 Responsible AI Toolbox

Besmira Nushi, Rahee Ghosh Peshawaria, Wenxin Wei
Minsoo Thigpen, Mehrnoosh Sameki
Microsoft

Contact: rai-toolbox@microsoft.com
Agenda

Introduction to the Responsible AI Toolbox

 Responsible AI Dashboard - Demo

 Responsible AI Mitigations and Tracker - Demo

 Vision and Language Tasks – Demo & QA

 User Insights, Challenges, and Opportunities

QA
Thanks to our v-team!
Microsoft’s Responsible AI Principles

Common need: Evaluation of system performance across demographic groups, key use cases, and operational factors.
The path to deploying reliable machine learning systems is still unpaved.

Software Engineering for ML: A Case Study
ICSE 2019
Key Challenge: Tool Fragmentation

Desiderata for Tool Integration

Learnability
Discoverability
Sharing Insights & Data
Current Tools: Open-source Building Blocks

- InterpretML – interpret.ml
- Error Analysis – erroranalysis.ai
- Fairlearn – fairlearn.github.io
- DiCE – github.com/interpretml/dice
- EconML – aka.ms/econml
- DoWhy – github.com/microsoft/dowhy
- BackwardCompatibilityML – github.com/microsoft/BackwardCompatibilityML
Introducing: Responsible AI Toolbox
Responsible AI Toolbox

An open-source framework for **accelerating** and **operationalizing Responsible AI** via a set of **interoperable** tools, libraries, and customizable dashboards.

- **Responsible AI Dashboard**
  - responsibleaitoolbox.ai

- **Interpretability Dashboard**
  - interpret.ml

- **Fairness Dashboard**
  - fairlearn.org

- **Error Analysis Dashboard**
  - erroranalysis.ai

- **Responsible AI Mitigations and Tracker**
  - responsibleaitoolbox.ai
Current Model Debugging & Improvement Approaches

Measure Error

Add data
Increase architecture size
Find better parameters

Compare error
Evaluating machine learning models
aka the problem with aggregated metrics

AI-powered scans can identify people at risk of a fatal heart attack almost a DECADE in advance 'by looking at the entire iceberg and not just the tip'

- The AI predicted heart risk with 90% accuracy, according to data
- Current medical scans are only able to see 'the tip of the iceberg'
- It could benefit around 350,000 in Britain, cardiologists believe
- Government funding will fast track the tech into the NHS in two years
Why isn’t this sufficient?

Benchmark

ML Model

89% Accurate

Different regions fail for different reasons
Emotion Recognition

[Howard et al., ARSO 2017]
Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems

TABLE I. DEEP LEARNING RECOGNITION RATES ACROSS THE DIFFERENT STIMULI SETS (IN %): (Fe)AR, (An)Gry, (Ha)PPY, (Sa)D, (Ne)UTrAL, (Su)RPRISEd, (Di)GUST, (Co)NTEMP

<table>
<thead>
<tr>
<th></th>
<th>Fe</th>
<th>An</th>
<th>Di</th>
<th>Ha</th>
<th>Ne</th>
<th>Sa</th>
<th>Su</th>
<th>Co</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIMH-ChEFS</td>
<td>13</td>
<td>43</td>
<td>100</td>
<td>100</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dartmouth</td>
<td>25</td>
<td>35</td>
<td>55</td>
<td>100</td>
<td>99</td>
<td>64</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>Radboud</td>
<td>33</td>
<td>54</td>
<td>100</td>
<td>100</td>
<td>95</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEPS</td>
<td>5</td>
<td>50</td>
<td>10</td>
<td>95</td>
<td>92</td>
<td>52</td>
<td>81</td>
<td></td>
</tr>
</tbody>
</table>
GenderShades Study

Follow up case study

[Buolamwini and Gebru, FAccT 2018]
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

[Lu et al., ICLR DebugML 2018]
Error terrain analysis for machine learning: Tool and visualizations
Performance discrepancies in the real world

Safety

Fairness

Trust
Concepts of disaggregated evaluation

Cohort (aka data slices, regions, subgroups, clusters):
Subsets of data created by adding filters to the overall test or train datasets.
Examples:
“age > 40 and residency= ‘Florida’”
“gender=female and ‘diabetes’ in pre_existing_conditions”

Performance discrepancy (ratio or difference):
- Discrepancy between all data vs. cohort of interest
- Discrepancy between two cohorts of interest
  Example: WA residents vs NY residents
- The best and worst performance across combinations of features.
  Example: the best and worst performance for combinations of gender and age
- Discrepancy between cohorts with the best and worst performance
Cohort design considerations

1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery
Cohort design considerations

1. Ground truth filters vs. Automated metadata filters

2. Consider the application-based cost of error

3. Cohort size in the train/test data may not reflect real-world usage

4. Automated vs. manual high-error cohort discovery

Credit risk assignment example

- 20% false positives for small loans (e.g. < $5000)
- 5% false positives for larger loans (e.g. > $20,000)
Cohort design considerations

1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery

Race representation in UCI Income Dataset
Cohort design considerations

1. Ground truth filters vs. Automated metadata filters
2. Consider the application-based cost of error
3. Cohort size in the train/test data may not reflect real-world usage
4. Automated vs. manual high-error cohort discovery

Automated discovery
Useful for quick discovery of cohorts with significantly higher error rates

Visualization based on Responsible AI Dashboard: https://github.com/microsoft/responsible-ai-toolbox
Cohort design considerations

1. Ground truth filters vs. Automated metadata filters

2. Consider the application-based cost of error

3. Cohort size in the train/test data may not reflect real-world usage

4. Automated vs. manual high-error cohort discovery

Manual discovery
Useful for exploring errors on known important cohort definitions.

Visualization based on Responsible AI Dashboard:
https://github.com/microsoft/responsible-ai-toolbox
Disaggregation: from evaluation to debugging

Data and model debugging
- Imbalance
- Noise
- Missing values
- Distribution shifts
- Spurious correlations
- Wrong labels

Disaggregated evaluation
Discrepancy metrics

Disaggregated training data metrics
e.g. class imbalance etc.

Different cohorts may have very different class imbalances which may or may not align with the overall class balance ratios in the training data.
Disaggregated model comparison

Baseline Model
80% accurate

Candidate Model
85% accurate

Model Updates may lead to new mistakes and lost trust.
Incompatibility Sources

**Optimization Stochasiticity**
- Stochastic batches in gradient descent
- Model initialization
- Random data augmentation
- Distributed training

**Label Noise**
- Semi-supervised learning with noisy data
- Human labeling error

**Distributional Shifts**
- Training data is not a representation of the real world
- Bias in data collection
- The concept definition changes
- Domain transfer

**Model Class**
- Fundamental architectural changes
Compatibility is not built-in

### Updates in Practice

**[Bansal et al., AAAI 2019]**
Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff

**[Srivastava et al., KDD 2020]**
An empirical analysis of backward compatibility in machine learning systems

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Dataset</th>
<th>Perf. v1</th>
<th>Perf. v2</th>
<th>Compatibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>Recidivism</td>
<td>0.68</td>
<td>0.72</td>
<td>72%</td>
</tr>
<tr>
<td></td>
<td>Credit Risk</td>
<td>0.72</td>
<td>0.77</td>
<td>66%</td>
</tr>
<tr>
<td></td>
<td>Mortality</td>
<td>0.68</td>
<td>0.77</td>
<td>40%</td>
</tr>
<tr>
<td>Multi-layered Perceptron</td>
<td>Recidivism</td>
<td>0.59</td>
<td>0.73</td>
<td>53%</td>
</tr>
<tr>
<td></td>
<td>Credit Risk</td>
<td>0.70</td>
<td>0.80</td>
<td>63%</td>
</tr>
<tr>
<td></td>
<td>Mortality</td>
<td>0.71</td>
<td>0.84</td>
<td>76%</td>
</tr>
</tbody>
</table>

High-stake decision-making

Backward compatibility scores available at:
[https://github.com/microsoft/BackwardCompatibilityML](https://github.com/microsoft/BackwardCompatibilityML)

Low compatibility
Percentage of predictions that remain correct.
Targeted Debugging for Machine Learning

Identify

Diagnose

Mitigate

Track, Compare, Validate

Responsible AI Tracker
Debugging Machine Learning Models

**Identify**
- Fairlearn: Fairness Assessment
- Error-Analysis: Error Analysis

**Diagnose**
- InterpretML: Interpret and Debug Models
- Counterfactual: Diverse Counterfactual Explanations for Debugging
- Exploratory-Data-Analysis: Understand Dataset Characteristics

**Mitigate**
- Fairlearn: Unfairness Mitigation Algorithms
- Responsible AI Mitigations: Enhance your dataset and retrain model

**Compare & Validate**
- Responsible AI Tracker
- Model Comparison

- Backward Compatibility
The future of data science productivity and tools

code  data  model  visualizations
Responsible AI Dashboard
ML Debugging and Causal Decision-Making
Responsible AI Dashboard

An open-source framework for accelerating and operationalizing Responsible AI via a set of interoperable tools, libraries, and customizable dashboards.

**Identify**
- Error Analysis
  - Identify cohorts with high error rate versus benchmark and visualize how the error rate distributes
- Fairness Assessment
  - Aggregate a variety of fairness assessment metrics, showing model prediction distributions

**Diagnose**
- Model Interpretability
  - Interpret and debug models
- Counterfactual Analysis and What If
  - Generate diverse counterfactual explanations for debugging. Perform feature perturbations
- Exploratory Data Analysis
  - Understand dataset characteristics

**Mitigate**
- Unfairness Mitigation
  - Mitigate fairness issues (via Fairlearn.org)
- Data Enhancements
  - Enhance your dataset and retrain model

**Compare**
- Model Comparison
- Backward Compatibility

**Make Decisions**
- Causal Inference
  - Understand the causal impact of your features on real-world outcomes
- Counterfactual Analysis
  - Generate diverse counterfactual explanations for providing actionable insights to users
Identify

Error Analysis
Fairness Assessment
Error Analysis

Rigorous performance evaluation and testing is often needed to deploy models in production.

Analyze and debug model errors

Benchmark

ML Model

89% Accurate

Different regions fail for different reasons

Analyzing and debugging model errors
Fairness in AI

There are many ways that an AI system can behave unfairly.

A voice recognition system might fail to work as well for women as it does for men.

A model for screening job application might be much better at picking good candidates among white men than among other groups.

Avoiding negative outcomes of AI systems for different groups of people

Learn more
Diagnose

Interpretability
Counterfactuals
Data Exploration
Interpretability

**Understand overall model predictions**
What are the top K important factors impacting your overall model predictions?

**Understand individual model predictions**
What are the top K important factors impacting your model predictions for a single sample?
Counterfactuals

**Debug model predictions**
Enable data scientists and model evaluators to debug models by understanding the closest datapoints with different prediction outcomes.

**Make responsible model-driven decisions**
Answer end-users’ questions such as “what can I do to get a different outcome from the AI model?”
Mitigate

Model Fairness mitigations
Data mitigations
Take Action

Causal Inference
Counterfactual Analysis
Causal Inference

**Understand overall causal effects**
Answer real-world “what if” questions about how an outcome would have changed under different policy choices.

**Explore individual causal effects**
Inform personalized interventions, such as a targeted promotion to customers or an individualized treatment plan. Learn about how an individual with a particular set of features respond to a change in a causal feature, or treatment.

**Extract a treatment policy**
Build policies for future interventions. Identify what parts of your sample experience the largest responses to changes in causal features, or treatments, and construct rules to define which future populations should be targeted for particular interventions.
A comprehensive single-pane-of-glass experience with a variety of model and data exploration capabilities such as Error Analysis, Model Explanations, Fairness metrics, and Data Exploration.
YAML-powered workflow: Introducing CLI experience to generate an RAI dashboard as part of an automated pipeline workflow using YAML.

Customizable: Specify which RAI components you want to generate to fit your scenario.

No code wizard: Introducing end-to-end on-demand generation of the dashboard from AML studio workspace UI.

Reporting: Export a PDF report of your RAI insights to share with business stakeholders.
Generate key summaries of Responsible AI insights by exporting to PDF. Share with technical and non-technical stakeholders to aid in compliance review.
Responsible AI Dashboard for Vision and Language
Responsible AI dashboard support for image and text data

Newly available:

**Common issues for image models**
- Misalignment of bounding boxes
- Object overlap
- Spurious correlations
- Labeling errors

**Common issue for text models**
- Problems with grounding
- Linguistic shortcuts
### What is new?

<table>
<thead>
<tr>
<th>Icon</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td>Rich visualizations for vision and text</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td>Meta-data support and ingestion</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image" /></td>
<td>Interpretability for vision and text</td>
</tr>
<tr>
<td><img src="image4.png" alt="Image" /></td>
<td>Consistent design with customizable debugging workflows</td>
</tr>
</tbody>
</table>


Meta-data for cohort design in Vision

- **Ground truth**
  - Demographics
  - Synthetics
  - Bounding Box Info

- **System data**
  - Camera Settings
  - Time of day
  - Location

- **Inferred attributes**
  - Brightness, Noise
  - Objects
  - Auto Captions
Meta-data for cohort design in Language

ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT, formerly known as ACM FAT*) is a peer-reviewed academic conference series about ethics and computing systems.[1] Sponsored by the Association for Computing Machinery, this conference focuses on issues such as algorithmic transparency, fairness in machine learning, bias, and ethics from a multi-disciplinary perspective. The conference community includes computer scientists, statisticians, social scientists, scholars of law, and others.[2]

The conference is sponsored by Big Tech companies such as Facebook, Twitter, and Google, and large foundations such as the Rockefeller Foundation, Ford Foundation, MacArthur Foundation, and Luminate.[3] Sponsors contribute to a general fund (no "earmarked" contributions are allowed) and have no say in the selection, substance, or structure of the conference.[4]

Ground truth

System data

Inferred attributes

Text Length
Gendered Words
Parse Tree
Complexity

Steps in Interaction
User Scenario
Telemetry

Sentiment
Toxicity
Topics
Entities

responsible-ai-toolbox/tree/main/nlp_feature_extractors
Responsible AI Mitigations and Tracker
Responsible AI Mitigations

```
pip install raimitigations
```

**Identify**
- Model has higher error for cohort X (e.g. old houses, children)

**Diagnose**
- Cohort X has a different class imbalance than the rest of the data
- Features that are informative for the whole data, are not useful for cohort X
- Numerical features are scaled for the whole data and not for cohort X
- Missing values for cohort X

**Mitigate**
- Balance the data for cohort X
- Select features
- Create new ones
- Feature scaling
- Value Imputation
An overview

`pip install raimitigations`

A rich set of mitigations focusing on **data quality** as it relates to the quality of ML models.

A simple interface for mitigation steps that follows the `.fit()` and `.transform()` convention.

Function calls adapted for responsible AI by extending existing calls either with **target features or cohorts**.

Possible to create **different models for different cohorts**, or **post process** predictions for improving predictions in a cohort.
Library Components

https://github.com/microsoft/responsible-ai-toolbox-mitigations
Library Workflow

Data → Mitigation Pipeline → Mitigated Dataset → Model

- **Scale**
- **Impute**
- **Rebalance**

Cohort Manager

- Cohort 0 → Scale → Impute → Mitigated Dataset
- Cohort 1 → Rebalance → Scale → Impute → Mitigated Dataset

Performance Metric
Targeted Mitigations

Data mitigation strategy

<table>
<thead>
<tr>
<th>Common</th>
<th>Separate Same type</th>
<th>Separate Different types</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Blanket mitigation</strong></td>
<td>Applies the same mitigation type to all cohorts and uses all data as context.</td>
<td><strong>Targeted mitigation</strong></td>
</tr>
<tr>
<td>Trains a single model for all cohorts.</td>
<td>Trains a single model for all cohorts.</td>
<td>Trains a single model for all cohorts.</td>
</tr>
<tr>
<td><strong>Targeted mitigation</strong></td>
<td>Applies the same mitigation type to all cohorts and uses all data as context.</td>
<td><strong>Targeted mitigation</strong></td>
</tr>
<tr>
<td>Trains different models for different cohorts</td>
<td>Trains different models for different cohorts.</td>
<td>Trains different models for different cohorts.</td>
</tr>
</tbody>
</table>

Identify

Diagnose

Mitigate
Responsible AI Tracker
https://github.com/microsoft/responsible-ai-toolbox-tracker

Managing and linking model improvement artefacts for cleaner data-science practices: code, models, visualizations, data.

Disaggregated model evaluation and comparison, for tracking both performance improvements and declines.

Initial integration with the Responsible AI Mitigations library. More to be done for e2e model improvement.

Initial integration with mlflow.
<table>
<thead>
<tr>
<th>Notebook</th>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 estimators.ipynb</td>
<td>baseline</td>
<td>0.789</td>
</tr>
<tr>
<td>balance all data.ipynb</td>
<td></td>
<td>0.827</td>
</tr>
<tr>
<td>target balance per cohort.ipynb</td>
<td></td>
<td>0.849</td>
</tr>
<tr>
<td>balance per cohort.ipynb</td>
<td></td>
<td>0.793</td>
</tr>
</tbody>
</table>

**Visualization reports**

### Metrics
- Accuracy
- Precision
- Recall
- F1 Score
- Log Loss
- ROC AUC

<table>
<thead>
<tr>
<th>Notebook</th>
<th>Cohort</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Log Loss</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 estimators.ipynb</td>
<td>adult-test-sampl.csv</td>
<td>0.789</td>
<td>1</td>
<td>0.134</td>
<td>0.237</td>
<td>0.433</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>0.611</td>
<td>1</td>
<td>0.149</td>
<td>0.259</td>
<td>0.635</td>
<td>0.797</td>
</tr>
<tr>
<td></td>
<td>Not married</td>
<td>0.936</td>
<td>1</td>
<td>0.054</td>
<td>0.102</td>
<td>0.266</td>
<td>0.836</td>
</tr>
<tr>
<td>balance all data.ipynb</td>
<td>adult-test-sampl.csv</td>
<td>0.827</td>
<td>0.626</td>
<td>0.272</td>
<td>0.673</td>
<td>0.436</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>0.683</td>
<td>0.616</td>
<td>0.215</td>
<td>0.702</td>
<td>0.443</td>
<td>0.784</td>
</tr>
<tr>
<td></td>
<td>Not married</td>
<td>0.947</td>
<td>0.936</td>
<td>0.237</td>
<td>0.378</td>
<td>0.276</td>
<td>0.837</td>
</tr>
<tr>
<td>target balance.ipynb</td>
<td>adult-test-sampl.csv</td>
<td>0.849</td>
<td>0.613</td>
<td>0.495</td>
<td>0.615</td>
<td>0.022</td>
<td>0.888</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>0.73</td>
<td>0.803</td>
<td>0.542</td>
<td>0.647</td>
<td>0.388</td>
<td>0.792</td>
</tr>
<tr>
<td></td>
<td>Not married</td>
<td>0.947</td>
<td>0.977</td>
<td>0.231</td>
<td>0.374</td>
<td>0.272</td>
<td>0.81</td>
</tr>
<tr>
<td>balance per cohort.ipynb</td>
<td>adult-test-sampl.csv</td>
<td>0.793</td>
<td>0.563</td>
<td>0.674</td>
<td>0.614</td>
<td>0.111</td>
<td>0.846</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>0.728</td>
<td>0.72</td>
<td>0.663</td>
<td>0.69</td>
<td>0.431</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Not married</td>
<td>0.846</td>
<td>0.27</td>
<td>0.737</td>
<td>0.395</td>
<td>0.262</td>
<td>0.876</td>
</tr>
</tbody>
</table>

Learning to mitigate for Responsible AI

- code
- data
- model
- visualizations

Track

Learn and Improve
- automl
- meta-learning
- code generation

- cohort A
- cohort B

Impute
- Scale
- Rebalance
- Scale
User Insights, Challenges, and Opportunities
Responsible AI as an open-source opportunity

Transparency

Research & Education

Integration with OSS frameworks
Adoption Challenges

- Disaggregated evaluation, reliability and ML criticality
- Choosing the right metrics (domain expertise)
- Integration of RAI, ML tools with other tools in the ML Lifecycle
- Wide range of ML expertise and problem domains
- Responsible AI pre- and post-production
## Insights - What works?

- Co-design with users/customers
- Vertical solutions (e.g. Responsible AI for Healthcare)
- Customization and flexibility (metrics, components)
- Transparency, reproducibility, reusability of evaluation pipelines
- Processes, Culture, Education – beyond tools
RESPONSIBLE AI MATURITY MODEL

Mapping your organization’s goals on the path to responsible AI

Level 1: Latent
Level 2: Emerging
Level 3: Developing
Level 4: Realizing
Level 5: Leading

Mihaela Vorvoreanu • Amy Heger • Samir Passi • Shipi Dhanorkar • Zoe Kahn • Ruotong Wang
Aether Central UX Research & Education • Microsoft

https://aka.ms/raimm
V1 • May 17, 2023
Stay tuned

https://github.com/microsoft/responsible-ai-toolbox
https://github.com/microsoft/responsible-ai-toolbox-mitigations
https://github.com/microsoft/responsible-ai-toolbox-tracker

- Extend the Responsible AI Dashboard for Generative AI
- More functionality around model comparison and monitoring
- Scalability investments and distributed mitigations
- Learning to mitigate for Responsible AI
Useful links

**Responsible AI Toolbox**  
[https://github.com/microsoft/responsible-ai-toolbox](https://github.com/microsoft/responsible-ai-toolbox)

**Responsible AI Tracker**  
[https://github.com/microsoft/responsible-ai-toolbox-tracker](https://github.com/microsoft/responsible-ai-toolbox-tracker)

**Responsible AI Mitigations**  
[https://github.com/microsoft/responsible-ai-toolbox-mitigations](https://github.com/microsoft/responsible-ai-toolbox-mitigations)

**Responsible AI: The research collaboration behind new open-source tools offered by Microsoft**  

**Responsible AI Dashboard Deep Dive Blogs**  
- Responsible AI dashboard: A one-stop shop for operationalizing Responsible AI in practice: [Tech Community blog](https://techcommunity.microsoft.com)  
- Responsible AI Dashboard in Azure Machine Learning: [Tech Community blog](https://techcommunity.microsoft.com)  
- Debug Object Detection Models with the Responsible AI Dashboard: [Tech Community blog](https://techcommunity.microsoft.com)

**Responsible AI Mitigations and Tracker: New open-source tools for guiding mitigations in Responsible AI**  
[aka.ms/rai-mitigationstracker-blog](aka.ms/rai-mitigationstracker-blog)
Questions?

rai-toolbox@microsoft.com