

AI Testing and Assurance

Bryan Jones

QAT Architect
2i



How do we
test AI?



Challenges

- The Oracle Problem
 - No Expected Results
 - Non-Deterministic
 - Probabalistic Answers
- The code doesn't represent the algorithm
- Data Complexity & Volume
- Self-Optimising
- System Complexity
- Attempting to Mimic Human Abilities
- Bias
- Hallucinations/Confabulation
- Test Coverage?



We test.
You impress.

Test Levels

Familiar

- Unit or Component testing,
- Integration testing
- End-to-End testing
- UAT

New

- Model Testing
- Data Testing
- Monitor in live



We test.
You impress.

Component Test

“Automation without Requirements”

Forced
Zeros

Blanks &
Nulls

Out of
Range
Values

Formatting
Errors

Duplications

Data Test

“The code is not the algorithm,
The data defines the behaviour”

- Bias & Fairness
 - Don't forget Proxies
- Data Diversity and Representativeness
- Data Labelling and Annotation
- Data Splitting Problems
- Statistical Analysis



We test.
You impress.

Integration Test

Familiar but ...

Probabilistic!

Chained Probabilities multiply

35% cat, 55% dog, 15% blueberry muffin



Integration Test

Familiar but ...

Probabilistic!

Chained Probabilities multiply

35% cat, 55% dog, 15% blueberry muffin

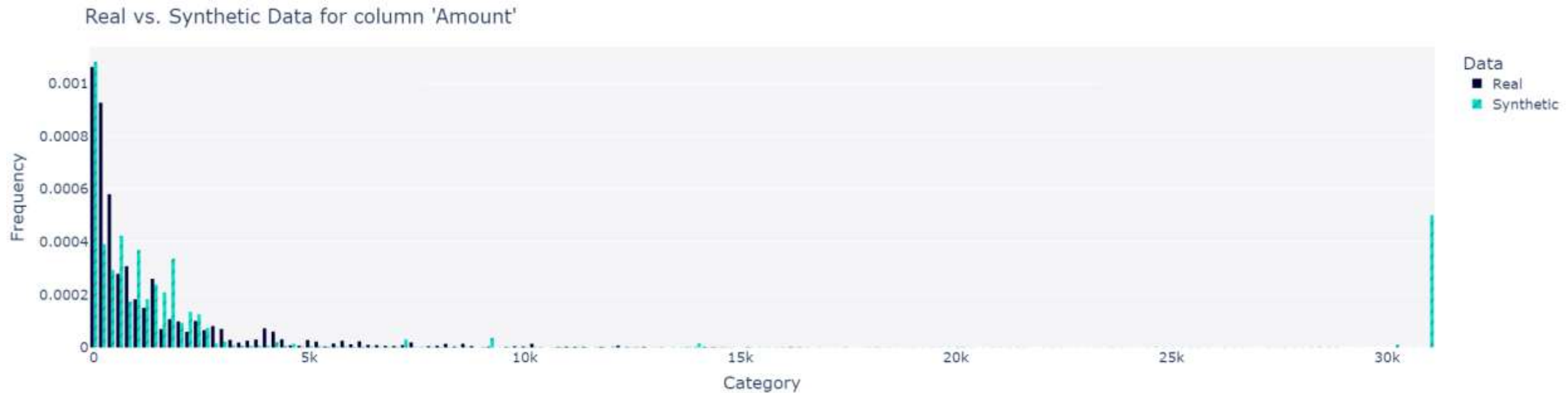


Model Testing

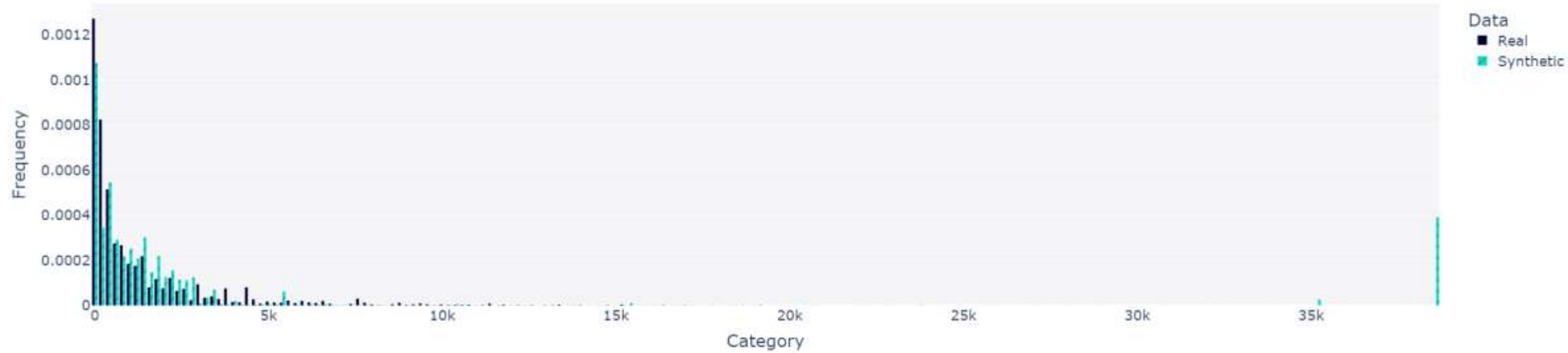
- Statistical in nature
- Large data volumes
- Complex to interpret
- Data Scientists are our friends

Still benefits from a Tester's curiosity, critical thinking, system thinking, and Question Asking!

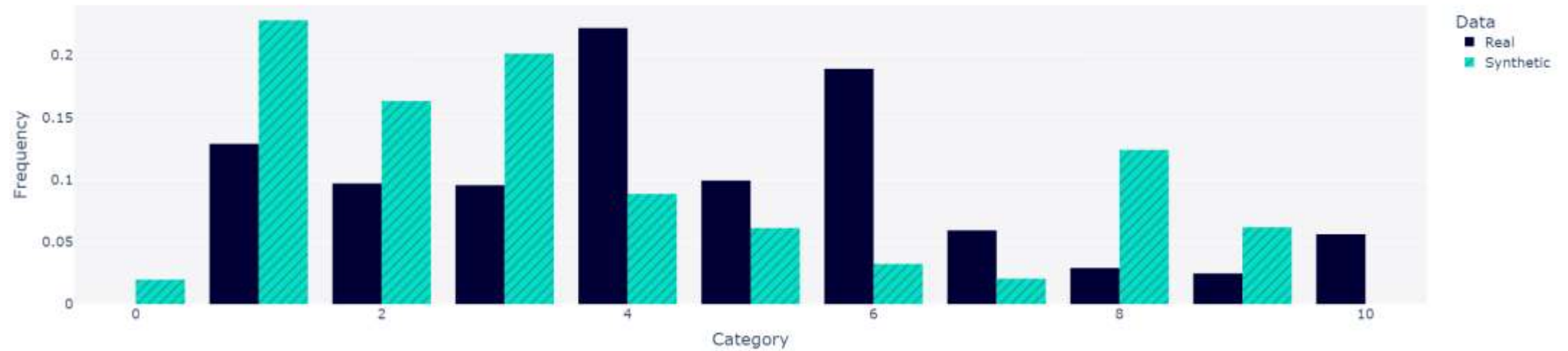
Evaluating the Synthetic Data



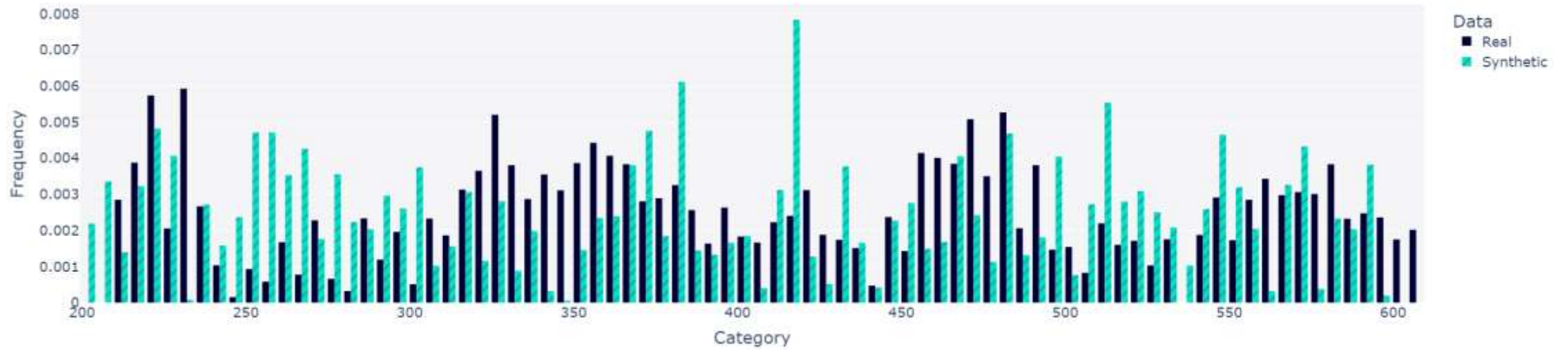
Real vs. Synthetic Data for column 'Balance'



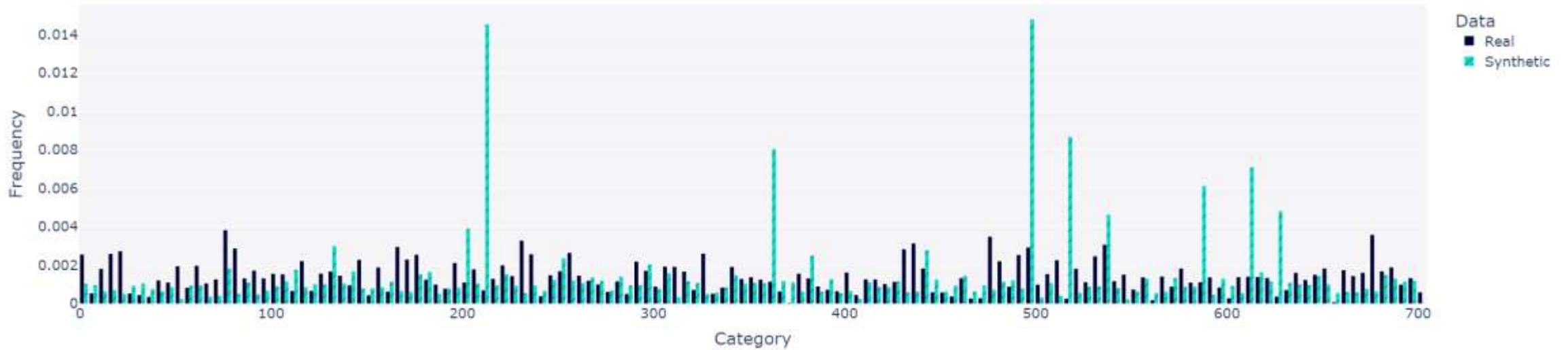
Real vs. Synthetic Data for column 'CountryKey'



Real vs. Synthetic Data for column 'RecipientAccountID'



Real vs. Synthetic Data for column 'SenderAccountID'



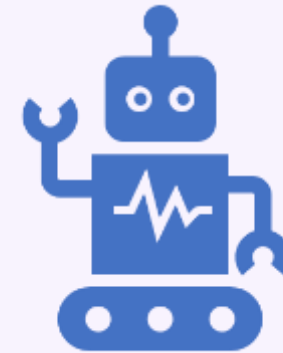
End-to-End

- Very Familiar
- Beware of complexity!
- Involve Experts to help decide what is a “Right” answer

UAT



Pretty much the same



But watch out for Automation
Bias!

Monitor in Live

- Performance
- Drift
 - Context, Concept or Model Drift
 - Data Drift



All the other
usual stuff

- A11y
- Usability
- Security – Check out Adversarial Testing
- Performance – but ...
 - e.g. *Accuracy, Precision, Recall, F1-Score, Confusion Matrix, ROI (or value add), Programmability, Energy/Power, Throughput/latency, GOPS, frame rate, delay, cost, footprint

*Accuracy measures how well your AI model performs on new or unseen data. Precision indicates the relevance of results to your target audience or problem. Recall shows how comprehensive the results are. F1-score is a measure of the balance between precision and recall.

Test Techniques

Explainability

A/B Testing

Parallel Testing

Statistical Analysis

Exploratory Testing

Use Experts

Pairwise/Orthogonal Testing

More Test Techniques

Metamorphic Testing

Adversarial Testing

Model Backtesting

Dual Coding/Algorithm Ensemble

Coverage Data

Cross Validation

Affordances Modelling

The Holy
Trinity



Quality Assurance



Testing

Quality
Engineering

The 4 Factors of the AssureAi Quality Score

Accuracy

Ensuring AI systems produce accurate, reliable outputs that meet user requirements.

- Data Integrity
- Model Validation
- Continuous Monitoring
- Feedback Loops
- Compliance and Standards

Robustness

Ability of AI systems to maintain performance under varying conditions and against adversarial attacks.

- Adversarial Testing
- Data Diversity
- Fault Tolerance
- Scalability
- Security Measures

Explainability

Making AI decisions understandable to humans, including how and why decisions are made.

- Transparent Algorithms
- Feature Importance
- User-Centric Design
- Audit Trails
- Documentation

Performance

The speed of AI systems in executing tasks.

- Benchmarking
- Optimization
- Scalability Testing
- Latency Reduction
- Resource Management

Quality Assurance of AI - Accuracy

Definition of Quality: Ensuring AI systems produce accurate, reliable outputs that meet user requirements.

Data Integrity

- Rigorous evaluation of the data and governance to ensure the quality of training datasets.

Model Validation

- Metrics and cross-validation to assess the model's ability.

Continuous Monitoring

- Establish monitoring systems to track AI performance and identify degradation over time.

Feedback Loops

- Incorporate user feedback to continuously improve AI system outputs and user satisfaction.

Compliance and Standards

- Adhere to industry standards and regulations to ensure quality and build trust.

Quality Assurance of AI - Robustness

Definition of Robustness: Ability of AI systems to maintain performance under varying conditions and against adversarial attacks.

Adversarial Testing

- Evaluate the AI's resilience to adversarial attacks.

Data Diversity

- Ensure training data encompasses a wide range of scenarios to improve model robustness.

Fault Tolerance

- Mechanisms for AI systems to handle errors gracefully and maintain functionality.

Scalability

- Design AI systems to be scalable, maintaining robustness as workload or data volume increases.

Security Measures

- Apply security best practices to protect AI systems from external and internal threats.

Quality Assurance of AI - Explainability

Definition of Explainability: Making AI decisions understandable to humans, including how and why decisions are made.

Transparent Algorithms

- Evaluate how transparent the model is for their decision-making processes.

Feature Importance

- Highlight which features are most influential in the AI's decision-making process.

User-Centric Design

- Develop explanation interfaces that cater to the target audience's level of expertise.

Audit Trails

- Maintain logs of AI decisions to facilitate review and analysis for continuous improvement.

Documentation

- Document the design choices, data sources, interfaces and known issues

Quality Assurance of AI - Performance

Definition of Performance: The speed of AI systems in executing tasks.

Benchmarking

- Compare AI system performance against industry benchmarks or competitor systems.

Scalability Testing

- Evaluate the AI system's ability to handle increased loads effectively.

Resource Management

- Use of hardware and software resources to ensure high performance.

Optimisation

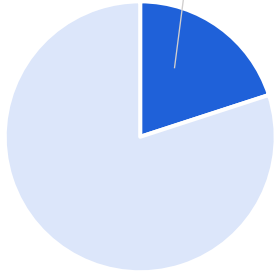
- Techniques to enhance computational efficiency without compromising output quality.

Latency Reduction

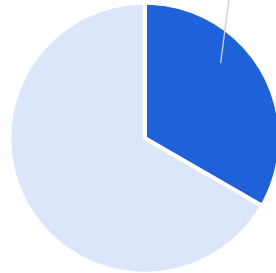
- Strategies to minimize response times, improving user experience.

AssureAi Scores

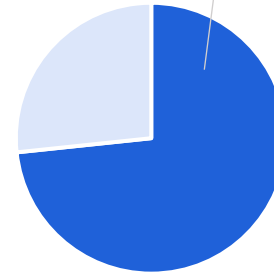
Accuracy



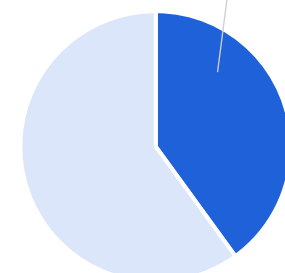
Explainability



Performance



Robustness



| Parameters | Score |
|--------------------------|--------|
| Data Integrity | Silver |
| Model Validation | NPR |
| Continuous Monitoring | NPR |
| Feedback Loops | NPR |
| Compliance and Standards | Gold |

| Parameters | Score |
|------------------------|--------|
| Transparent Algorithms | Bronze |
| Feature Importance | Silver |
| User-Centric Design | NPR |
| Audit Trails | NPR |
| Documentation | Silver |

| Parameters | Score |
|---------------------|--------|
| Benchmarking | Bronze |
| Optimisation | Gold |
| Scalability Testing | Gold |
| Latency Reduction | Silver |
| Resource Management | Silver |

| Parameters | Score |
|---------------------|--------|
| Adversarial Testing | Bronze |
| Data Diversity | NPR |
| Fault Tolerance | Bronze |
| Scalability | Bronze |
| Security Measures | Silver |

NPR = Not production ready

We test.
You impress.



Useful Resources & Tools

- [HuggingFace](#)
- [Fairlearn](#)
- [AI Fairness 360](#)
- [What-If-Tool](#)
- Google Vertex AI
- Amazon SageMaker
- [MLflow](#)
- [Neptune.ai](#)
- [MonkeyLearn](#)
- AWS Comprehend
- [Scikit Learn](#)
- [SciPy](#)
- [Tensorflow
Extended \(TFX\)](#)
- [PyTorch](#)

Contact Us



Bryan Jones

Bryan.Jones@2itesting.com

LinkedIn:

/bryan-jones-mbcs-96953/

www.2itesting.com

Q&A

#EuroSTARConf