Trusted AI: Bringing Trust and Transparency into AI through Open Source

Todd Moore
VP, IBM Open Technology and Developer Advocacy
So what does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

Did anyone tamper with it?

Is it fair?

Is it easy to understand?

Is it accountable?
Robust AI Example: Self Driving Vehicles

Adversarial machine learning

Adversarial machine learning can be used to “trick” machine learning models into providing incorrect predictions, often with devastating consequences e.g. self driving vehicles.

Figure 2. Illustration of an adversary generating a dynamic target segmentation for hiding pedestrians.
Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe’s COMPAS algorithm.

Defendants with low risk scores are released on bail.

It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.

---

IBM also has a long history in the open source ecosystem

and

We are leveraging this to bring Trust and Transparency into AI though Open Source..
Trusted AI Lifecycle through Open Source

Pillars of trust, woven into the lifecycle of an AI application

- Did anyone tamper with it? (ROBUSTNESS)
- Is it fair? (FAIRNESS)
- Is it easy to understand? (EXPLAINABILITY)
- Is it accountable? (LINEAGE)

Adversarial Robustness 360
↳ (ART)
- github.com/IBM/adversarial-robustness-toolbox
- art-demo.mybluemix.net

AI Fairness 360
↳ (AIF360)
- github.com/IBM/AIF360
- aif360.mybluemix.net

AI Explainability 360
↳ (AIX360)
- github.com/IBM/AIX360
- aix360.mybluemix.net

In the works!
AI Fairness 360 (AIF360)

https://github.com/IBM/AIF360

AIF360 toolkit is an open-source library to help detect and remove bias in machine learning models. **AIF360 translates algorithmic research from the lab into practice.** Applicable domains include finance, human capital management, healthcare, and education.

The AI Fairness 360 Python package includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models.

**Toolbox**
Fairness metrics (70+)
Fairness metric explanations
Bias mitigation algorithms (10+)

http://aif360.mybluemix.net/
AIX Explainability 360

AIX360 toolkit is an open-source library to help explain AI and machine learning models and their predictions. This includes three classes of algorithms: local post-hoc, global post-hoc, and directly interpretable explainers for models that use image, text, and structured/tabular data.

The AI Explainability360 Python package includes a comprehensive set of explainers, both at global and local level.

Toolbox
Local post-hoc
Global post-hoc
Directly interpretable

http://aix360.mybluemix.net
We are also making these capabilities around Trusted AI available to businesses through Watson OpenScale. Watson OpenScale tracks and measures trusted AI outcomes across its lifecycle, and adapts and governs AI to changing business situations — for models built and running anywhere.

**Measure and track AI outcomes**
Track performance of production AI and its impact on business goals, with actionable metrics in a single console.

**Govern, detect bias and explain AI**
Maintain regulatory compliance by tracing and explaining AI decisions across workflows, and intelligently detect and correct bias to improve outcomes.

![Insights Dashboard](image)
IBM is joining Linux Foundation AI

Trust and Transparency into AI though Open Source
We would like to partner with community to build Trusted and Transparent AI

To collaborate, look at the corresponding projects here
codait.org
or
https://github.com/topics/trusted-ai

and reach out via github or send an email to singhan@us.ibm.com