Building Trustworthy Intelligent Systems for Real-World Deployment

Adrian Mercer, Leila Rahman

Article type: Editorial

Volume 1, Issue 1 (2026)

Read Current Issue →

Abstract

Intelligent systems are moving rapidly from controlled experimental environments into operational settings where their decisions influence people, infrastructure, machines, and physical processes. Yet strong performance on curated benchmarks provides only limited evidence that such systems will remain dependable when confronted with distribution shift, incomplete data, sensor degradation, adversarial manipulation, human interaction, or changing operational conditions. Trustworthy deployment therefore requires a broader evidentiary standard than predictive accuracy alone. This editorial argues that operational trustworthiness is an emergent system-level property produced through the interaction of robustness, uncertainty awareness, safety engineering, explainability, cybersecurity, human oversight, governance, and lifecycle monitoring. Particular attention is given to the benchmark-to-deployment gap, failure-condition validation, inappropriate human reliance, adversarial vulnerabilities, and accountability across complex development and supply chains. The field should prioritize realistic stress testing, evaluation under distribution shift, uncertainty-aware decision architectures, transparent deployment reporting, continuous monitoring, human-in-the-loop safeguards, incident learning, and interoperable safety standards. Progress should ultimately be judged by whether intelligent systems remain reliable, understandable, and controllable under the conditions in which society actually depends on them.

Keywords: trustworthy artificial intelligence; robust machine learning; operational reliability; distribution shift; explainable AI; autonomous systems; lifecycle assurance

Citation

Mercer, A., & Rahman, L. (2026). Building Trustworthy Intelligent Systems for Real-World Deployment. Journal of AI, Automation & Intelligent Systems, 1(1). https://eldranacademic.com/jaais/articles/1/

Received	Revised	Accepted	Published
18 November 2025	12 January 2026	30 January 2026	12 February 2026

Copyright

© 2026 Adrian Mercer, Leila Rahman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) ↗. Use, distribution or reproduction in other forums is permitted, provided the original authors and copyright owners are credited and the original publication in this journal is cited in accordance with accepted academic practice.

Disclaimer

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, the publisher, the editors or the reviewers. Any product evaluated in this article, or claim made by its manufacturer, is not guaranteed or endorsed by the publisher.

1.0 Introduction

1.1 Context and Problem Framing

Artificial intelligence, machine learning, robotics, and autonomous decision technologies are increasingly embedded in operational systems rather than confined to laboratory demonstrations. Predictive models inform clinical workflows, intelligent controllers interact with industrial processes, autonomous machines perceive and respond to dynamic physical environments, and learning-enabled components increasingly support transportation, logistics, agriculture, infrastructure management, and other consequential activities. These applications transform the meaning of system performance because an error outside the laboratory may propagate through organizational processes, affect physical assets, alter human decisions, or create safety consequences that cannot be captured by a single test-set score. Engineering practice has consequently begun to frame trustworthy artificial intelligence in terms that extend beyond accuracy to include reliability, safety, resilience, transparency, accountability, and appropriate human control (Tabassi, 2023; Thiebes et al., 2021). Software-engineering experience similarly demonstrates that producing a high-performing machine-learning model is only one component of building a dependable production system (Amershi et al., 2019), while safety research on cyber-physical systems shows that operational assurance must address interactions among software, sensing, actuation, environments, and failure processes (Corso et al., 2021).

The central difficulty is that experimental success and operational trustworthiness answer different questions. Curated benchmarks commonly assume comparatively stable data-generation processes, known task boundaries, predefined performance metrics, and evaluation samples sufficiently similar to development data. Real deployments violate these assumptions. WILDS, for example, documents substantial performance degradation under naturally occurring distribution shifts across domains including healthcare, wildlife monitoring, and satellite imagery (Koh et al., 2021). Deep networks may also exploit predictive shortcuts that perform well within benchmark distributions while failing when superficial correlations change (Geirhos et al., 2020). Predictive confidence can become unreliable under dataset shift (Ovadia et al., 2019), and production-readiness work has long emphasized testing data dependencies, infrastructure behavior, model quality, monitoring, and operational degradation rather than treating offline accuracy as sufficient evidence (Breck et al., 2017). A model that scores highly under independent and identically distributed testing can therefore remain unsafe, brittle, poorly calibrated, difficult to interpret, or operationally unusable.

1.2 Why the Issue Matters

The unresolved problem is consequently not simply how to produce more accurate algorithms but how to establish credible evidence that learning-enabled systems will behave acceptably across their operational envelope. Assurance must extend from data and model development through integration, deployment, monitoring, modification, and eventual retirement. Machine-learning testing already spans properties, components, workflows, and application-specific failure modes that conventional software testing alone cannot adequately capture (Zhang et al., 2022), while lifecycle assurance frameworks stress the need to connect development evidence with continuing operational evidence (Ashmore et al., 2021). For autonomous systems, the problem is still more demanding because uncertainty in perception or decision-making may be converted directly into physical action, requiring safety reasoning that integrates engineering, ethical, and legal dimensions (Burton et al., 2020). The

central position advanced here is therefore that benchmark accuracy is necessary evidence of capability but insufficient evidence of deployability. Trustworthy intelligent systems require system-level assurance under realistic uncertainty, explicit failure conditions, human interaction, adversarial pressure, and lifecycle change.

2.0 Current Landscape

2.1 Performance-Centred AI Evaluation

The trustworthy-intelligence landscape now encompasses multiple technical and institutional approaches that were once investigated largely in isolation. Robust machine learning seeks performance stability under perturbation and changing data distributions; out-of-distribution detection attempts to identify inputs departing from the model's training support; uncertainty estimation seeks to distinguish confident knowledge from unreliable prediction; and adversarial machine learning investigates intentional manipulation of data, models, or inference processes. Early out-of-distribution detection research demonstrated that neural networks require explicit mechanisms for recognizing unfamiliar inputs rather than merely generating predictions for them (Hendrycks & Gimpel, 2017), while adversarial-machine-learning research has established that apparently small or strategically designed perturbations may expose vulnerabilities not reflected in conventional accuracy testing (Biggio & Roli, 2018). Dataset-shift detection provides another operational layer by attempting to identify changes in input populations before silent performance deterioration becomes consequential (Rabanser et al., 2019).

2.2 Robustness and Uncertainty

Explainability and interpretability have developed in parallel. Post-hoc techniques such as LIME provide local explanations for otherwise opaque predictions (Ribeiro et al., 2016), but high-stakes deployment has also generated a stronger argument for using inherently interpretable models where feasible instead of assuming that explanations can reliably compensate for black-box complexity (Rudin, 2019). Documentation frameworks broaden transparency beyond individual predictions. Model cards record intended uses, evaluation conditions, and limitations (Mitchell et al., 2019), while datasheets for datasets make provenance, composition, collection assumptions, and recommended uses more explicit (Gebru et al., 2021). Internal algorithmic auditing further extends accountability across development processes rather than locating responsibility exclusively in the final model (Raji et al., 2020).

2.3 Explainability and Monitoring

These technical mechanisms increasingly sit within formal risk and management frameworks. The NIST AI Risk Management Framework treats trustworthiness as context dependent and multidimensional rather than reducible to a single technical property (Tabassi, 2023). ISO/IEC 23894 provides guidance for integrating AI-related risk management into organizational processes, while ISO/IEC 42001 establishes requirements for an AI management system (International Organization for Standardization & International Electrotechnical Commission, 2023a, 2023b). For embodied systems, assurance requirements become more stringent because learning components operate within feedback loops involving sensors, actuators, physical dynamics, communication networks, and people. ISO 21448, addressing safety of the intended functionality in road vehicles, exemplifies the need to consider hazards arising even when components have not technically failed but system capability is inadequate for a particular condition (International Organization for Standardization, 2022). Trustworthy deployment is thus moving toward an integrated discipline of technical robustness, safety engineering, documentation, governance, and continuing operational verification.

2.4 Human Oversight and Operational Integration

3.0 Critical Challenges

3.1 The Benchmark-to-Deployment Gap

The first major challenge is the persistent gap between benchmark performance and deployment performance. Aggregate test accuracy can conceal dependence on contextual shortcuts, underrepresented subgroups, or environmental conditions absent from the benchmark. Hidden stratification in medical imaging, for example, can generate clinically important failure patterns even when aggregate performance appears strong (Oakden-Rayner et al., 2020). External validation of a widely implemented sepsis prediction model likewise demonstrated that performance observed outside the original development context may differ substantially from expectations based on prior validation (Wong et al., 2021). Such findings illustrate a broader functional problem: demonstrating that an AI component can produce outputs does not establish that it performs the intended real-world function reliably within the surrounding system (Raji et al., 2022).

Distribution shift compounds this problem because deployed environments are not static. Changes in populations, equipment, weather, workflows, policies, sensors, user behavior, upstream software, or data collection can progressively invalidate assumptions established during development. Healthcare provides a particularly clear illustration because differences among institutions, clinical processes, disease prevalence, and patient populations can undermine transportability even when algorithmic performance appeared credible during development (Kelly et al., 2019). The same principle applies to manufacturing systems experiencing equipment wear, autonomous vehicles encountering unusual environmental conditions, or agricultural robots operating across changing seasons and terrain. Static validation therefore provides evidence about a particular evaluation distribution rather than a permanent guarantee of system reliability.

3.2 Distribution Shift and Rare Events

A third challenge is inadequate validation under failure conditions. Conventional testing tends to emphasize expected inputs and nominal operation, whereas dependable autonomous and cyber-physical systems require evaluation of rare but consequential combinations of sensor errors, communication faults, environmental ambiguity, degraded actuators, uncertain perception, and unexpected behavior by other agents. Black-box safety validation research demonstrates the importance of systematically searching for failure trajectories rather than waiting for them to appear in ordinary testing (Corso et al., 2021). The problem becomes especially acute when combinatorial operational spaces make exhaustive testing impossible. Assurance must consequently establish credible coverage arguments, risk-informed scenario selection, fail-safe behavior, and acceptable residual risk rather than assuming that the absence of observed failures implies safety.

3.3 Safety and Accountability

Opacity and human interaction create another layer of risk. Explanations may improve intelligibility, but they can also be mistaken for proof of correctness. Human operators can become over-reliant on automated recommendations, particularly when systems are usually accurate and failures are infrequent. Research on automation has long distinguished calibrated trust from indiscriminate reliance (Lee & See, 2004). Experimental evidence indicates that cognitive forcing mechanisms can reduce inappropriate reliance on AI advice (Buçinca et al., 2021), while human-robot research has shown that people may follow robotic guidance even after observing poor robot performance (Robinette et al., 2016). Human oversight is therefore not inherently protective; its effectiveness depends on interface design, authority structures, workload, training, information quality, and whether the human can detect and meaningfully correct machine failure.

3.4 Reproducibility and Evidence Quality

Finally, cybersecurity, model drift, and fragmented accountability challenge lifecycle trustworthiness. Adversarial vulnerabilities can target training data, model behavior, interfaces, or sensor inputs, while ordinary non-malicious drift can gradually reduce performance. At the organizational level, failures may originate far upstream from the deployed model through poor data practices, undocumented assumptions, third-party components, or changing dependencies. Sambasivan et al. (2021) describe how neglected data work can produce cascading consequences in high-stakes AI development. When datasets, foundation components, integration software, cloud services, operators, and application owners belong to different organizations, responsibility for monitoring and corrective action can become diffuse. Trustworthiness therefore requires accountable lifecycle processes as much as sophisticated algorithms.

4.0 Discussion and Priorities

4.1 Integrative Discussion

These challenges indicate that trustworthiness is not a property that can be attached to a model after training. It emerges from relationships among model behavior, data quality, uncertainty, hardware and software architecture, human interaction, organizational controls, cybersecurity, and the deployment environment. Lifecycle assurance is therefore more appropriate than one-time certification because evidence generated during development must remain connected to evidence gathered after deployment and after system modification (Ashmore et al., 2021). The relevant unit of evaluation is not merely the trained model but the socio-technical system in which that model acquires inputs, generates decisions, influences action, and is monitored or overridden.

This systems perspective also exposes trade-offs that simplified notions of trustworthy AI can obscure. Higher predictive complexity may improve benchmark accuracy while reducing interpretability or increasing computational latency. Human control may limit some autonomous failure modes but can also create slow responses, automation bias, or ambiguous responsibility. Real-time systems may be unable to use computationally expensive uncertainty or explanation methods at every decision point. Adaptive systems can respond to changing distributions, yet online adaptation alters the very artifact that was originally validated, creating tension between responsiveness and assurance stability. In high-stakes contexts, the relevant design question is therefore not whether interpretability, autonomy, or adaptation is universally preferable but which combination provides defensible performance for a defined operational context.

The same principle applies to safety and cybersecurity. A system may be statistically robust but unsafe because its failures occur in hazardous physical states; explainable but insecure because its interfaces remain vulnerable; or heavily governed but operationally unreliable because drift is detected too late. The NIST framework's multidimensional treatment of trustworthy characteristics reflects this interdependence (Tabassi, 2023). In embodied systems, engineering assurance must additionally account for the transition from informational errors to physical consequences (Burton et al., 2020). Trustworthy intelligent systems should therefore be evaluated through structured claims about intended use, operating conditions, known limitations, uncertainty, failure response, human roles, security controls, and continuing evidence that these claims remain valid.

Complementary scholarship informing this synthesis includes International Organization for Standardization & International Electrotechnical Commission. (2023a); International Organization for Standardization & International Electrotechnical Commission. (2023b); and International Organization for Standardization. (2022). These works broaden the evidentiary base considered across the field.

4.2 Research and Practice Priorities

A first priority is to replace benchmark-only evaluation with standardized stress testing that reflects realistic operating variability. Evaluations should deliberately include temporal, geographic, demographic, environmental, institutional, and hardware-related shifts rather than assuming independent and identically distributed test conditions. WILDS demonstrates the value of benchmark suites constructed around naturally occurring distribution shifts (Koh et al., 2021), but application-specific testing must go further by defining operational envelopes and examining performance near their boundaries. Meaningful progress would be demonstrated by reproducible evaluations showing not merely average accuracy but degradation profiles, subgroup behavior, recovery mechanisms, and conditions under which the system should abstain or transfer control.

A second priority is uncertainty-aware and failure-aware decision architecture. Systems should distinguish between high-confidence routine cases and conditions in which evidence is weak, inputs are anomalous, or system competence is doubtful. Uncertainty estimates must themselves be evaluated under distribution shift rather than assumed to remain calibrated (Ovadia et al., 2019). Safety validation should systematically search for plausible failure trajectories and combinations of disturbances, especially in autonomous and cyber-physical systems. Evidence for deployment should include performance under sensor degradation, missing information, network interruption, actuator limitations, adversarial perturbation, and realistic edge cases rather than nominal-operation tests alone.

A third priority is transparent reporting of deployment conditions and system dependencies. Model cards, dataset documentation, and related reporting mechanisms should evolve from descriptive supplements into operational artifacts that specify intended populations, environmental assumptions, excluded conditions, uncertainty characteristics, known failure modes, external dependencies, update policies, and monitoring responsibilities (Gebru et al., 2021; Mitchell et al., 2019). Such reporting should support reproducibility across sites and make explicit which aspects of validation evidence are transferable and which require local re-evaluation.

Lifecycle monitoring constitutes a fourth priority. Deployed systems require baselines for detecting drift, changes in calibration, abnormal inputs, cybersecurity events, and deterioration in downstream outcomes. Shift-detection methods can provide early warnings (Rabanser et al., 2019), but monitoring frameworks must connect detection to predefined response procedures: investigation, rollback, model suspension, human escalation, or revalidation. Incident reporting should likewise become a source of shared engineering knowledge, allowing recurrent failure mechanisms to inform benchmarks, standards, and system design.

A fifth priority is to engineer human oversight rather than merely require a human to remain nominally "in the loop." Interfaces should communicate uncertainty and limitations in forms that enable calibrated reliance, and organizations should establish clear authority for intervention. Evidence should demonstrate that human-machine teams perform acceptably under workload, time pressure, disagreement, and automation failure, not simply that users prefer the interface. A sixth priority is convergence around interoperable assurance and governance practices. AI risk-management guidance, management-system standards, domain safety standards, cybersecurity requirements, and technical verification methods should be connected into coherent evidence frameworks. For high-risk deployment, the burden of evidence should increase with the severity and irreversibility of potential harm. Reproducible performance, realistic validation, traceable documentation, defined human control, security testing, monitoring capacity, and documented responses to failure should become components of a credible deployment case rather than optional additions.

4.3 Implications for the Journal's Scope

For the Journal of AI, Automation & Intelligent Systems (JAAIS), the transition from proof-of-concept capability to operational credibility defines an important scholarly frontier. High benchmark performance remains scientifically valuable, but manuscripts concerned with deployable intelligent systems should increasingly explain the

conditions under which reported performance is expected to hold. Research that integrates robustness testing, uncertainty quantification, failure analysis, safety validation, interpretable decision-making, cybersecurity, human-machine collaboration, or post-deployment monitoring can provide substantially stronger evidence than isolated improvements in test accuracy.

JAAIS is therefore well positioned to encourage contributions that connect algorithmic innovation with systems engineering and empirical validation. Particularly relevant work includes reproducible evaluations under distribution shift, validated intelligent-control and robotic systems, safety assurance for autonomous platforms, real-world human-AI studies, monitoring and adaptation methods, and technically grounded governance or assurance frameworks. Studies that report negative results, failure boundaries, cross-site validation, or conditions under which an intelligent system should not be deployed are also scientifically consequential. As lifecycle-assurance research emphasizes, credible intelligent systems are built through evidence accumulated across development and operation rather than through a single performance demonstration (Ashmore et al., 2021).

4.4 Conclusion

The central challenge facing intelligent-systems engineering is no longer simply whether increasingly capable models can achieve impressive results under controlled evaluation. It is whether those capabilities can be translated into systems that remain dependable when data change, sensors deteriorate, users behave unexpectedly, adversaries interfere, components are updated, and operational constraints depart from laboratory assumptions. Benchmark accuracy establishes capability within a defined experimental frame; it does not establish safe and reliable behavior throughout the life of a deployed system.

Trustworthy deployment therefore requires a change in both evidence and engineering practice. Robustness, explainability, uncertainty awareness, safety, cybersecurity, human oversight, monitoring, documentation, and governance must be treated as interacting elements of one assurance problem. The strongest future contributions will define operational conditions explicitly, test credible failures deliberately, monitor systems continuously, and make residual uncertainty visible rather than hiding it behind aggregate performance scores. For intelligent systems that increasingly participate in consequential decisions and physical processes, scientific maturity will ultimately be measured less by how well they perform when the world resembles the benchmark than by how responsibly and predictably they behave when it does not.

5.0 Declarations

5.1 Author Contributions

Adrian Mercer and Leila Rahman contributed to the conceptual development, literature synthesis, critical interpretation, drafting and revision of the editorial.

5.2 Funding

No external funding was received for the preparation of this editorial.

5.3 Competing Interests

The authors declare that they have no conflicts of interest or competing interests relevant to this editorial.

5.4 Data Availability

No original datasets were generated or analysed for this editorial.

5.5 AI Use Disclosure

Generative AI was used to support literature organization and language drafting during preparation of this editorial. The authors remain responsible for independently checking the accuracy of the arguments, citations, references, and final submitted text in accordance with the publisher's policy.

6.0 References

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H. C., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. ACM Computing Surveys, 54(5), Article 111, 1–39. https://doi.org/10.1145/3453444

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. 2017 IEEE International Conference on Big Data (Big Data), 1123–1132. https://doi.org/10.1109/BigData.2017.8258038

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), Article 188, 1–21. https://doi.org/10.1145/3449287

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. Artificial Intelligence, 279, 103201. https://doi.org/10.1016/j.artint.2019.103201

Corso, A., Moss, R. J., Koren, M., Lee, R., & Kochenderfer, M. J. (2021). A survey of algorithms for black-box safety validation of cyber-physical systems. Journal of Artificial Intelligence Research, 72, 377–428. https://doi.org/10.1613/JAIR.1.12716

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations.

International Organization for Standardization & International Electrotechnical Commission. (2023a). Information technology —Artificial intelligence—Guidance on risk management (ISO/IEC 23894:2023). International Organization for Standardization.

International Organization for Standardization & International Electrotechnical Commission. (2023b). Information technology —Artificial intelligence—Management system (ISO/IEC 42001:2023). International Organization for Standardization.

International Organization for Standardization. (2022). Road vehicles—Safety of the intended functionality (ISO 21448:2022). International Organization for Standardization.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I. S., Beery, S. M., Leskovec, J., Kundaje, A., . . . Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. Proceedings of the 38th International Conference on Machine Learning, 139, 5637–5664.

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

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. https://doi.org/10.1145/3287560.3287596

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Proceedings of the ACM Conference on Health, Inference, and Learning, 151–159. https://doi.org/10.1145/3368555.3384468

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32, 13991–14002.

Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. Advances in Neural Information Processing Systems, 32, 1394–1406.

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). The fallacy of AI functionality. 2022 ACM Conference on Fairness, Accountability, and Transparency, 959–972. https://doi.org/10.1145/3531146.3533158

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. https://doi.org/10.1145/3351095.3372873

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 101–108. https://doi.org/10.1109/HRI.2016.7451740

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article 39, 1–15. https://doi.org/10.1145/3411764.3445518

Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. Electronic Markets, 31, 447–464. https://doi.org/10.1007/s12525-020-00441-4

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2022). Machine learning testing: Survey, landscapes and horizons. IEEE Transactions on Software Engineering, 48(1), 1–36. https://doi.org/10.1109/TSE.2019.2962027