Testing AI: A New Test Specialism

Presented by: Dr Stuart Reid

14/09/2022
AI – Importance

Worldwide AI market to exceed €500 billion by 2024

Growing trend: In 2021 56% report some AI adoption (50% in 2020)

92% of businesses are increasing investments in AI and data

26% of companies have AI systems in widespread production

92% of large companies are achieving returns on their AI investments

AI and machine learning are the top IT investment priority in Europe

▲ Research firm IDC
* 2022 survey of senior data and technology executives by NewVantage Partners
▽ Global survey: The state of AI in 2021 | McKinsey
⊕ The Economist Intelligence Unit, 2021
60% believe AI will profoundly change their daily lives in the next 3–5 years
   * (Germany – 44%, UK – 46%, India – 74%, Korea – 76%, China – 80%) *

40% will not share their information with an AI system ⊕

33% do not trust recommendations or decisions from AI systems ⊕
   * but only 41% are aware social media uses AI

50% trust companies using AI as much as they trust other companies
   * (Germany – 42%, UK – 35%, India – 68%, Korea – 46%, China – 76%) *

28% are willing to trust AI systems in general ⊕
   * healthcare AI is trusted more than HR AI
   * 57% would be more willing to use AI if assurance mechanisms are in place

* Ipsos for the World Economic Forum
⊕ KPMG/Univ. of Queensland
AI and Jobs

WEF predict 85 million jobs lost to machines, but 97 million new roles by 2025 *

but only 22% believe that jobs will not be lost ⊖

> 70% of employees are happy with AI for task automation, but far less for HR support, such as evaluating employees ⊖

Data scientist was the second best job in the US in 2021 ▲

AI specialist was the fastest-growing job category in 2020 ▼
The Path to the AI Test Specialism

AI/ML Knowledge

AI/ML-specific potential defects

AI/ML Development

failures → lack of trust

AI/ML Development & Testing

awareness, failures → lack of trust

AI/ML Specialist Testing

Testing Knowledge

AI-specific test techniques

generic test techniques
ML – Step One – Set-Up

ML – Step One

Select Data

Prepare Data

ML Algorithm

Acquire Algorithm

Select Algorithm

Understand Problem

Development Data
ML – Step Two – Model Creation

ML – Step Two

- Development Data
- ML Algorithm
- ML Model
ML – Step Three – Operation

Prepare Data

PRODUCTION DATA

ML MODEL

PREDICTIONS

ML – Step Three
ML – All Three Steps

Select Data → Prepare Data → ML Algorithm → ML Model → Predictions

- Select Data
- Prepare Data
- Development Data
- ML Algorithm
- Acquire Algorithm
- Predictions
- ML Model
- Production Data
- Prepare Data
ML – The Three Specialist Test Areas

Select Data

Prepare Data

Prepare Data

Input Data Testing

DEVELOPMENT DATA

ML ALGORITHM

Acquire Algorithm

Select Algorithm

ML MODEL

PREDICTIONS

PRODUCTION DATA

Model Testing

Framework Testing
Input Data Testing

1. Select Data
2. Prepare Data

Development Data

Input Data Testing

Prepare Data

Production Data

ML Algorithm

ML Model

Predictions

Acquire Algorithm

Select Algorithm
Scope of Input Data Testing

Design, Implement and Test Pipeline

Development Data Pipeline

Raw Data

Development Data Pipeline

Production Data Pipeline

Raw Data

Testing of Trustworthy Systems

#UCAAT
Example Development Data Pipeline

- **RAW DATA**
  - data acquisition
  - data pre-processing
  - feature engineering

- **DEVELOPMENT DATA PIPELINE**

- **acquire pre-gathered datasets**
### Input Data Defect Types

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples/Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unrepresentative training data</td>
<td>- focused on a subset of use cases</td>
</tr>
<tr>
<td></td>
<td>- datasets that do not provide coverage of all regions in the data space</td>
</tr>
<tr>
<td>Biased training data</td>
<td>- missing data</td>
</tr>
<tr>
<td>Data pipeline</td>
<td>- wrong data types</td>
</tr>
<tr>
<td></td>
<td>- out of range data</td>
</tr>
<tr>
<td></td>
<td>- outliers in data</td>
</tr>
<tr>
<td></td>
<td>- incorrectly labelled data</td>
</tr>
<tr>
<td>Data governance rules broken</td>
<td>- sub-optimal feature selection</td>
</tr>
<tr>
<td>Data acquisition</td>
<td>- internally inconsistent</td>
</tr>
<tr>
<td></td>
<td>- skewed through data augmentation</td>
</tr>
<tr>
<td></td>
<td>- imbalanced by insufficient coverage of all target classes</td>
</tr>
<tr>
<td>Data acquisition</td>
<td>- data from untrustworthy sources</td>
</tr>
<tr>
<td></td>
<td>- insecure data input channels</td>
</tr>
</tbody>
</table>
Input Data Testing Types

Data Governance Testing
Data Pipeline Testing
Data Provenance Testing
Data Sufficiency Testing
Data Representativeness Testing
Data Outlier Testing
Dataset Constraint Testing
Label Correctness Testing
Feature Testing
  - Feature Contribution Testing
  - Feature Efficiency Testing
  - Feature-Value Pair Testing
Unfair Data Bias Testing
Input Data Testing Types

Data Governance Testing
- Data Pipeline Testing
- Data Provenance Testing
- Data Sufficiency Testing
- Data Representativeness Testing

Data Outlier Testing
- Dataset Constraint Testing
- Label Correctness Testing
- Feature Testing
  - Feature Contribution Testing
  - Feature Efficiency Testing
  - Feature-Value Pair Testing
- Unfair Data Bias Testing

40% check for under-representation of protected characteristics
37% check for skewed/biased raw data
30% check for skewed/biased data later in the ML workflow
Framework Testing

Select Data

Prepare Data

Prepare Data

DEVELOPMENT DATA

ML ALGORITHM

Select Algorithm

Acquire Algorithm

ML MODEL

PREDICTIONS

PRODUCTION DATA

Framework Testing
Scope of Framework Testing

**Select and Implement Algorithm**

- **Draft Model**
- **Tuned Model**
- **Deployed Model**

**Training Data** → **Training** → **Tuning** → **Validation Data** → **Evaluation** → **Testing**
AI Development Frameworks

IBM Watson Studio
- a suite of tools that support the development of AI solutions

Keras
- a high-level open-source API, written in the Python language, capable of running on top of TensorFlow and CNTK

Apache MxNet
- a deep learning open-source framework used by Amazon for Amazon Web Services (AWS)

CNTK
- the Microsoft Cognitive Toolkit (CNTK) is an open-source deep-learning toolkit

TensorFlow
- an open-source ML framework based on data flow graphs for scalable machine learning, provided by Google

PyTorch
- an open-source ML library operated by Facebook, for apps using image processing and natural language processing (NLP). Supports both Python and C++ interfaces

scikit-learn
- an open-source software machine learning library for the Python programming language
Development Framework Defect Types

Development framework
- sub-optimal selection
- design defect
- implementation defect
- user interface defect
- development library defect
  - e.g. defect in CNTK, PyTorch
- API defect
  - e.g. API to a library or interface between Keras and TensorFlow
- deployment defect

ML algorithm
- sub-optimal selection
- design defect
- implementation defect
- lack of explainability
- documentation defect

Training, evaluation and tuning
- poor allocation of data to training, validation and testing datasets
- poor selection of evaluation approach (e.g. n-fold cross-validation)
- sub-optimal hyperparameter selection
Development Framework Testing Types

Framework Configuration Testing
Model Explainability Testing
ML Algorithm Testing
- Code Review
- Static Analysis
- Dynamic Unit Testing
- API Testing
- Library Implementation Testing
- Model Structure Testing
- Algorithm Bias Testing

Deployment Optimization Testing
Model Deployment Testing
Training Performance Testing
Training Data Recoverability Testing
Model Reproducibility Testing
Model Roll-Back Testing
Framework Security Testing
Framework Suitability Review
Development Framework Testing Types

Framework Configuration Testing
Model Explainability Testing
ML Algorithm Testing
  - Code Review
  - Static Analysis
  - Dynamic Unit Testing
  - API Testing
  - Library Implementation Testing
  - Model Structure Testing
  - Algorithm Bias Testing

Deployment Optimization Testing
Model Deployment Testing
Training Performance Testing
Training Data Recoverability Testing
Model Reproducibility Testing
Model Roll-Back Testing
Framework Security Testing
Framework Suitability Review

Testing of Trustworthy Systems
Model Testing

- Select Data
- Prepare Data
- Development Data
- ML Algorithm
- Acquire Algorithm
- Select Algorithm
- Prepare Data
- Production Data
- ML Model
- Predictions

Testing of Trustworthy Systems
Scope of Model Testing

**ML Model**

- **Tuned Model**
  - Test Data
  - Test Results

- **Deployed Model**
  - Production Data
  - Predictions

---

Testing of Trustworthy Systems

#UCAAT
Model Defect Types

Model form
- unsuitable model selected
- inappropriate model structure

Functional
- wrong function learnt by the model
  (design defect in the model)
  (implementation defect in the model)
- failure to achieve required performance measures (e.g. lack of accuracy)
- API defect
- adversarial examples

Non-functional
- performance efficiency defect
- ethical requirement missed
- biased/unfair ML model
- users not satisfied with model
- unacceptable concept drift

Pre-trained model
- documentation defect
- API defect
Model Testing Types

Functional Testing
- A/B Testing
- Adversarial Testing
- API Testing
- Back-to-Back Testing
- Boundary Value Analysis
- Combinatorial Testing
- Exploratory Testing
- Fuzz Testing
- Metamorphic Testing
- Model Performance Testing
  - Alternative Model Testing
  - Performance Metric Testing
- Model Validation Testing
- Operational Testing
  - Drift Testing
  - Regression Testing

Functional Testing (continued)
- Overfitting Testing
- Reward Hacking Testing
- Scenario Testing
- Side-Effects Testing
- Smoke Testing
- White-Box Testing of Neural Networks

Non-Functional Testing
- Ethical System Testing
- Model Bias Testing
- Model Documentation Review
- Model Suitability Review
- Performance Efficiency Testing
Model Testing Types

**Functional Testing**
- A/B Testing
- Adversarial Testing
- API Testing
- Back-to-Back Testing
- Boundary Value Analysis
- Combinatorial Testing
- Exploratory Testing
- Fuzz Testing
- Metamorphic Testing
- Model Performance Testing
  - Alternative Model Testing
  - Performance Metric Testing
- Model Validation Testing
- Operational Testing
  - Drift Testing
  - Regression Testing

**Functional Testing (continued)**
- Overfitting Testing
- Reward Hacking Testing
- Scenario Testing
- Side-Effects Testing
- Smoke Testing
- White-Box Testing of Neural Networks

**Non-Functional Testing**
- Ethical System Testing
- Model Bias Testing
- Model Documentation Review
- Model Suitability Review
- Performance Efficiency Testing

50% test ML model performance internally before deployment
Supporting the AI Test Specialism

AI/ML Knowledge → AI/ML Development
   → AI/ML Development & Testing
      → AI/ML Specific Testing
         → Testing Knowledge

Conferences & Seminars → Standardization → AI Product Certification → AI/ML Testing Training Courses

failures → lack of trust
Any further questions?