sulaiti et al., 2023).

In summary, transparency is not yet fully understood as it requires the treatment of issues on many different levels with respect to both documentation, communication, traceability, and auditability (Larsson et al., 2023). Further research in these areas is therefore required as trust in AI stands and falls with the essential concept of transparency.

Benevolence

Benevolence implies the perceived or actual intention of an AI system to act in the interests of its users, not promoting the goals of its designers at their expense. The concept comes originally from trust and leadership theory, where it is seen as a necessary property of a good leader, but can also be applied in the context of human-computer interaction: an AI, even if it is seen as capable and honest, may still be distrusted if it is not perceived as benevolent. Benevolence is therefore essential for trust in adopting AI-support by human users. Recent work shows that humans in an organizational setting tend to trust "AI managers" less than human managers as they assume AI managers to be less benevolent, even when the system is perceived as capable (Li & Bitterly, 2024).

Similar studies in the public sector also found that people tend to trust AI's ability more than its benevolence. Cooperative game experiments have shown that even humans tend to treat their AI partners less favourably than human partners, even if the AI partners were programmed to act in the humans' interest. So, benevolence alone does not ensure fair cooperation (Novozhilova et al., 2024; Karpus et al., 2021).

It is therefore not sufficient that AI agents "behave well" and do not harm humans, but they also need to be perceived as being benevolent, i.e. being in line with the goals of their users, and being transparent, accountable and trustworthy. Otherwise, users will suspect malevolence, misuse, or a hidden agenda, undermining trust in such agents.

In short, benevolence is still an unsolved issue in many ways. Users often treat AI agents as if they were human agents, attributing falsely both benevolence and malevolence to them even if the agents do not have such intentions. Users' perceptions, and not the agents' behaviour, determine whether users trust them. What determines these perceptions, though, seems often arbitrary and is not well understood. Further research in this area is therefore quintessential if we want to build trustworthy AI systems.

Competence

In the context of trust in AI, competence refers to the actual, or perceived ability of an AI system to perform tasks effectively, accurately, and reliably. Competence is one of the foundational dimensions of trustworthiness, together with integrity and benevolence. These factors determine whether users are willing to rely on the system (McKee, Bai, & Fiske, 2023). When people see AI as competent, they also tend to attribute more authority to it; conversely, perceived incompetence leads users to avoid the system and to critically check its output.

Recent research highlights that competence is a complex issue as it has both a technical and a social dimension. McKee et al. (2023), for example, show that people can attribute both warmth and competence to AI systems, in the same way as they would do to humans. Their findings show that highly autonomous AI agents are viewed as more competent but less warm, i.e. there is a trade-off between efficiency and perceived empathy. In the workplace domain, Nielsen (2024) found that proactive AI systems, i.e. systems that aid rather than just responding to user input, can, somewhat unexpectedly, even reduce users' self-perceived competence, causing lower satisfaction and engagement. This shows that perceptions of competence in AI systems can negatively affect users' own sense of ability and control.

Competence also involves the user's ability to understand and utilize AI effectively. A recent study among Latvian public-sector employees reveals that users' knowledge of AI systems and their confidence and literacy in using them, is positively correlated with trust in AI-generated outputs (Lāma & Lastovska, 2025). Thus, competence needs to be evaluated both at the system level and the user level.

Ultimately, competence can contribute towards moral judgments about AI. Nevertheless, users often rate AI systems as being more competent than moral (Oliveira et al., 2024). They usually do not grant an AI system the ability of behaving ethically. So, even systems that are observed as competent are still not trusted morally. This raises the question whether AI systems can be moral at all, i.e. possess conscience in the way human beings do. This issue, though vital for trust in AI, is still not understood.

Helpfulness

By helpfulness we mean the degree to which an AI system assists users in completing tasks, solving problems, or achieving goals in a way that is effective, timely, and appropriate in its context. A system that is perceived as helpful will not only deliver correct output but also support user in a suitable way.

Research shows that perceived helpfulness increases user engagement, trust, and adoption. For instance, a randomized controlled trial found that pharmacists using an AI prototype for medication verification observed the AI-generated content to a larger extent where such content was available and accurate, but that inappropriate AI advice led to increased cognitive processing time and reduced workflow efficiency (Tsai et al., 2025). Helpful AI can therefore streamline human tasks, whereas unhelpful AI may hinder performance.

But even helpfulness has an ethical dimension: a recent study on AI alignment showed that AI systems also need to be seen as harmless and honest. A system that is only "helpful" can still lead to unwanted outcomes if the

context is misunderstood (Dahlgren Lindström, Methnani, Krause, & Ericson, 2025). In other words, helpfulness alone does not guarantee positive outcomes.

A closely related factor is that of perceived usefulness. Research on generative AI tools showed that perceived usefulness was a significant predictor of users' intention to use the system (Kim & Park, 2025). Therefore, designing AI that is genuinely helpful requires not only functional performance but also good communication of value. All in all, even this issue is not fully understood and requires further research.

Integrity

Integrity refers to the principle that AI systems should be developed and used in ways that are honest and consistent with ethical norms, and accountable for their outcomes. An AI system operating with integrity respects data provenance, maintains fairness, avoids misleading users, preserves trust, and aligns with societal values.

Research shows the critical role of integrity in AI. For example, (Abujaber & Nashwan, 2024) emphasises the foundational value of integrity arguing that applications must ensure methodological soundness, responsible data use, and disclosure of bias or limitations. Even in academic publishing such concerns are becoming increasingly relevant: AI-assisted tools can today automatically generate research papers, which raises the question of authorship, and undermines scientific credibility. A study by Pellegrina and Helmy (2025) discusses the use of AI for detecting errors and misconduct in scientific papers, pointing to integrity as essential for maintaining the credibility of research. Integrity is also relevant in many other domains calling for governance frameworks built on integrity (Limongi, 2024).

Integrity in AI is important for several reasons: users need to be confident that AI systems behave reliably, ethically, and honestly without hidden agendas or undue bias. They also expect that such systems support transparency, auditability, and accountability (Limongi, 2024). Finally, systems lacking integrity may unintentionally drift into harmful behaviour, misleading users, or enabling misuse.

Nevertheless, integrity is still an open issue. It demands more than just regulatory compliance: it requires also that ethical values, organisational culture, and continuous oversight are built into all areas of AI. For instance, reliable detection of misuse of AI in publishing is still immature (Pellegrina & Helmy, 2025). Even in healthcare integrity needs to go beyond algorithms, also covering data provenance, user consent and reporting of limitations (Abujaber & Nashwan, 2024). Without systematic design for integrity, high performance alone will not uncover ethical flaws.

Authenticity

Authenticity refers to the capacity of an AI system to be perceived as genuine, true to its source, and aligned with human expectations, as opposed to being perceived as deceptive, manipulated, or misrepresentative. It relates both to how the system behaves and to how the users interpret its behaviour in relation to how humans would behave in a similar situation.

Authenticity is relevant because it establishes trust and user involvement, and ultimately user acceptance of AI systems. For example, in a study of chatbots in Chinese e-commerce, perceived authenticity of the chatbot significantly predicted user acceptance and perceived usefulness (Marjerison, et al., 2025). In the area of culture and creativity, the concept of "synthetic authenticity" has emerged: people interacting with AI-generated creative content expressed higher trust when the content followed traditional aesthetic principles, transparent disclosure and user involvement, i.e. authenticity is mediated by design and context (Wang & Adzharuddin, 2025).

Nevertheless, achieving authenticity in AI is still not sufficiently understood. Some researchers have applied authentication theory, stating that authenticity, rather than being a fixed attribute, is established by users engaging in "negotiation" to assess whether an artifact is authentic or not, based on signs of origin, intention, and context (Farooq & de Vreese, 2025). This means that authentic AI systems must not only perform well, but show their origins, align with human norms, and be transparent about their generation.

Authenticity is also one of the core mechanisms in responsible AI systems, together with control and transparency. Rivera, et al. (2025) put forward an "Authenticity-Control-Transparency" theory of responsible AI, where authenticity mechanisms ensure that AI architectures, algorithms and affordances reflect ethical design decisions

and human-centred values. Without addressing authenticity, AI systems may be effective but still appear untrustworthy or alien.

In practice, many AI systems fail to offer authenticity, because they hide their AI-generation, try to imitate humans without admitting so, or producing content that is just “too perfect” and therefore unnatural. They risk undermining user trust and legitimacy. In summary, authenticity is a key antecedent for trust in AI systems that designers and policymakers should treat seriously. Further research is required, though, to lay a foundation for this.

Conclusions

A review of contemporary research on AI ethics and human-AI interaction in the 2020ies shows that trust in AI depends on a large web of interrelated factors (antecedents or prerequisites) the most relevant being: predictability, stability, transparency, benevolence, competence, helpfulness, integrity, and authenticity. Together, these form the psychological and technical foundations of users’ willingness to trust, adopt, and collaborate with AI systems.

Predictability concerns the extent to which AI behaviour is consistent and intelligible over time, supporting user control and risk management. Stability complements this by emphasizing robustness under changing inputs and contexts; instability erodes trust and safety. Transparency ensures that the reasoning, data, and limitations behind AI decisions are accessible and interpretable, forming the basis of accountability and informed oversight.

Moving from functional to relational dimensions, benevolence captures users’ perception that an AI acts in their best interest, aligning its goals with human welfare. Competence refers to the system’s ability to perform effectively and reliably, while helpfulness extends competence into active user support, ensuring that AI meaningfully advances human goals.

With respect to ethics, integrity means honesty, consistency, and adherence to moral and procedural standards, particularly in sensitive contexts such as healthcare and research. Finally, authenticity implies that the AI-generated output is genuine and transparent, i.e. that users perceive it as sincere, traceable, and aligned with human values.

Together, these attributes suggest that trustworthy AI is not defined by a few simple factors alone but by an intricate web of factors that are both highly inter-related and subject to their own respective sub-factors. The current paper tries to investigate some of the complexity of these factors.

References

- Abujaber, A. A., & Nashwan, A. J. (2024). Ethical framework for artificial intelligence in healthcare research: A path to integrity. World Journal of Methodology, 14(3), 94071.

- Ahmad, K. Maabreh, M., Ghaly, M., Khan, K., Qadir, J., Al-Fuqaha, A. (2022). Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges. Computer Science Review, 43, 2022, 100452.

- Al-Sulaiti, G., Sadeghi, M. A., Chauhan, L., Lucas, J., Chawla, S., Elmagarmid, A. (2023). A Pragmatic Perspective on AI Transparency at Workplace. AI & Ethics, 4, 189-200.

- Arbelaez Ossa, L., Lorenzini, G., Milford, S.R., Shaw, D., Elger, B. S., Rost, M. (2024). Integrating ethics in AI development: a qualitative study. BMC Med Ethics 25, 10.

- Beckers S. (2018) AAAI: An Argument Against Artificial Intelligence. In: Müller V. (eds) Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 44. Springer, Cham

- Dahlgren Lindström, A., Methnani, L., Krause, L., Ericson, P., de Rituerto de Troya, Í., Coelho Mollo, D., & Dobbe, R. (2025). Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through reinforcement learning from human feedback. Ethics and Information Technology, 27(2), 28.

- European Commission (2019). *Ethics Guidelines for Trustworthy AI*. Directorate-General for Communications Networks, Content and Technology, High-Level Expert Group on Artificial Intelligence, EU Publications Office, Luxembourg. https://data.europa.eu/doi/10.2759/346720

- Farooq, A., & de Vreese, C. (2025). Deciphering authenticity in the age of AI: how AI-generated disinformation images and AI detection tools influence judgements of authenticity. AI & Society, Springer, 2025.

- Gordon, M., Duh, K., & Kaplan, J. (2021). Data and Parameter Scaling Laws for Neural Machine Translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing 2021, pp. 5915–5922. Association for Computational Linguistics.

- Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three Levels of AI Transparency. Computer, 56 (2), 93-100.

- Karpus, J., Krüger, A., Verba, J. T., Bahrami, B. Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. iScience, Volume 24, Issue 6, 25 June 2021.

- Kijewski, S., Ronchi, E. & Vayena, E. The rise of checkbox AI ethics: a review. AI Ethics 5, 1931–1940 (2025).

- Kim, J., & Park, S. (2025). Factors affecting human–AI collaboration: trust and perceived usefulness as mediators. Information, 16(10), 856.

- Kocak, B., Yardimci, A. H., Yuzkan, S., Keles, A., Altun, O., Bulut, E., Bayrak, O. N., Okumus, A. A. (2023). Transparency in Artificial Intelligence Research: A Systematic Review of Availability Items Related to Open Science in Radiology and Nuclear Medicine. Academic Radiology, 30(10), 2254-2266.

- Lāma, G., & Lastovska, A. (2025). AI competence and sentiment: A mixed-methods study of attitudes and open-ended reflections. Frontiers in Artificial Intelligence, 8, Article 1658791.

- Larsson, S., Haresamudram, K., Högberg, C., Lao, Y., Nyström, A., Söderlund, K., & Heintz, F. (2023). Four Facets of AI Transparency. In S. Lindgren (Ed.), Handbook of Critical Studies of Artificial Intelligence (pp. 445-455). Edward Elgar Publishing.

- Lei, Y., & Ying, Y. (2020). Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent. In Proceedings of the 37th International Conference on Machine Learning. arXiv:2006.08157.

- Li, M., & Bitterly, T. B. (2024). How Perceived Lack of Benevolence Harms Trust of Artificial Intelligence Management. Journal of Applied Psychology, 109(11), 1794-1816.

- Limongi, R. (2024). The use of artificial intelligence in scientific research with integrity and ethics. Review of Artificial Intelligence in Education, 5(0), Article 22.

- Liu, B. (2021). In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human-AI Interaction. Journal of Computer-Mediated Communication, 26(6), 384-402.

- Liu, P., Liu, Z., Gao, Z.-F., Gao, D., Zhao, W. X., Li, Y., Ding, B., & Wen, J.-R. (2024). Do Emergent Abilities Exist in Quantized Large Language Models: An Empirical Study. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation 2024, pp. 5174–5190. European Language Resources Association.

- Marjerison, R. K., Dong, H., Kim, J.-M., Zheng, H., Zhang, Y., & Kuan, G. (2025). Understanding user acceptance of AI-driven chatbots in China's e-commerce: The roles of perceived authenticity, usefulness, and risk. Systems, 13(2), 71.

- Mayer, R. C., Davis, J. H. and Schoorman, F. D. (1995). An integrative model of organizational trust. The Academy of Management Review 20, 3, pp. 709-734.

- McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. iScience, 26(9), 107603.

- Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M. (2021). Revisiting the Calibration of Modern Neural Networks. In Advances in Neural Information Processing Systems, Volume 34, 2021. Curran Associates, Inc.

- Nielsen, L. (2024). When AI-based agents are proactive: Implications for competence and system satisfaction in human-AI collaboration. Business & Information Systems Engineering, 66(2), 241–256.

- Novozhilova, E., Mays, K., Paik, S., & Katz, J. E. (2024). More Capable, Less Benevolent: Trust Perceptions of AI Systems across Societal Contexts. Machine Learning and Knowledge Extraction, 6(1), 342-366.

- Oliveira, M., Brands, J., Mashudi, J., Liefooghe, B., & Hortensius, R. (2024). Perceptions of artificial intelligence systems' aptitude to judge morality and competence amidst the rise of chatbots. Cognitive Research: Principles and Implications, 9(1), 47.

- Pellegrina, D., & Helmy, M. (2025). AI for scientific integrity: detecting ethical breaches, errors, and misconduct in manuscripts. Frontiers in Artificial Intelligence, 8.

- Rivera, A., Abhari, K., & Xiao, B. (2025). Responsible AI design: The Authenticity-Control-Transparency (ACT) theory. Journal of the Association for Information Systems, 26(5), 1337-1389.

- Soloff, J. A., Foygel Barber, R., & Willett, R. (2024). Bagging Provides Assumption-free Stability. Journal of Machine Learning Research, 25:1-35.

- Soudi, M., Bauters, M. AI Guidelines and Ethical Readiness Inside SMEs: A Review and Recommendations. DISO 3, 3 (2024).

- Subbaswamy, A., Adams, R., & Saria, S. (2021). Evaluating Model Robustness and Stability to Dataset Shift. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. arXiv:2010.15100.

- Tsai, C.C., Kim, J. Y., Chen, Q., Rowell, B., Yang, X. J., Kontar, R., Whitaker, M., Lester, C. (2025). Effect of artificial intelligence helpfulness and uncertainty on cognitive interactions with pharmacists: Randomized controlled trial. Journal of Medical Internet Research, 27, e59946.

- Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022). Trust in human-AI interaction: Scoping out models, measures, and methods. Ithaca: Cornell University Library. arXiv:2205.00189.

- Wang, C., & Adzharuddin, N. A. (2025). Synthetic authenticity and audience trust in AI-generated intangible cultural heritage: A qualitative multimodal study of Chinese digital heritage platforms. E-Journal of Media and Society, 8(2), 1-10.

- Wang, C., Yang, C., Deng, H., & Zhang, S. (2023). Calibration in Deep Learning: A Survey of the State-of-the-Art. Published in Artificial Intelligence Review. arXiv:2308.01222.

- Wang, T-H., Xiao, W., Chahine, M., Amini, A., Hasani, R., & Rus, D. (2023). Learning Stability Attention in Vision-based End-to-end Driving Policies. In Proceedings of the 5th Annual Conference on Learning for Dynamics and Control, Proceedings of Machine Learning Research vol 211:1–13, 2023.

- Wang, Y., Wan, Y., & Wang, Z. (2017). Using experimental game theory to transit human values to ethical AI. Ithaca: Cornell University Library. arXiv:1711.05905.

- Webster, J., Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly Vol. 26 No. 2, pp. xiii-xxiii.