MWSUG 2025 - Paper AE62

SAS Trustworthy AI Examples

Caleb Petterson, SAS Institute Inc., Minneapolis, MN

ABSTRACT

The accelerating adoption of artificial intelligence (AI) brings significant risks related to model opacity, unfair outcomes, and fragility in production. To address these challenges, we propose a scenario-based framework for implementing Trustworthy AI principles. This paper demonstrates the construction of 15 end-to-end machine learning pipelines across education, finance, healthcare, and public sectors, each targeting a specific challenge within the core disciplines of explainability, fairness, and robustness. We illustrate how techniques such as surrogate models, SHAP, fairness constraints, adversarial training, and drift detection can be systematically evaluated and applied. Our findings indicate that the effectiveness of a Trustworthy AI technique is highly context-dependent, and we provide a practical blueprint for organizations seeking to operationalize responsible AI using SAS® software.

INTRODUCTION

Trustworthy AI (TAI) defines systems that are transparent, fair, and robust. As AI systems are deployed in increasingly critical domains, the need for TAI has become a business and ethical imperative. However, a significant gap exists between the theoretical principles of TAI and their practical implementation. This paper addresses this gap by introducing a framework for building and evaluating TAI pipelines using artificial scenarios as controlled testbeds. We will detail our approach of constructing 15 distinct pipelines to stress-test techniques across the three pillars of explainability, fairness, and robustness. The primary goal is to provide a practical, repeatable methodology for SAS® users to integrate TAI practices into their own workflows.

A FRAMEWORK FOR TRUSTWORTHY AI: THE THREE PILLARS

TAI is a multidisciplinary framework that cannot be reduced to a single feature. We dissect three of its core pillars, each addressing a critical dimension of trust (Molnar, 2022). Explainability ensures that model decisions can be understood by human stakeholders, answering the "why" behind a prediction. Fairness involves measuring and mitigating unwanted biases to ensure equitable outcomes across different demographic groups (Barocas et al., 2019). Robustness ensures models remain reliable and secure when faced with real-world challenges like adversarial attacks, data drift, and noisy data (Goodfellow et al., 2014). This framework provides the structure for our scenario-based analysis.

METHODOLOGY: ARTIFICIAL SCENARIOS AS TESTBEDS

To systematically evaluate Trustworthy AI techniques, we developed a series of artificial scenarios. These are end-to-end machine learning pipelines designed to simulate real-world problems in a controlled environment, similar to a wind tunnel for testing aircraft.

- Scenario Design: Each scenario is built on a public dataset and isolates a specific trustworthiness challenge:
    1. Students Dropout and Academic Success Dataset
    2. German Credit (Statlog) Dataset
    3. CDC Diabetes Dataset
    4. Heart Disease Dataset
    5. Adult Income Dataset
- The 15 pipelines are organized by industry and discipline, as shown in Table 1.
- All pipelines were constructed using SAS® software, leveraging procedures and custom code for data preparation, model training, bias mitigation, and explainability analysis (SAS Institute Inc., 2023a).

Industry	Explainability Scenario	Fairness Scenario	Robustness Scenario
Education	Distilling Dropout Rules (**Surrogate Trees**)	Mitigating **International Status** Bias in Scholarships (**Demographic Parity**)	Adapting to Curriculum Drift (**AdaBN**)
Finance	Visualizing Loan Amounts (**PDP/ICE**)	Mitigating **Age** Bias in Credit Scoring (**Equalized Odds**)	Strengthening Loans (**Adversarial Training**)
Healthcare (diabetes)	Explaining Risk (**SHAP**)	Mitigating **Income** Bias in Screening (**Demographic Parity**)	Enhancing Diabetes Risk Predictions (**Adaptive Imputation, MC Dropout**)
Healthcare (heart disease)	Interpreting Diagnosis (**LIME**)	Mitigating **Sex** Bias in Diagnosis (**FNR Parity**)	Improving Heart Disease Predictions (**Domain Adaptation**)
Public	Uncovering Income Logic (**RuleFit**)	Mitigating **Racial** Bias in Employment (**FNR Parity**)	Safeguarding Income Classification (**Noise Correction**)

Table 1. Taxonomy of Trustworthy AI Scenarios by Industry and Discipline

Experimental Application and Results

This section details the application of our framework, including how techniques were selected and evaluated within scenarios.

Applying Explainability for Black-Box Scenarios

Explainability techniques provide a lens into the model's decision-making process. Their utility varies depending on the stakeholder's need for global model understanding versus local explanations.

Global Explainability with Surrogate Models

In the student dropout scenario, a complex random forest model was approximated using a shallow decision tree surrogate. This technique distilled the model's logic into simple, actionable "if-then" rules that academic advisors could use directly, translating opaque prediction weights into an intervention plan.

Local Explainability with SHAP and LIME

For clinical risk prediction in healthcare, local techniques with SHAP (SHapley Additive exPlanations) quantified the contribution of each feature (e.g., glucose level, BMI) to an individual patient's diabetes risk score, providing clinicians with auditable justifications to foster trust in model predictions.

Ensuring Fairness with Context-Aware Mitigation

Our experiments demonstrated that the choice of fairness technique must be guided by the specific type of bias and the desired fairness metric. We applied techniques across the pre-, in-, and post-processing spectrum, with tools like Fairlearn (Bird et al., 2020).

Pre-processing with Correlation Removal

To prevent a model from leveraging proxies for sensitive attributes, we used correlation removal techniques to sanitize training data. In the public sector scenario for employment services, this helped reduce the model's reliance on features correlated with race.

In-processing with Exponentiated Gradient

In scenarios where a fundamental fix was needed, such as reducing false negative rates (FNR) for female patients in heart disease diagnosis, we applied in-processing techniques like an exponentiated gradient algorithm. This method directly optimizes for fairness constraints during model training and creates a more inherently fair model.

Post-processing with Threshold Optimization

For credit scoring, where the primary concern was balancing error rates across age groups, we achieved equalized odds by applying post-processing threshold optimization. This technique adjusts decision thresholds for different groups without altering the underlying model scores.

Engineering Robustness against Real-World Challenges

Robustness techniques ensure model performance endures over time. We focused on domain generalization and adversarial robustness.

Adaptation to Drift with Adaptive BatchNorm

In the education scenario, we simulated curriculum drift, where a model trained on historical student data becomes obsolete. Using performance monitoring to detect decay, we applied adaptive BatchNorm (AdaBN) to rapidly recalibrate the model's internal statistics on new data, thereby restoring accuracy without a full retraining.

Generalization with Adversarial Training

For loan amount prediction, we hardened models against manipulation using projected gradient descent (PGD) adversarial training (Goodfellow et al., 2014). This involved challenging the model during training with perturbed inputs, making it resistant to attempts by applicants to exploit the system for inflated loan offers.

Discussion

The results from our 15 pipelines lead to several key insights and considerations for practitioners.

The Importance of Context

The most significant finding is that there is no single "best" Trustworthy AI technique. The optimal choice is highly context-dependent. For example, post-processing was effective for achieving equalized odds in credit scoring, but in-processing was more appropriate for addressing the critical risk of false negatives in medical diagnosis. The stakeholder's need also dictates the choice of explainability technique.

Implications for Practitioners

The scenario-based framework provides a practical blueprint for organizations. We recommend that teams looking to operationalize TAI start by identifying their highest-risk failure modes and then use a similar testbed approach to evaluate and validate mitigation strategies before full-scale deployment.

Conclusion

This paper presents a practical, scenario-based framework for implementing TAI. By constructing and analyzing 15 pipelines, we have demonstrated that techniques for explainability, fairness, and robustness are most effective when selected based on the specific context and stakeholder requirements. The framework provides a blueprint for SAS® users to systematically integrate these disciplines into their AI lifecycle, moving from principles to practice.

REFERENCES

SAS Institute Inc. (2023). Trustworthy AI blog series. SAS Blogs. Retrieved August 22, 2025, from https://blogs.sas.com/content/tag/trustworthy-ai-toolkit/

SAS Institute Inc. (2023). The SAS® trustworthy AI life cycle. Cary, NC: SAS Institute Inc. Retrieved August 22, 2025, from https://github.com/sassoftware/sas-trustworthy-ai-life-cycle

SAS Institute Inc. (2023). SAS® Viya® platform: Programming documentation. Cary, NC: SAS Institute Inc. https://github.com/sassoftware/sas-viya-workbench-examples

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI (MSR-TR-2020-32). Microsoft Research. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Molnar, C. (2022). Interpretable machine learning (2nd ed.). Retrieved September 9, 2025, from https://christophm.github.io/interpretable-ml-book/

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Retrieved September 9, 2025, from https://fairmlbook.org/

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv Preprint, arXiv:1412.6572. https://arxiv.org/abs/1412.6572

ACKNOWLEDGMENTS

The authors would like to thank the Trustworthy AI team in the Data Ethics Practice division at SAS® for their collaboration and insights in developing the scenarios and techniques discussed in this paper.

RECOMMENDED READING

- *SAS® Trustworthy AI Examples (https://github.com/sassoftware/sas-trustworthy-ai-examples)*
- *SAS® Trustworthy AI Life Cycle (https://github.com/sassoftware/sas-trustworthy-ai-life-cycle)*
- *Trustworthy AI Blog Series (https://blogs.sas.com/content/tag/trustworthy-ai-toolkit/)*
- *Fairlearn (https://fairlearn.org/)*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Caleb Petterson
University of Minnesota Twin Cities
612-443-7314
pette184@umn.edu
https://www.linkedin.com/in/caleb-petterson-494098290/