feedzai

Responsible AI & The TRUST Framework

Move fast and FIX things:
Build AI that people TRUST

Pedro Bizarro
Co-Founder and
Chief Science Officer

feedzai

3 Reasons

Why talk about TRUST in AI?

Reason 1

The Market wants Trustworthy AI

75% of US and UK consumers would not purchase from organizations they don't trust with their personal data.

BusinessWire, 75% of US & UK consumers are not comfortable purchasing from brands with poor data ethics

82% of American consumers are concerned about AI's ethical implications.

Markkula Center for Applied Ethics, Ethics in the Age of AI

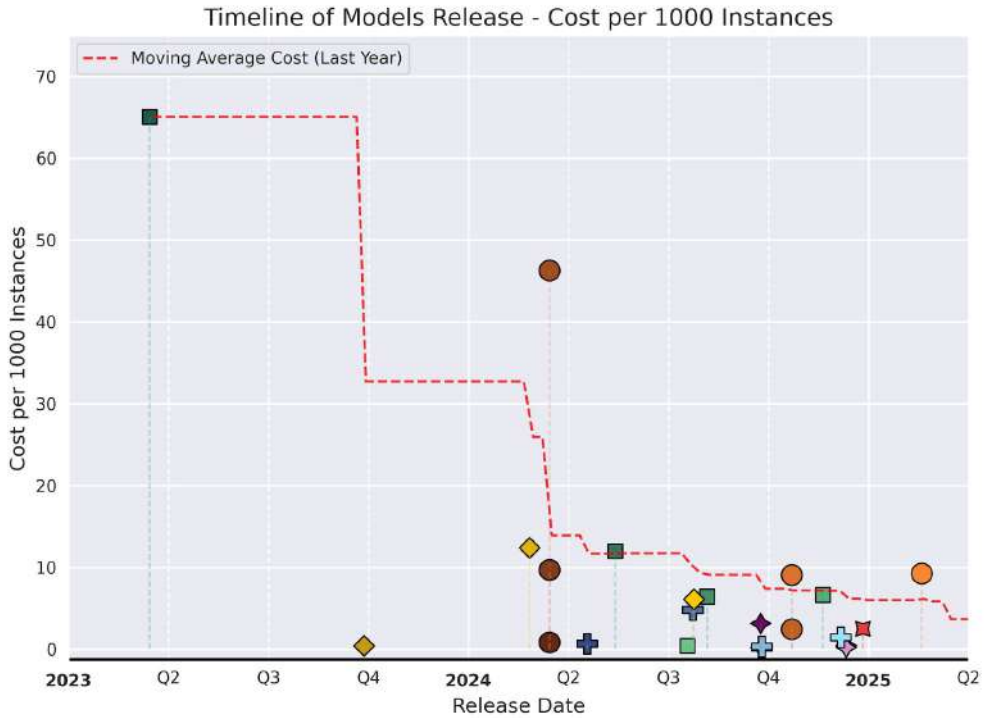
56% of Fortune 500 companies cite AI as a risk factor in their annual reports.

Accenture, Thrive with responsible AI: Embedding trust can unlock value

Reason 2

AI is commodity. TRUST is king.

150x cost reduction in 2 years
[our own benchmarks]



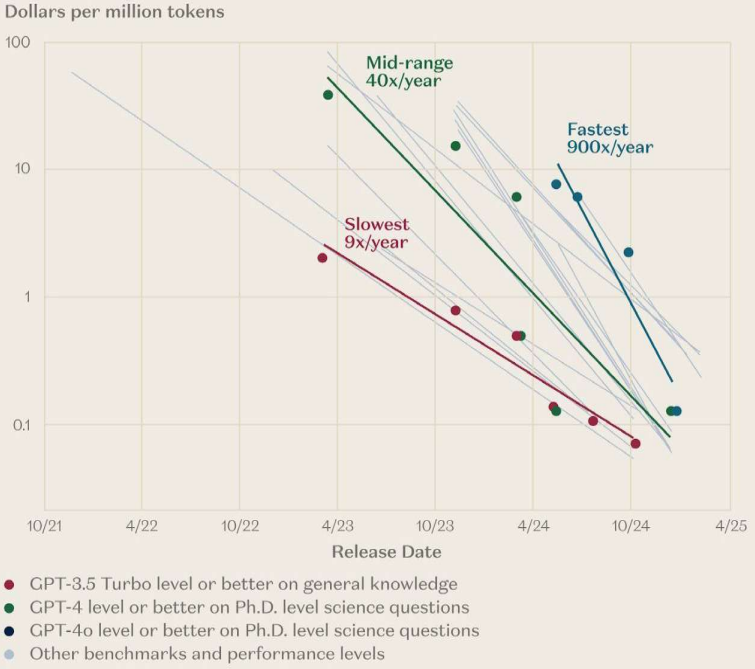
Reason 2

AI is commodity. TRUST is king.

9x to 900x cost reduction **per year**
[By Epoch AI]

The Cost of Doing Business

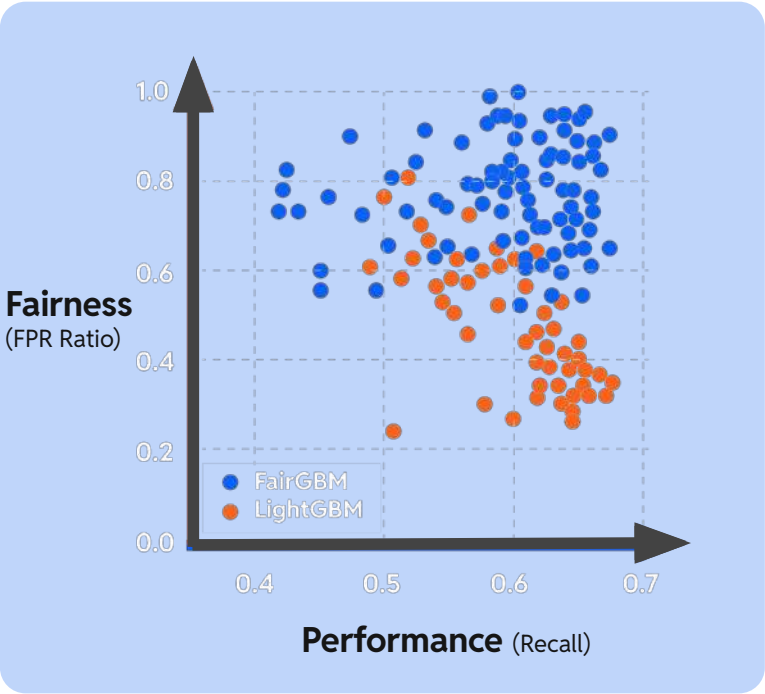
The price per token (for prompts and responses) for AI models at a given level of intelligence. The least "intelligent" models showed a roughly 9x decrease in cost per year, while the most capable ones dropped in price by roughly 900x per year.



Note: Data are shown on a logarithmic scale.
Source: Epoch AI

Reason 3

We can have TRUST *and* Performance



For a long time it was assumed
that you had to choose between, e.g.,
either fairness **or** performance,
but that is not the case anymore.

feedzai

How did we get to the TRUST framework?

Client needs and client problems with competitors, incumbents, or in-house solutions

"It's all black-box. We don't know **how it was trained or why the scores are generated**."

"Model is trained with old data, **degrades quickly**, it's legacy before it goes live."

"The **model is much worse in some groups** than in others, and we can't fix it."

"We delete data that could be useful; we are **afraid of data leaks & social engineering**."

"We create good features & models, but they **don't perform the same in production**."

The TRUST Framework

feedzai

Building Trust in High-Performance AI

Responsible AI **enhances value**: not a loss, but an **investment**.

Responsible AI **safeguards users** and **protects brand reputation**.

Trust in AI systems must be **intentionally built from the start**.

Before considering the TRUST Framework...

... is AI even necessary for your use case?

Genuine Necessity

Does the problem genuinely **require an AI solution**?

Could a different, or simpler, human-centered approach achieve the desired outcome responsibly?

Cost-Benefit Analysis

Are potential **benefits** of deploying AI clearly **greater than the risks**?

Are there dangers of over-automation and unintended consequences such as negative feedback loops?

The TRUST Framework

feedzai

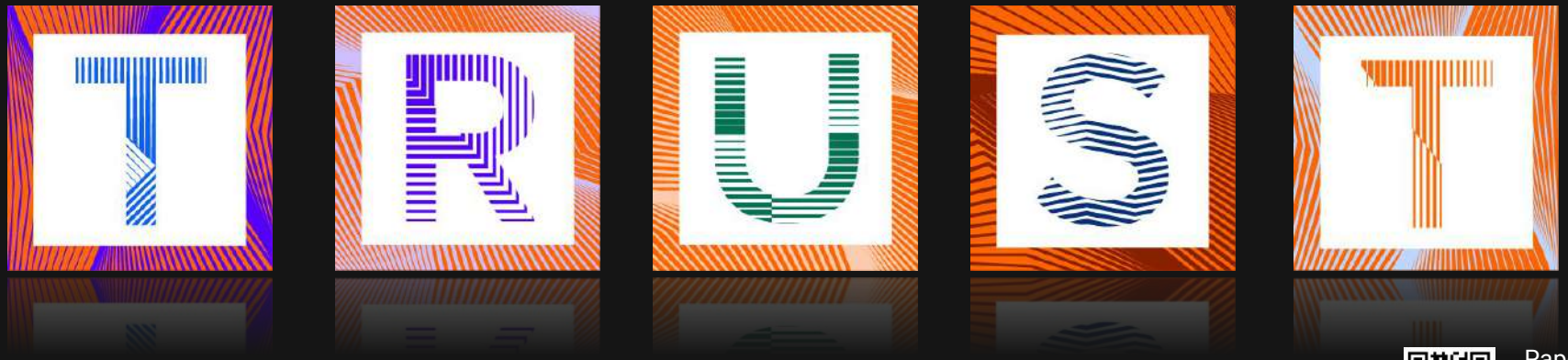
Transparent

Robust

Universal

Secure

Tested



Papers
Videos
Code

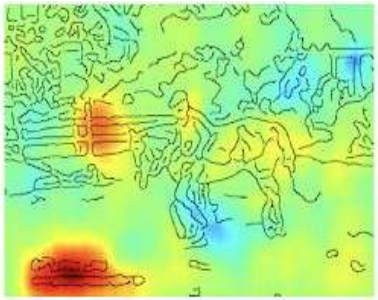
feedzai

Transparent: Motivation

Helps reveal if you can trust the AI system

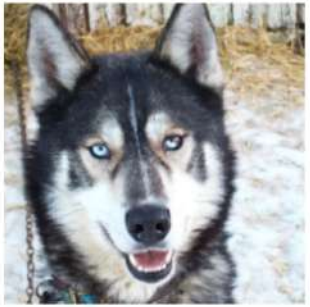
Model **correctly** classified image as a horse
by hidden mistake

Input image



Red pixels are the most important for the model

Model **incorrectly** classified husky as a wolf
by hidden mistake



Input image

Most important pixels for "wolf" misclassification

These models learned spurious correlations (photographer copyright, or the existence of snow)

Montavon, G., Samek, W., & Müller, K. R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 2018
Ribeiro, M. T., Singh, S., & Guestrin, C. Why should I trust you?" Explaining the predictions of any classifier. *ACM SIGKDD* 2016

The TRUST Framework

Transparent

Themes and Elements of Transparent AI

Explainable

Explanations

Reveal the key factors or attributes used by models to reach predictions.

Human-AI Control

Allow certain users to interact, understand, and adjust AI decisions.

AI Presence

Indicate clearly when users are engaging with an AI system.

Governance

Model Governance

Document the model's type, architecture, hyperparameters, configuration, performance, and compute needs.

Data Provenance

Maintain clear records of data origin, selection, cleaning, transformation, and sampling.

Process Governance

Provide policies and processes that guide development, deployment, and use of AI systems, including ethical, regulatory, and accountability considerations.

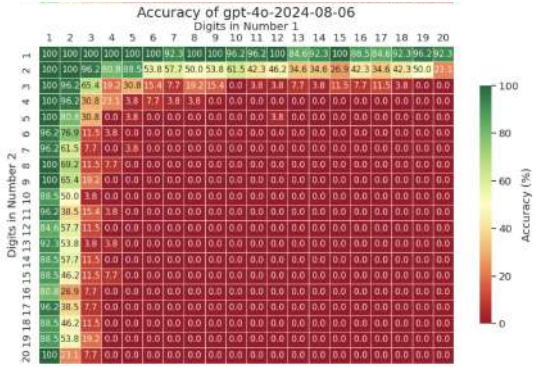
feedzai

Robust: Motivation

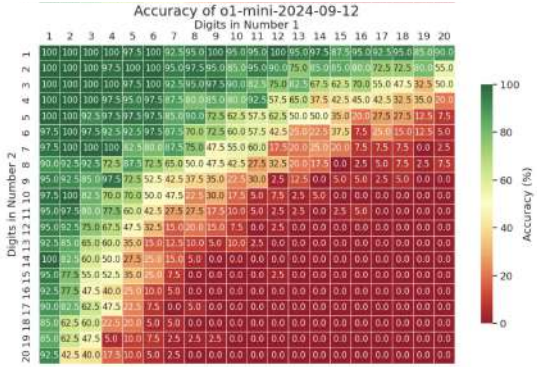
Can we trust an AI system if it is not robust?

Example: Are OpenAI's models good at multiplying two numbers?

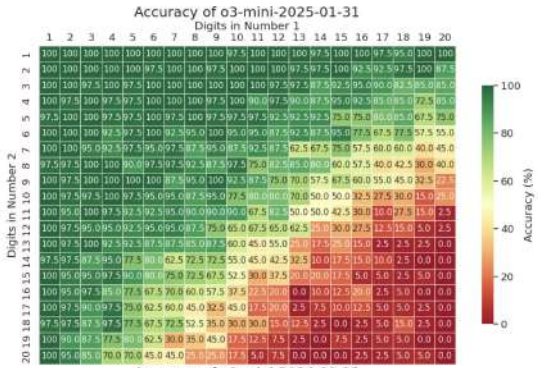
gpt-4o struggles beyond 4x4



o1-mini struggles beyond 7x7



o3-mini struggles beyond 10x10



Yuntian Deng (@yuntiandeng) on X

feedzai

Robust: Motivation

Can we trust an AI system if it is not robust?

How many R's are there in Ruberrberrry?

Attempt 1

Let's count them:

1. Ruberrberrry
2. Ruberrberrry
3. Ruberrberrry

There are a total of 5 "R"s in "Ruberrberrry".

Attempt 2

Let's count them:

1. Ruberrberrry

There are four instances of the letter "R".

Attempt 3

Let's count them:

1. Ruberrberrry

There are six "R"s in "Ruberrberrry".

Run on ChatGPT 4o on March 4, 2025

The TRUST Framework

Robust

Themes and Elements of Robust AI

Consistent

Performance is stable, delivering accurate, reproducible and consistent results aligned with training and avoiding erratic behavior.

Adaptable

Tolerates changes in data and environmental conditions, handling unexpected or adversarial inputs with minimal disruption.

The TRUST Framework

Universal: Motivation

Possible high negative impact

POLICY

AI landlord screening tool will stop scoring low-income tenants after discrimination suit / Screening system SafeRent will also pay around $2.3 million to Massachusetts residents with housing vouchers as part of a class action settlement.

The Verge, 11/20/2024

Upstart says it's improving AI models after report finds race approval disparities

Nat. Mortgage News, 09/26/2024

Revealed: bias found in AI system used to detect UK benefits fraud

Exclusive: Age, disability, marital status and nationality influence decisions to investigate claims, prompting fears of 'hurt first, fix later' approach

The Guardian, 12/06/2024

Social Media Tells You Who You Are. What if It's Totally Wrong?

Wired, 10/11/2024

BIASED TECHNOLOGY: THE AUTOMATED DISCRIMINATION OF FACIAL RECOGNITION

ACLU, 02/20/2024

feedzai

Universal: Motivation

Complex problem with no magic solution

Sources of Bias...

... include **data**, **cleaning**, **sampling**, **transformations**, **labels**, **training choices**, **humans-in-the-loop**, **post-processing**, and more.

Intentions not enough

If you are not controlling for bias, it is **almost inevitable** that you will get a **biased AI system**.

Simple Solutions?

Eliminating protected attributes doesn't solve the problem. It makes it worse, because it prevents us from measuring bias.

The TRUST Framework

Universal: Motivation

Must be educated and make difficult choices

Incompatibilities in Fairness

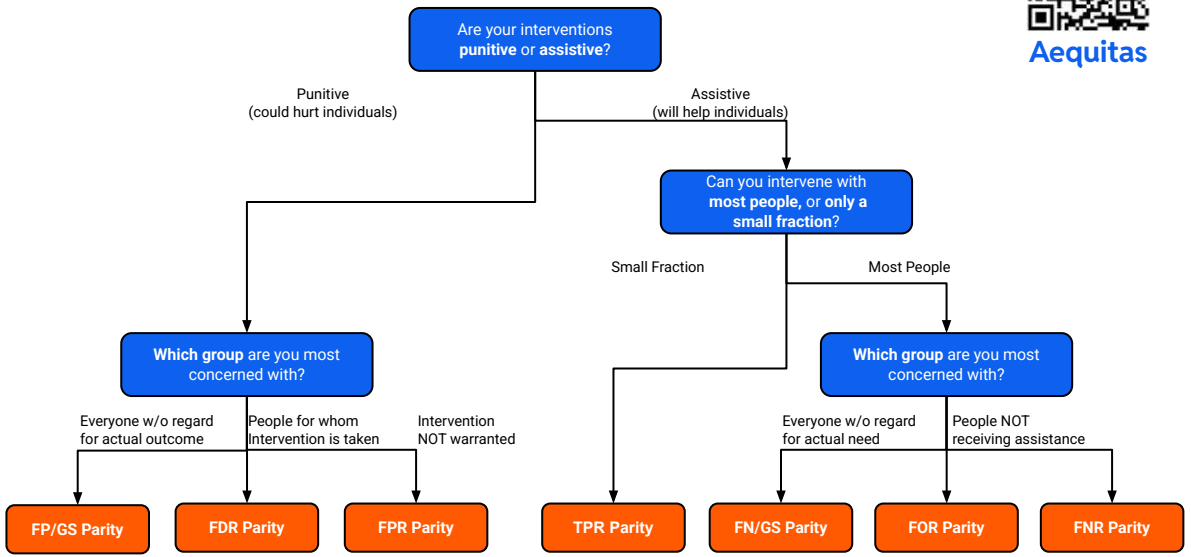
It is not possible to optimize for **multiple fairness metrics at once** in settings with different prevalences (*e.g.*, disease rates in different age groups).

Need to **choose one fairness metric, depending on the type of problem**.

Use the Fairness Tree to help choose the most appropriate fairness metric.

Aequitas Fairness Tree

Aequitas



P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. Rodolfa, R. Ghani "Aequitas: A bias and fairness audit toolkit."

The TRUST Framework

Universal

Themes and Elements of Universal AI

Fair

Equitable Outcomes

Reduce disparate treatment by not discriminating against groups based on protected attributes.

Regular Audits

Conduct regular audits to detect biases and validate bias mitigation strategies.

Inclusive

Inclusive Development

Involve diverse teams and stakeholder input to integrate multiple perspectives on fairness.

Designed for Everyone

Designed and accessible to everyone including people with disabilities.

The TRUST Framework

Secure: Motivation

Data leaks, System malfunctions

Generative AI Under Attack: Flowbreaking Exploits Trigger Data Leaks

Forbes, 11/26/2024

Study: 77% of Businesses Have Faced AI Security Breaches

AI systems are particularly vulnerable to security breaches, which is why shoring up your defenses is key in 2024.

Tech.co, 03/22/2024

Sensitive DeepSeek data exposed to web, cyber firm says

Reuters, 01/30/2025

The TRUST Framework

Secure

Themes and Elements of Secure AI

Data Protection

Data Privacy

Handle sensitive data with strict confidentiality, complying with privacy regulations.

Data Consent

Provide clear processes for data owners to control how their information is used or deleted.

Data Integrity

Ensure that data is authentic and correct, and that is not lost, nor tampered with.

System Protection

System Security

Run audits, penetration tests or vulnerability scans to guard against unauthorized access and to comply with industry or regulatory controls.

System Availability

System, services, data, and information are monitored and always available to authorized parties, ensuring correct and secure usage.

feedzai

Tested: Motivation

Was the AI system tested for these use cases?

Pricey AI Pins are being returned quicker than they are sold: 'It just doesn't work'

New York Post, 08/09/2024

AI has high data center energy costs — but there are solutions

MIT Management, 01/07/2025

Lawyer Used ChatGPT In Court —And Cited Fake Cases. A Judge Is Considering Sanctions

Forbes, 01/08/2023

NEWS

Misinformation researcher admits ChatGPT added fake details to his court filing / Jeff Hancock says he used GPT-4o to help with citations – and didn't realize the tool 'hallucinated' new ones.

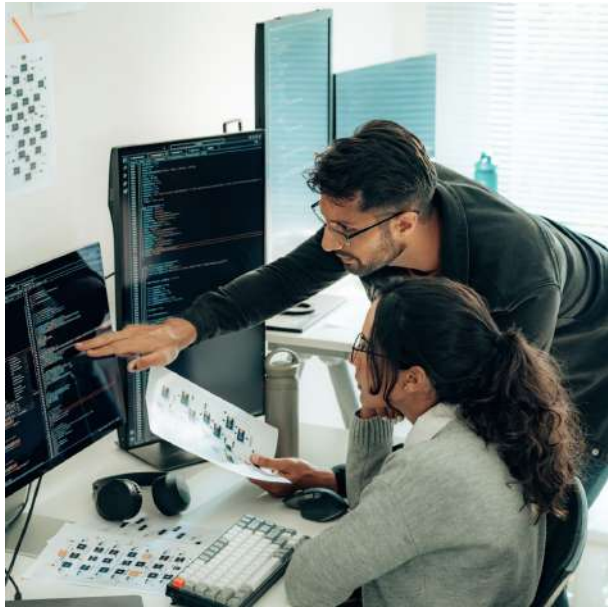
The Verge, 12/04/2024

Tested: Motivation

Real-world fraud detection explanations

Experiment: Test decision accuracy and response time of fraud analysts with progressively more information:

1. **Transactional Data Alone**
2. **Transactional Data + AI Model Score**
3. **Transactional Data + AI Model Score + Explanations**



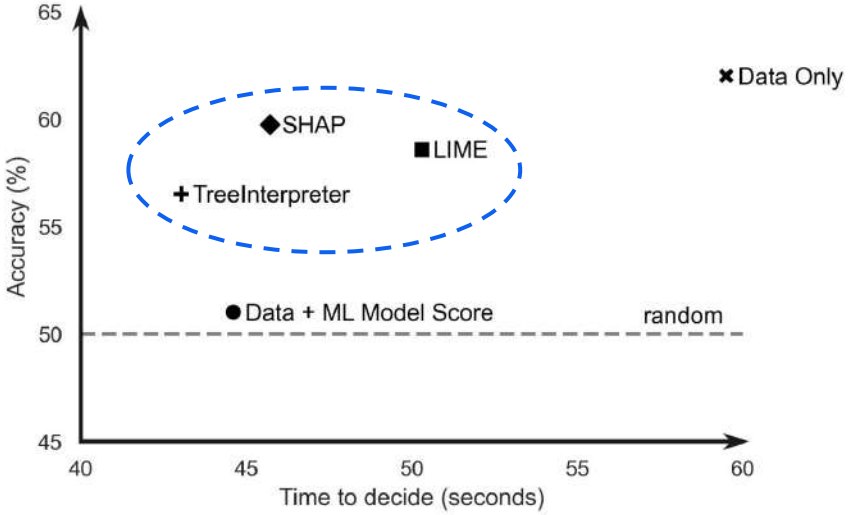
Jesus et al. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. *ACM FAccT 2021*

Tested: Motivation

Real-world fraud detection explanations

Key Finding: real-world fraud analysts, **responded faster, but with lower accuracy** when given algorithmic explanations compared to when given transactional data only.

The true value of product features is determined by real-world performance, not just theoretical benefits.



The TRUST Framework

Tested

Themes and Elements of Tested AI

Silent bugs | Heisenbug | Data leakage | Data misses

User Aligned

Intended Uses

Monitor and test to confirm that system decisions align with user intended outcomes in cases the system was designed for.

Societal Benefit

Design with societal benefit in mind, considering a wide range of factors and anticipating the impact on all stakeholders

Complete Testing

Integration Testing

Test considering **all system components**, including models, settings, rules, lists, explanations, other systems, and user interactions.

Continuous Testing

Continuously monitor and test the entire development lifecycle to prevent system or data issues to become AI issues in **extreme cases**.

Cost Efficiency

Minimize time, energy, and computational resources while considering the full lifecycle cost and prevent "cost leaks."

Oversight and Assessment

Take advantage of independent evaluations to learn and confirm adherence to best practices.

The TRUST questionnaire

TRUST Framework v2025.09

feedzai

Transparent

Questions to evaluate Transparent AI

- ☑️ Are AI decisions explained clearly and understandably, including in generative models?
- ☑️ Is AI presence clearly disclosed, especially when users interact with AI-generated content?
- ☑️ Can users understand and, where appropriate, influence AI outcomes?
- ☑️ Are there established policies for responsible AI development and deployment?
- ☑️ Are model design choices and data processes fully documented and accessible?

  (e.g., via a model cards, identifying the model purpose, intended users, data provenance, sampling, filtering, performance metrics, post-training decisions, known limitations, ethical considerations?)

feedzai

Robust

Questions to evaluate Robust AI

- ☑️ Does the system maintain reliable performance across varying conditions?
- ☑️ Are results stable and predictable over time?
- ☑️ Are stress tests and adversarial simulations performed before deployment?
- ☑️ Is there a structured process for monitoring and correcting performance degradation?
- ☑️ Can the system detect and respond to data drift, anomalies, or malicious activity, including prompt-based misuse?

feedzai

Universal

Questions to evaluate Universal AI

- ☑️ Are fairness goals clearly defined and embedded in model development?
- ☑️ Are regular bias audits conducted?
- ☑️ Are there documented strategies to mitigate unfair or harmful outcomes?
- ☑️ Was the AI developed with diverse perspectives?
- ☑️ Is the system accessible, inclusive, and equitable across demographics, languages, and abilities, including in AI-generated content?

feedzai

Secure

Questions to evaluate Secure AI

- ☑️ Are regular security audits and penetration tests conducted, including 3rd-party risks from integrations?
- ☑️ Is the system continuously monitored for security threats, including prompt injection?
- ☑️ Does the system comply with data privacy regulations (eg, GDPR, CCPA, industry-specific standards)?
- ☑️ Do users have control over their data, including consent and deletion rights?
- ☑️ Is sensitive data protected through encryption, access controls, and formal agreements with 3rd-party AI providers?

feedzai

Tested

Questions to evaluate Tested AI

- ☑️ Has the system undergone thorough user acceptance and integration testing, covering inputs, prompts, or edge cases?
- ☑️ Is the AI deployed strictly within its validated scope, with safeguards to ensure system boundaries are upheld?
- ☑️ Does testing confirm that the AI system enhances decision-making and delivers measurable value?
- ☑️ Are shadow environments used for A/B testing and validation before full deployment?
- ☑️ Are performance, efficiency, third-party integrations, and harmful content regularly benchmarked and monitored post-deployment?

feedzai

TRUST work at Feedzai

Selected publications and projects

The TRUST Framework

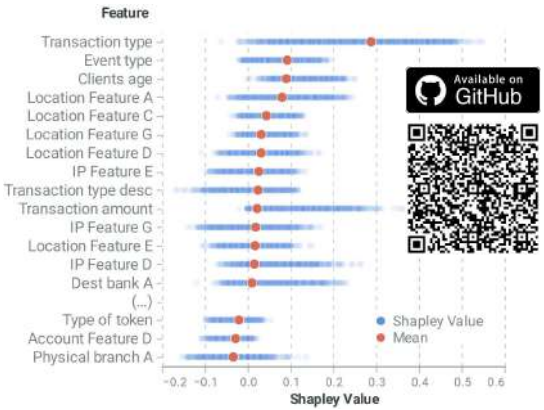
Transparent

Learn about Feedzai Research

Transparent AI @ Feedzai

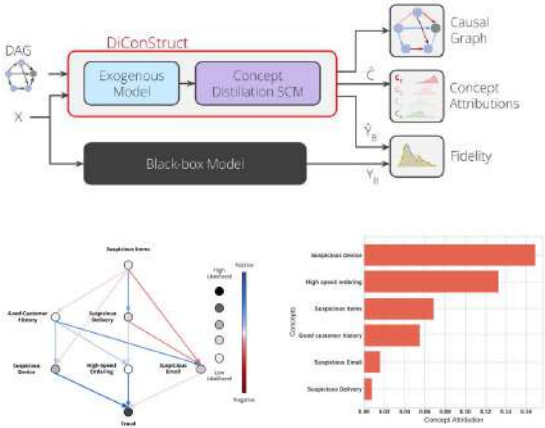
TimeSHAP

Explaining Recurrent Deep Learning Models



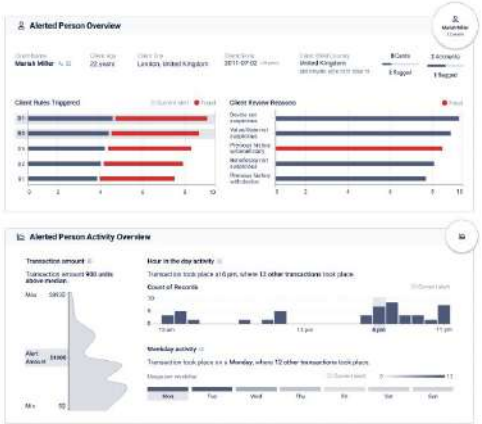
DiConStruct

Causal Explanations for Black-Box Models



Show Me What's Wrong!

Charts and Text to Guide Data Analysis



Bento et al. TimeSHAP: Explaining recurrent models through sequence perturbations. *SIGKDD* 2021
Moreira et al. DiConStruct: Causal Concept-based Explanations through Black-Box Distillation. *CLeaR* 2024
Feliciano et al. "Show Me What's Wrong!": Combining Charts and Text to Guide Data Analysis. *NLVIZ/IEEE VIS* 2024

feedzai research

The TRUST Framework

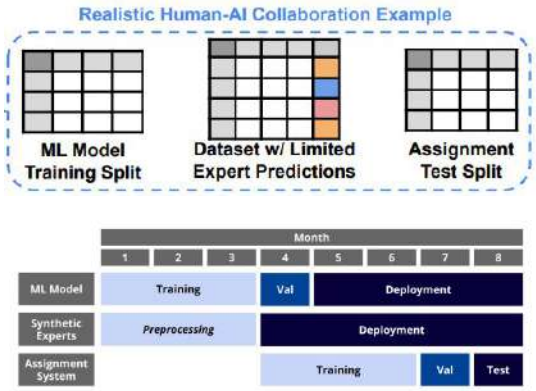
Robust

Learn about
Feedzai Research

Robust AI @ Feedzai

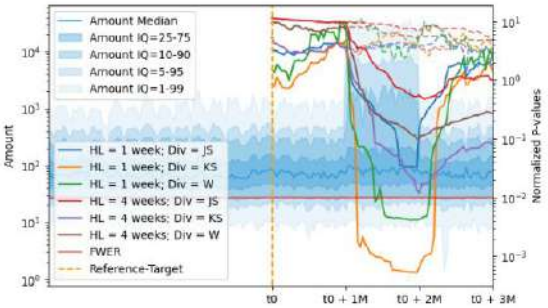
FiFAR

Learning to Defer in Fraud Detection with AI



Feature Monitoring

Lightweight Automation for Data Streams



Data+Shift

Visual Investigation of Data Drift

Alves et al. FiFAR: A fraud detection dataset for learning to defer. *ICAIF 2023.*
Conde et al. Lightweight Automated Feature Monitoring for Data Streams. *KDD 2022.*
Palmeiro et al. Data+Shift: Supporting visual investigation of data distribution shifts by data scientists. *EuroVis 2022.*

feedzai
research

The TRUST Framework

Robust

Robust AI @ Feedzai

A Good Example Supporting Robust Model Design

Challenge: Few high-quality, freely available tabular datasets exist for AI stress-testing, especially in sensitive areas such as fraud detection.

Solution: Feedzai developed the **open source** **Bank Account Fraud** (BAF) Tabular Dataset: a large, set of synthetic bank transactions reflecting natural behavior variations, seasonal shifts, and activity spikes.

Impact: It helps researchers test fraud models in real-world conditions, producing more robust models.

Proof-point: Feedzai's models maintained consistent performance even during major disruptions, such as the early COVID-19 pandemic.

The TRUST Framework

Universal

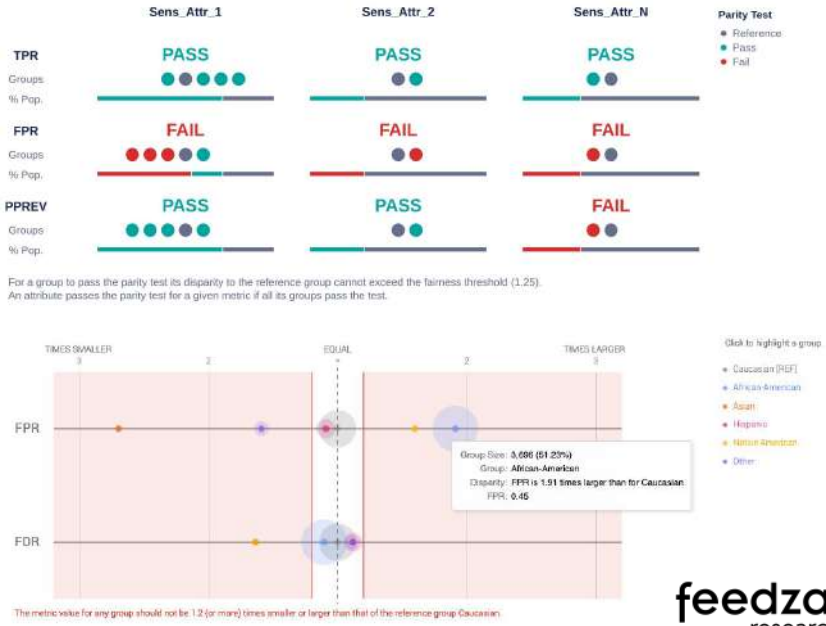
Universal AI @ Feedzai

Example of Bias Mitigation at Feedzai

To determine whether an AI system is unbiased, the first step is to conduct a **bias audit**.

Aequitas Flow (Feedzai's open-source bias audit tool contribution) calculates various fairness metrics across groups defined by protected attributes, identifying any performance gaps.

Available on GitHub



feedzai research

The TRUST Framework

Universal

Universal AI @ Feedzai

Example of Bias Mitigation at Feedzai

Several bias mitigation techniques are applied at Feedzai, including **pre-processing** techniques:

- **Diverse sampling:** Ensure all groups are well-represented in data.
- **Balancing data:** Oversample underrepresented groups or undersample overrepresented ones.
- **Data augmentation & reweighting:** Use techniques like Feedzai's **Fair-OBNC** to enhance fairness.

Learn more about Fair-OBNC

feedzai research

The TRUST Framework

Universal

Universal AI @ Feedzai

Example of Bias Mitigation at Feedzai

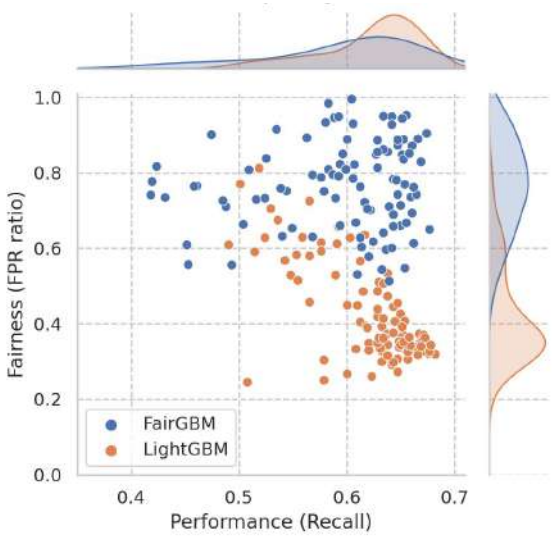
...through **in-processing**...

- **Built-In Fairness:** Use fairness-aware models, such as Feedzai's **FairGBM**, to inject fairness regularization in training, and strike a balance across accuracy and fairness.

Available on GitHub

- **Fairness-aware Hyperparameter Optimization:** optimize the fairness-accuracy trade-offs with model agnostic Feedzai's **Fair-AutoML** algorithms

Available on GitHub



feedzai research

The TRUST Framework

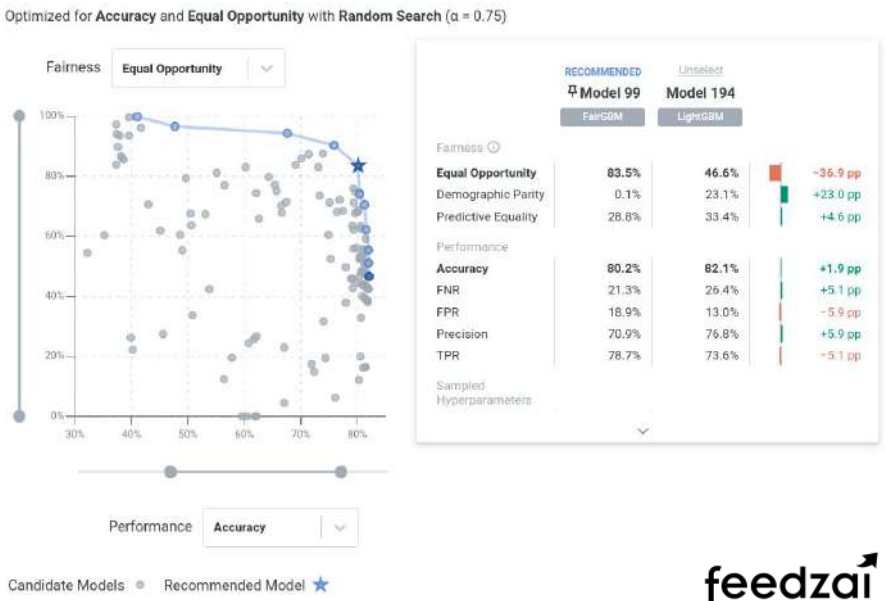
Universal

Universal AI @ Feedzai

Example of Bias Mitigation at Feedzai

...to **post-processing**...

- **Fairness-Aware Tuning:** Visualize and select the model that strikes a good balance between performance and fairness, e.g., using Feedzai's **Fairband** *(Fintech Breakthrough Awards, 2021)*



The TRUST Framework

Secure

Secure AI @ Feedzai

Feedzai continuously runs audits, penetration tests, and vulnerability scans

594

industry standard controls implemented & monitored, for **operational, technical & managerial** scopes

+50

client **audits**, **penetration tests,** or **vulnerability scans** per year

3

yearly independent audits (**PCI DSS**, **SOC 2 Type II** and **ISO 27001**)

The TRUST Framework

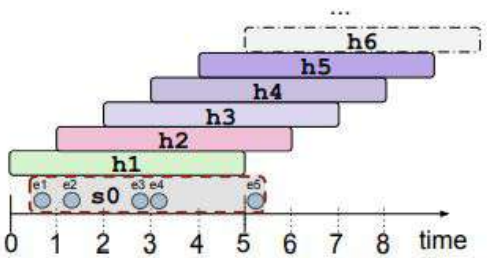
Tested

Learn about Feedzai Research

Tested AI @ Feedzai

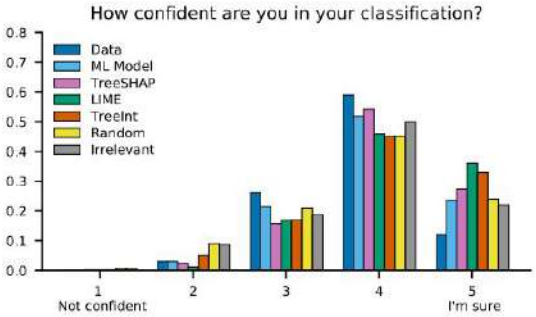
Railgun

Streaming Windows Under Strict Requirements



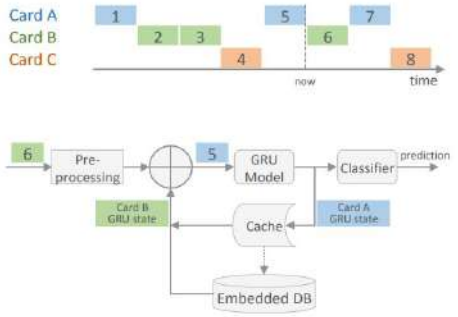
Xplainzai

App-Grounded Evaluation of Explainers



Interleaved RNNs

for Sequential Fraud Detection



Gomes et al. Railgun: managing large streaming windows under MAD requirements. *VLDB 2021*
Jesus et al. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. *ACM FAccT 2021*
Martin et al. A Case Study on Designing Evaluations of ML Explanations with Simulated User Studies. *ICLR, 2023*
Branco et al. Interleaved sequence RNNs for fraud detection. *ACM SIGKDD 2020*

feedzai research

The TRUST Framework

Tested

Tested AI @ Feedzai

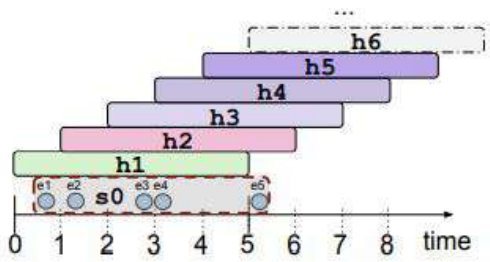
Real-world Scenario: Streaming Windows

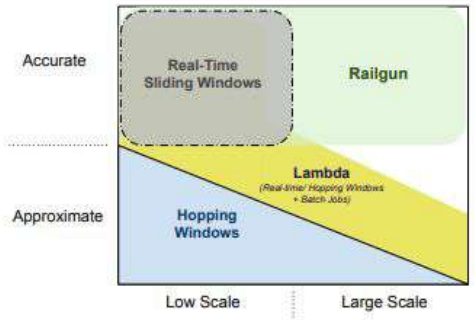
Feedzai Railgun: High-throughput, real-time fraud detection with strict accuracy & latency requirements

Objective: Build a truly scalable, low-latency streaming system with strict accuracy that works in mission critical environments

Experiment: Benchmark the system using real data under extreme loads of **up to 1M events/sec**

Key Finding: Scores under stress at **millisecond-level latencies** (<250ms @ 99.9%) with **low memory use**, scaling nearly **linearly**.





Gomes et al. Railgun: managing large streaming windows under MAD requirements. *VLDB 2021*

feedzai research

feedzai

Move fast and FIX things:
Build AI that people TRUST

Thank you!

Papers
Videos
Code