Can We Trust AI at Work?

A Practical Guide to Reliable, Explainable, Secure, Fair, and Accountable Artificial Intelligence in Industry

Jeffrey A. Young
Research to Reality

August 2026

Abstract

Artificial intelligence is increasingly used to recommend, rank, predict, detect, summarize, generate, and automate decisions in business. That does not mean an AI system is automatically trustworthy, nor does a high laboratory score prove that it will remain dependable after deployment. This paper translates and expands the technical material in *Trustworthiness of AI Systems in Industry: Assessment and Enhancement* for a non-technical audience. Its central argument is that trustworthiness is not one property and cannot be established by accuracy alone. Depending on context, relevant concerns include validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy, and fairness. These dimensions align closely with the NIST AI Risk Management Framework (AI RMF) (National Institute of Standards and Technology, 2026b; Tabassi, 2023). The paper explains reliability, robustness, interpretability, privacy protection, fairness, model drift, adversarial threats, human oversight, testing, monitoring, and governance without requiring a technical background. It also identifies claims in the source report that require correction, qualification, or stronger evidence.

Contents

1 Introduction: Useful AI Is Not Automatically Trustworthy

Artificial intelligence can produce impressive results and still be unsuitable for an important decision. A model can be accurate on average while failing badly for a small group of customers. It can work well during development and deteriorate after deployment because the world changes. It can generate a plausible explanation without proving that its decision is correct. It can also be technically accurate while violating privacy expectations, creating security risks, or encouraging people to rely on it outside the purpose for which it was tested.

The source report correctly treats trust as a multidimensional problem involving reliability, robustness, interpretability, and ethical governance. It introduces ensemble learning, SHAP, LIME, adversarial testing, and differential privacy. The managerial question underneath these techniques is more important than any individual algorithm: *What evidence should an organization demand before allowing AI to influence people, money, operations, or safety?*

NIST provides a strong framework for answering that question. The AI RMF describes trustworthy AI as valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed (National Institute of Standards and Technology, 2026b; Tabassi, 2023). NIST also stresses context: characteristics must be balanced according to the system's use. There is no universal "trustworthiness score."

A better question than "Can AI be trusted?" is therefore: *Can this AI system be trusted for this task, in this environment, for these users, under these conditions, with these controls, and with consequences the organization understands and can manage?*

2 Trustworthiness in Plain Language

Trust should not be treated as a personality characteristic of a machine. It is better understood as a conclusion supported by evidence. The required evidence depends on consequences. A movie recommender can tolerate mistakes that would be unacceptable in medical care, credit, employment, industrial control, or critical infrastructure.

For non-technical readers, the major dimensions can be translated into seven questions:

1. **Validity and reliability:** Does it perform the intended task consistently under realistic conditions?
2. **Safety:** If it is wrong, what can happen, and can it fail safely?
3. **Security and resilience:** Can it withstand malicious manipulation, unusual inputs, outages, and operational stress?
4. **Accountability and transparency:** Is ownership clear, and can the organization explain what the system is for, how it is governed, and who is responsible?
5. **Explainability and interpretability:** Can appropriate users understand enough about an output to use it responsibly?
6. **Privacy:** Are personal and sensitive data protected throughout the lifecycle?
7. **Fairness:** Has the organization looked for harmful disparities and established a process to address them?

NIST treats AI as socio-technical: outcomes arise from models, data, people, organizations, processes, and environments together (Tabassi, 2023). An excellent model inside a poorly governed process can still produce an untrustworthy system.

3 Accuracy: Necessary but Not Sufficient

The technical source discusses accuracy, precision, recall, F1 score, and AUC-ROC. These are useful, but no single metric answers the trust question. Consider a rare-event detector. A system that predicts the common outcome almost every time may appear highly accurate while missing many events the business actually cares about. Conversely, increasing detection may create costly false alarms.

The correct metric depends on the consequences of each kind of error. Management should ask what a false positive means, what a false negative means, who experiences the consequences, and whether the evaluation data resemble deployment conditions.

Evaluation data should be separated from development data. Relevant subgroups, edge cases, and operating scenarios should be examined. NIST specifically calls for systems to be demonstrated as valid and reliable and for limits on generalization beyond development conditions to be documented (National Institute of Standards and Technology, 2026a). A performance percentage without context is not a trustworthiness assessment.

4 Reliability, Drift, and the Reality of Deployment

Reliability concerns consistency over time and across conditions. A system may perform well in a demonstration and behave differently when data sources change, sensors degrade, users change behavior, policies change, or the economic environment shifts.

This is often called distribution shift or data drift. Imagine a demand model trained on several years of purchasing behavior. A supply shock, new competitor, regulatory change, or change in customer behavior may weaken historical relationships. The model has not necessarily "broken"; its environment has changed.

Deployment must therefore be followed by monitoring. Depending on the application, organizations may track changes in inputs, output distributions, error rates, subgroup performance, unusual usage, and business outcomes. When correct answers become known slowly, periodic expert review may be needed.

The source recommends recalibration and retraining. That is reasonable but should not become "retrain on a schedule." Retraining can introduce new errors or biases. A mature process treats it as a controlled change: evaluate the candidate model, compare it with the current version, document the reason, obtain appropriate approval, and preserve rollback capability.

5 Robustness and Adversarial Threats

Reliability asks whether performance remains consistent. Robustness asks what happens when conditions become difficult, unexpected, noisy, or deliberately hostile.

Research has shown that carefully designed changes to inputs can cause machine-learning models to make incorrect predictions (Papernot et al., 2017; Szegedy et al., 2014). NIST's current adversarial-machine-learning taxonomy covers evasion, poisoning, privacy, and, for generative AI, misuse attacks, while emphasizing that mitigations have limitations (Vassilev et al., 2025).

For business leaders, the broader lesson is that AI expands the attack surface. Threats can arise during data collection, training, deployment, user interaction, or integration with other software. Generative systems add concerns such as malicious instructions, data leakage, and misuse.

Robustness testing should resemble realistic stress testing. Test missing fields, malformed data, extreme values, ambiguous requests, novel categories, dependency failures, and malicious inputs.

A model that works only under ideal conditions is not ready for an environment in which ideal conditions cannot be guaranteed.

6 Ensembles: Why Multiple Models May Help

The source discusses ensemble learning, in which predictions from multiple models are combined. Random forests, boosting, and voting systems are examples. Ensembles can improve predictive performance or stability in many applications (Goodfellow et al., 2016).

But more models do not automatically create more trust. Models trained on the same biased data can preserve the same bias. Correlated errors do not become independent evidence merely because multiple models vote. More complexity can also make systems harder to explain and operate.

The defensible conclusion is conditional: ensembles are one technical option that *may* improve performance or robustness, but improvement must be demonstrated on relevant validation and test data. The code example in the source PDF is illustrative; it does not itself prove improved generalization in an industrial setting.

7 Interpretability and Explainability

Two influential approaches in the source are LIME and SHAP. LIME explains an individual prediction by approximating a complex model locally with a simpler interpretable model (Ribeiro et al., 2016). SHAP connects feature-attribution explanations to Shapley values and assigns features contributions for particular predictions (Lundberg & Lee, 2017; Shapley, 1953).

These methods can help answer: Which factors pushed a prediction upward? Which pushed it downward? Why did the model treat two cases differently?

An explanation, however, is not proof of correctness. Feature attribution describes model behavior; it does not prove causality, fairness, or factual correctness. A bad model can be made easier to understand without becoming a good model.

NIST therefore treats explainability and interpretability as related to but distinct from other trustworthiness characteristics (National Institute of Standards and Technology, 2026b). The U.S. FDA, Health Canada, and U.K. MHRA likewise emphasize transparency and performance of the human-AI team for machine-learning-enabled medical devices (U.S. Food and Drug Administration, 2024). Explanations should be designed for the people who must act on them, not merely produced as reassuring graphics.

8 Privacy and Differential Privacy

AI may depend on large amounts of sensitive data. Privacy concerns what is collected, why it is needed, who can access it, how long it is retained, whether it can be linked to individuals, and what may be inferred from outputs.

The source correctly introduces differential privacy. Differential privacy is a mathematical framework designed to limit how much the inclusion or exclusion of one person's record can influence a randomized output (Dwork & Roth, 2014). In plain language, it seeks to learn population-level information while limiting what can be learned about a particular participant.

There are trade-offs. Stronger privacy can reduce utility, and privacy parameters require careful selection and accounting. Differential privacy is not a switch that makes an AI system private. It is a framework with assumptions and implementation requirements.

The same caution applies to federated learning. Keeping data decentralized may reduce some data movement, but federated learning is not synonymous with privacy. Privacy should be an architectural requirement, not a review performed after the model is complete.

9 Fairness and Harmful Bias

AI systems learn from data generated by institutions and societies. Those data can reflect historical inequality, incomplete representation, measurement problems, and outdated policy. A model can reproduce or amplify patterns an organization does not intend to preserve.

Fairness is not achieved simply by deleting protected attributes; other variables can act as proxies. Nor is fairness represented by one universal metric. Different fairness criteria can conflict, and the appropriate standard depends on context, affected populations, law, and the type of decision.

The source mentions reweighting and adversarial debiasing. These are legitimate families of techniques, but not universal solutions. Technical mitigation should follow a clear definition of the harm: Who can be harmed? What disparity matters? How will it be measured? What threshold triggers action? Who decides whether residual disparity is acceptable?

NIST deliberately describes trustworthy AI as fair with harmful bias *managed*, rather than promising a universally "unbiased" system (National Institute of Standards and Technology, 2026b; Tabassi, 2023).

10 Class Imbalance and Why SMOTE Requires Caution

The source recommends SMOTE or oversampling for imbalanced data. This requires qualification. SMOTE generates synthetic minority-class examples to alter the training distribution. It can be useful, but research has documented limitations. Synthetic patterns may diverge from the original minority distribution, with behavior affected by sample size, dimensionality, and neighborhood choices (Elreedy & Atiya, 2019). Other work has identified classifier-specific problems and settings in which no rebalancing can be competitive (Blagus & Lusa, 2013; Sakho et al., 2024).

Class imbalance should therefore not trigger an automatic recipe. Determine whether imbalance is actually harming the outcomes that matter, compare alternatives, and evaluate on untouched real-world test data. Synthetic training examples should never be allowed to make a test set look more representative than the real deployment population.

11 Security and Resilience

Traditional cybersecurity remains essential: access control, secure development, patching, logging, network protection, secrets management, and incident response still matter. AI adds new failure modes.

Attackers may manipulate training data, craft evasive inputs, extract sensitive information, exploit model interfaces, or misuse generative systems. NIST emphasizes lifecycle-wide attacks and the limitations of mitigations (Vassilev et al., 2025).

Threat modeling should identify assets, plausible attackers, entry points, consequences, and controls. This becomes especially important when generative AI can call tools. The risk is no longer merely an incorrect sentence. A connected system may retrieve documents, send messages, change records, or trigger workflows. The more authority an AI system receives, the more important authorization, isolation, logging, confirmation, and rollback become.

12 Human Oversight Is a Designed Control

Saying "a human reviews it" is not enough. Reviewers may defer to automation, become overloaded, or lack information needed to challenge a prediction. A nominal approval step can become a rubber stamp.

Effective oversight specifies what reviewers see, what warning signs matter, when escalation is required, and what authority they possess. The organization should measure the combined human-AI process rather than the model alone. Medical-device guidance's emphasis on the human-AI team illustrates this principle (U.S. Food and Drug Administration, 2024).

Human oversight is meaningful only when humans can investigate, override, pause, or escalate.

13 How Much Autonomy Is Appropriate?

There is no defensible universal answer to whether AI is ready to manage industrial processes without human oversight. Autonomy should be proportional to risk and evidence.

A practical autonomy ladder is:

1. **Assist:** AI supplies information; a person decides.
2. **Recommend:** AI proposes an action; a person approves.
3. **Act with review:** AI acts, but actions are routinely reviewed and reversible.
4. **Act within limits:** AI acts inside predefined boundaries and escalates exceptions.
5. **High autonomy:** AI acts with minimal routine review, supported by strong evidence, monitoring, and controls.

Moving upward should require stronger validation, safer failure modes, more mature governance, and clearer accountability. The question is not whether the technology looks advanced. It is whether the complete socio-technical system has demonstrated acceptable residual risk.

14 A Practical Lifecycle for Trustworthy AI

NIST organizes risk management around **Govern, Map, Measure, and Manage** (National Institute of Standards and Technology, 2026a; Tabassi, 2023). A non-technical organization can operationalize these ideas as follows.

14.1 Define the decision

Write down exactly what AI is allowed to predict, recommend, generate, or do. Avoid vague goals.

14.2 Identify affected people and assets

List customers, employees, partners, operators, communities, data, money, equipment, and processes that could be affected.

14.3 Define failure

Describe false positives, false negatives, unsafe actions, privacy failures, security failures, discriminatory outcomes, misleading explanations, and outages.

14.4 Set acceptance criteria before testing

Define minimum performance and maximum tolerable risk before seeing results. Otherwise teams may accept whatever score the model produces.

14.5 Validate realistically

Use deployment-relevant data. Keep evaluation data separate from development data. Test subgroups, edge cases, and unusual conditions.

14.6 Stress-test

Test malformed inputs, missing data, malicious behavior, shifts, dependency failures, and scenarios near system boundaries.

14.7 Design oversight

Specify who reviews, approves, overrides, investigates, pauses, and escalates.

14.8 Protect data and access

Apply privacy, cybersecurity, access-control, retention, and logging requirements appropriate to the use.

14.9 Deploy gradually

Where possible, begin with limited scope, shadow operation, or human approval before expanding autonomy.

14.10 Monitor and reassess

Track technical metrics and business outcomes. Define thresholds for investigation, rollback, retraining, or suspension.

15 Questions Executives Should Ask

Leaders do not need to become machine-learning engineers, but they should ask questions that expose weak evidence:

- What exact decision or action is the system allowed to influence?
- What data were used to build and test it?
- How similar are test conditions to deployment?
- What kinds of mistakes does it make, and who bears their cost?
- How does performance vary across important groups and scenarios?
- What happens when the model is uncertain or outside its tested range?
- What security threats were tested?

- What confidential or personal information can enter or leave?
- What explanations are available to operators and affected people?
- Who can override or stop it?
- What is logged?
- How is drift detected?
- What triggers reevaluation?
- Who remains accountable after deployment?

If these questions cannot be answered, the organization does not yet possess enough evidence to justify strong reliance.

16 Cost and the Myth of Free Trustworthiness

Trust controls cost money and time. More testing requires resources. Monitoring requires infrastructure. Privacy protections can reduce utility. Explainability adds complexity. Adversarial testing requires expertise. Human review reduces some labor savings.

These costs should be compared with the cost of failure: financial loss, regulatory exposure, customer harm, reputational damage, operational disruption, or security incidents. Assurance should therefore be risk-based. Low-consequence internal drafting can justify lighter controls than systems affecting medical care, employment, credit, safety, critical infrastructure, or major financial decisions.

Trustworthiness is not an optional layer added after innovation. It is part of the cost of deploying AI responsibly.

17 Validation Findings: Claims in the Source That Need Correction or Qualification

The source PDF is a useful technical outline, but several statements should not be repeated as established fact without qualification.

17.1 The EU AI Act citation is outdated

The source cites the European Commission's 2021 *proposal*. That reference is historically valid but no longer represents the current regulatory status. A current publication should cite the enacted AI Act and verify the applicable obligations and implementation dates before making compliance claims.

17.2 The Watson Health case study is inadequately supported

The source states broadly that Watson Health assists clinicians, provides detailed diagnostic reasoning through interpretability frameworks, and incorporates new medical research. The bibliography supplied in the PDF does not substantiate those specific operational claims. Public reporting also documented substantial challenges in Watson healthcare initiatives. This paper therefore does not use that paragraph as evidence of a successful trustworthy-AI deployment.

17.3 The fraud-detection "over 90" claim cannot be validated

The source says companies including Cognizant and Accenture used AI for fraud detection, "achieving over 90"; the sentence is incomplete and supplies neither the metric nor a supporting source. A percentage without a definition—accuracy, recall, precision, or another measure—is not meaningful evidence. It is omitted here.

17.4 The ensemble example does not prove improvement

The source's example combines Random Forest and Gradient Boosting and says the ensemble improves decision-making. The code is illustrative and reports no validated comparison proving superiority over its component models. The correct statement is that ensembles *can* improve performance; this must be demonstrated.

17.5 Adversarial training is not universal protection

Adversarial training is a legitimate defense family, but NIST's current taxonomy covers multiple attack classes and emphasizes mitigation limitations (Vassilev et al., 2025). It should not be presented as a complete security solution.

17.6 The style-transfer/GAN defense statement needs a traceable source

The source mentions style-transfer-based defenses using GANs without providing a clearly traceable supporting citation for that specific claim. It is not relied upon here.

17.7 "Disturbance Improver Algorithms" could not be reliably validated

The source names this as a noise-tolerant KNN variant. The supplied bibliography does not identify a primary source supporting the terminology. It should be removed unless a reliable primary citation is found.

17.8 SMOTE is not a fairness guarantee

SMOTE addresses training-set class imbalance; it does not establish fairness. Published analyses document important limitations (Blagus & Lusa, 2013; Elreedy & Atiya, 2019; Sakho et al., 2024).

17.9 The historical expert-system description is misleading

The source says expert systems reduced reliance on manual rules. Classical expert systems were generally built around explicitly encoded rules and knowledge. The larger transition from rule-based AI to statistical machine learning is valid, but that sentence should be corrected.

18 Conclusion

Trustworthy AI is not created by one algorithm, one accuracy score, one explanation method, or one ethics checklist. It emerges from model quality, data quality, safety, security, privacy, fairness, human oversight, governance, and continuous evaluation.

Ensembles, SHAP, LIME, adversarial testing, and differential privacy can contribute to trustworthiness. None establishes it alone. A model can be explainable and wrong, accurate and unfair,

private and unsafe, robust to ordinary noise and vulnerable to attack, or excellent in the laboratory and poor after the environment changes.

For organizations, this is not an argument against AI. It is an argument for disciplined adoption. NIST's framework is useful because it moves the conversation from vague confidence toward governance, mapping, measurement, and management (National Institute of Standards and Technology, 2026a; Tabassi, 2023). As systems receive more authority, organizations should demand stronger evidence.

The future of industrial AI will not be determined solely by whether machines can make decisions. It will also be determined by whether organizations can demonstrate when those decisions deserve reliance, detect when that reliance is no longer justified, and remain accountable for the consequences.

References

Blagus, R., & Lusa, L. (2013). Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(106). https://doi.org/10.1186/1471-2105-14-106

Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy* (Vol. 9). Now Publishers. https://doi.org/10.1561/0400000042

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

National Institute of Standards and Technology. (2026a). *Ai rmf core*. Retrieved August 17, 2026, from https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

National Institute of Standards and Technology. (2026b). *Trustworthy and responsible ai*. Retrieved August 17, 2026, from https://www.nist.gov/trustworthy-and-responsible-ai

Papernot, N., McDaniel, P., & Goodfellow, I. (2017). Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. https://doi.org/10.1145/3052973.3053009

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

Sakho, A., Malherbe, E., & Scornet, E. (2024). Do we need rebalancing strategies? a theoretical and empirical study around smote and its variants.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the theory of games ii* (pp. 307–317). Princeton University Press.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*.

Tabassi, E. (2023). *Artificial intelligence risk management framework (ai rmf 1.0)* (tech. rep. No. NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

U.S. Food and Drug Administration. (2024). *Transparency for machine learning-enabled medical devices: Guiding principles*. Retrieved August 17, 2026, from https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles

Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (tech. rep. No. NIST AI 100-2e2025). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2025