Human-AI Collaboration for Trust Management

Mito Akiyoshi
Senshu University
mito.akiyoshi@gmail.com

Abstract

Trust is one of the principles that human-AI teams must attain for the fulfillment of their mission. Explainable AI and the principle of computational reliabilism provide AI-intrinsic solutions for trust management. When human-AI collaboration breaks down, human-AI teams turn to common sense and intuition to recover trust. In addition, research on earlier innovations has shown that institutional and organizational mechanisms such as citizen advisory boards and standardization promote trust. This paper sketches a framework for deployable and actionable trust management mechanisms. To that end, it will:

- Identify three dimensions of trust.
- Examine the role of heterogeneous stakeholders in human-AI systems.
- Address the links among interpersonal trust, institutional trust, and trust in algorithms.
- Suggest that stakeholder heterogeneity is a multi-level and multi-faceted imperative for establishing trust in human-AI teams.

Dimensions of Trust

Trust is of course not a thing but a relation, and it involves three dimensions: a trusting entity (trustor), trusted entity (trustee), and a trust object, a domain or category in which trust is vested or from which it is withdrawn (Schilke, et al. 2021). With the proliferation of technologies that have agency capability, human trustors and trustees are increasingly working with non-human trustors and trustees. Suboptimal performance or unethical behavior of human-AI teams erodes trust in them. A breakdown of trust presents a special kind of risk for the team and the ecosystem it operates in (Akiyoshi 2022).

In a world where algorithms interact with humans to create and maintain trust, performative and ethical problems associated with faulty algorithms become a challenge for AI-human teams. These are further complicated when algorithms "go virtual", i.e. provide no tangible hardware interface, or are embedded deeply in service and products (Glikson and Woolley 2020). When suboptimal algorithms lead to low trust in a service or product, this can escalate to the broader withdrawal of trust from the organizations and institutions that have deployed them.

For example, ethical concerns such as fairness have to be instantiated as algorithm-specific characteristics like precision and recall (also known as sensitivity or the true positive rate). Fairness conditions in machine learning are relevant to trust management in human-AI collaboration, but there is a trade-off between precision and recall. One cannot have a classifier that is both perfectly precise and perfectly sensitive. Géron (2022) gives an example: "If you trained a classifier to detect videos that are safe for kids, you would probably prefer a classifier that rejects many good videos (low recall) but keeps only safe ones (high precision)" (Géron 2022: 194). Given the social implications of that kind of trade-off, trust in AI-human systems is more likely to become possible when decisions by algorithm specialists are accountable to the interests of broader groups of stakeholders.

Stakeholder Heterogeneity

But who are those stakeholders? Barocas et al. define five types of stakeholders in this domain (Barocas, Hardt, and Narayanan 2023):

- **Decision subjects** are the individuals, organizations, or other social entities that receive or are affected by decisions made by human-AI teams. If human decision subjects perceive errors or discrimination in those decisions, trust relationships are compromised, leading to challenges and rejections.
- **Data providers** are entities whose behavior supplies input for training machine learning models. In some cases, behavior patterns such as online search history and location

data are used without their consent or awareness; if discovered, this fact can also impact the perception of trustworthiness.

- **Domain experts** provide a model of professional judgement by presenting the problem and task and offering proper labeling of training examples.
- **Algorithm experts** translate domain experts' expertise into algorithms.
- **Policy makers** such as governments and corporations design policies and use centralized decisions achieved by algorithms.

When an algorithm's performance is aligned with human expectations, human-AI teamwork achieves a synergy and heterogenous stakeholders' needs are satisfied. Trust will probably result. For example, a human-AI team consisting of AI and pathologists can reduce errors in cancer detection more effectively than either a human-only or an AI-only solution (Wang et al. 2016). However, Vaccaro, Almaatouq, and Malone (2024) conducted a meta-study that complicates matters in an enlightening way. For a given task, they first compared human-only and AI-only performance and determined which actor performed better. They then compared that winner's performance with the performance of a collaborating AI-human team on the same task. If the human-AI team performed better than whichever individual actor had previously had the superior performance, only then could it be deemed an instance of "human-AI synergy" One implication of this study is that for tasks where AI-only teams perform better than human-only teams, the inclusion of humans can result in a performance loss. In those cases, poor performance of human-AI teams can exacerbate statistical discrimination and lead to a loss of trust among decision subjects (Akiyoshi 2022).

The Limits of AI-intrinsic Trust Solutions

Explainable AI and the principle of computational reliabilism provide AI-intrinsic solutions for trust management. The principle of computational reliabilism dismisses full transparency as a means to establish trust in algorithms (Durán and Jongsma 2021). Instead, it proposes to construct a good-enough algorithm rather than a perfectly comprehensible one. The principle is comprised of:

- A comparison with known solution (verification)
- A comparison with experimental data (validation)
- Robustness analysis
- A history of successful or unsuccessful implementation
- Expert knowledge

Yet the principle of computational reliablilism ignores some of the multiple stakeholders mentioned above. Only "the cognitive agent assessing the process" engages in the process of trust-building (Russo, Schliesser, and Wagemans 2023). In the formulation of Barocas et al., these are the domain experts and algorithm experts, while the principle neglects the interests and knowledge of decision subjects, data providers, and policy makers (Barocas, Hardt, and Narayanan 2023). Without the involvement of those missing actors, trust-building is at risk.

Trust relationships are further threatened by perverse instantiation (Bostrom 2016). This occurs when human-AI collaboration goes awry as a result of the difficulty of specifying and communicating the boundaries that reasonable humans would place on the desired goals. For example, preventing deaths from cancer is an ideal outcome, but murdering everyone to achieve it is not an appropriate instance of that outcome, at least for the humans involved. (Wooldridge 2021).

When a gap emerges between human expectations and the results of human-AI collaboration, algorithm experts often attempt to incorporate more human feedback to improve communication. They view common sense and intuition as valuable resources to enhance human-AI team reasoning and decision-making. Wooldridge states, "Many of us feel more comfortable if we know that a human being is making a decision that has serious consequences for another human being" (Wooldridge 2020:145). Similarly, Jarrahi suggests, "Humans perform better in the face of decisions that require an intuitive approach" (Jarrahi 2018:580).

Both the adoption of computational reliabilism and the incorporation of more human feedback are AI-intrinsic solutions that tend to ignore differences among heterogeneous stakeholders and within each of those groups. AI-intrinsic solutions by themselves do not address the issue of perverse instantiation and establish trust in human-AI collaboration.

In the case of statistical discrimination, for example, the use of commonly known facts could result in *more* discriminatory decisions by a human-AI team. In fact, to eliminate statistical discrimination in decisions, one needs expertise in the subject matter and in the issue of inequality in general, rather than just "common sense and intuition". Experts on inequality would oppose the claim that "more human feedback" prevents poor performance and trust erosion. Contrary to the idea that "many of us feel more comfortable if we know that a human being is making a decision that has serious consequences for another human being," the literature on audit studies has shown that human judges discriminate against racial and ethnic minorities (Massey and Denton 1993). Some people in some circumstances may thus feel *less* comfortable knowing that a human being alone is making a consequential decision. We need source and PDF files that can be used in a variety of ways and can be output on a variety of devices.

An Integrative Approach to Human-AI Collaboration

Common sense and intuition could be supplemental resources to improve human-AI team performance, but human feedback by itself is not a panacea for a trust crisis associated with perverse instantiation.

Human-AI teams may not achieve intended goals, or when they do, they may bring about unintended consequences. Furthermore, they may sacrifice the interests of decision subjects. Barocas et al. point out that a power shift occurs among stakeholders with the implementation of machine learning. Decision subjects, data providers, and domain experts lose power as decisions are taken away from them and move into the hands of algorithm experts and policymakers (Barocas, Hardt, and Narayanan 2023).

Such a power shift is one aspect of adverse instantiation: a gap between what was wished for and what was really desired becomes salient, with profound social consequences. The limits of AI-intrinsic solutions for trust management can be conceptualized as a special case of what economists call an externality. These are unintended consequences of a transaction that can affect various stakeholders, including entities external to that transaction. Industrial pollution of a public waterway is a classic example.

Common sense and intuition appear to be attractive tools for strengthening trust relationships, but their effectiveness is constrained by two key factors: 1) They are fraught with cognitive errors. In some cases, they can reduce or eliminate perverse instantiation, but in others, they can further complicate it. 2) Different stakeholders have different and sometimes competing notions of what common sense and intuition in fact consist of.

Note that these issues are distinct from the difficulty of translating human knowledge into a form that an algorithm can understand. The challenge of teaching common sense to algorithms is well-known, and the strategy of exhaustive enumeration of its contents to train an algorithm has long been abandoned (Wooldridge 2021). Rather, the underlying issue is the heterogeneity of the stakeholders.

The literature on earlier innovations establishes that trust is facilitated when diverse stakeholders convene to collectively discuss technical options in an open setting (Akiyoshi 2022). Since the construction of trustworthy trustees (e.g., human-AI teams) does not automatically entail trust in trust objects (i.e., their decisions), an institutional approach that supports input from heterogeneous stakeholders can compensate for the shortcomings of AI-intrinsic solutions. For example, MacKenzie's study of high-frequency trading illustrates how in that domain of human-AI collaboration trust is a dynamically negotiated and renegotiated accomplishment among various stakeholders (MacKenzie 2021).

To recapitulate, establishing trust among team components in human-AI collaboration is difficult, because of the presence of heterogeneous stakeholders. Algorithms are often a "black box" to humans, and achieving human-AI synergy is challenging. Potential solutions include computational reliabilism and human feedback, but they each have limits. An integrative approach supplements AI-intrinsic attempts by including the full range of relevant stakeholders in the trust-building process.

Acknowledgments

The author would like to thank William Lawless, Gerald Lombardi, and the participants in the Current and Future Varieties of Human-AI Collaboration Symposium at the 2025 AAAI Spring Symposium Series for their comments.

References

Akiyoshi, M. 2022. Trust in Things: A Review of Social Science Perspectives on Autonomous Human-Machine-Team Systems and Systemic Interdependence. *Frontiers in Physics* 10. doi.org/10.3389/fphy.2022.951296

Bostrom, N. 2016. *Superintelligence: Paths, Dangers, Strategies*. Reprint edition. Oxford: Oxford University Press.

Barocas, S, Hardt M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, Massachusetts: The MIT Press.

Durán, J. M. and Jongsma, K. R. 2021. Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI. *Journal of Medical Ethics* 47 (5): 329. doi.org/10.1136/medethics-2020-106820.

Géron, A. 2022. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 3rd edition. Sebastopol, California: O'Reilly Media.

Glikson, E. and Woolley, A. W. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14(2): 627–60. doi.org/10.5465/annals.2018.0057.

Jarrahi, M. H. 2018. Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making. *Business Horizons* 61(4): 577–86. doi.org/10.1016/j.bushor.2018.03.007.

MacKenzie, D. 2021. *Trading at the Speed of Light: How Ultrafast Algorithms Are Transforming Financial Markets*. Princeton University Press.

Massey, D. S. and Denton, N. A. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, Mass.: Harvard University Press.

Russo, F.; Schliesser, E.; and Wagemans, J. 2023. Connecting Ethics and Epistemology of AI. *AI & SOCIETY*, January. doi.org/10.1007/s00146-022-01617-6.

Schilke, O.; Reimann, M.; and Cook, K. S.; 2021. Trust in Social Relations. *Annual Review of Sociology* 47 (1): 239–59. doi.org/10.1146/annurev-soc-082120-082850.

Vaccaro, M.; Almaatouq, A.; and Malone, T. 2024. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nature Human Behaviour* 8 (12): 2293–2303. doi.org/10.1038/s41562-024-02024-1.

Wang, D.; Khosla A.; Gargeya, R; Irshad, H; Beck; A. H. 2016. Deep Learning for Identifying Metastatic Breast Cancer. arXiv preprint. arXiv.1606.05718. Ithaca, NY: Cornell University Library.

Wooldridge, M. 2021. *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going.* New York: Flatiron Books.