LFAI Trusted AI Committee

Animesh Singh

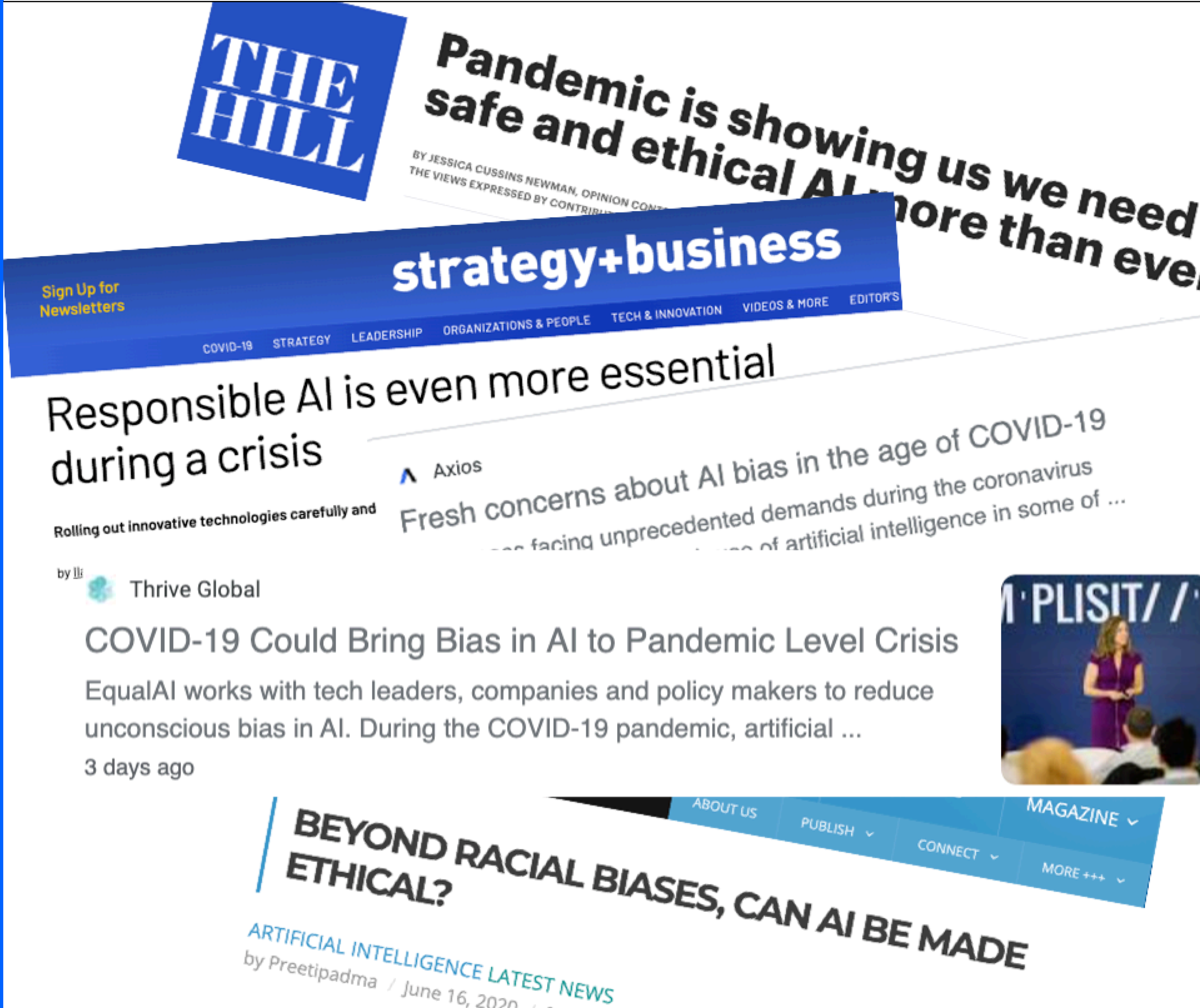
THE LINUX FOUNDATION

LF AI

We are have been working responsibly to bring

Trust and Transparency into AI..

…relevant more so in these unprecedented times….



THE HILL

Pandemic is showing us we need safe and ethical AI more than eve

BY JESSICA CUSSINS NEWMAN, OPINION CONT
THE VIEWS EXPRESSED BY CONTRIBU

Responsible AI is even more essential during a crisis

Rolling out innovative technologies carefully and

by Ili

Axios

Fresh concerns about AI bias in the age of COVID-19

facing unprecedented demands during the coronavirus
of artificial intelligence in some of ...

Thrive Global

COVID-19 Could Bring Bias in AI to Pandemic Level Crisis

EqualAI works with tech leaders, companies and policy makers to reduce unconscious bias in AI. During the COVID-19 pandemic, artificial ...

3 days ago

BEYOND RACIAL BIASES, CAN AI BE MADE ETHICAL?

ARTIFICIAL INTELLIGENCE LATEST NEWS

by Preetipadma / June 16, 2020

"We believe now is the time to begin a national dialogue on whether and how facial recognition technology should be employed by domestic law enforcement agencies"

Arvind Krishna

IBM CEO, June 2020 letter to US Congress

“If we fail to make **ethical** and **inclusive** artificial intelligence we risk losing gains made in civil rights and gender equity under the guise of machine neutrality.”

Joy Buolamwini
Gender Shades
MIT Media Lab

Joy Buolamwini – MIT Media Lab <https://www.media.mit.edu/people/joyab/overview/> (CC BY 4.0).

Trusted AI Lifecycle through Open Source

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

Did anyone tamper with it?

ROBUSTNESS

Adversarial Robustness 360

↳ (ART)

github.com/IBM/adversarial-robustness-toolbox

art-demo.mybluemix.net

Is it fair?

FAIRNESS

AI Fairness 360

↳ (AIF360)

github.com/IBM/AIF360

aif360.mybluemix.net

Is it easy to understand?

EXPLAINABILITY

AI Explainability 360

↳ (AIX360)

- github.com/IBM/AIX360

aix360.mybluemix.net

Is it accountable?

LINEAGE

AI Factsheets

LFAI Trusted AI Committee

https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee

Bring Trust, Transparency and Responsibility into AI

- ✓ Principles Working Group
- ✓ Technical Working Group

Chairs	Region	Company
Animesh Singh	*North America*	*IBM*
Souad Ouali	*Europe*	*Orange*
Jeff Cao	*Asia*	*Tencent*

IBM

AT&T

orange™

ERICSSON

Tech Mahindra

Tencent

HUAWEI

AI for people

ETHICAL ML INSTITUTE

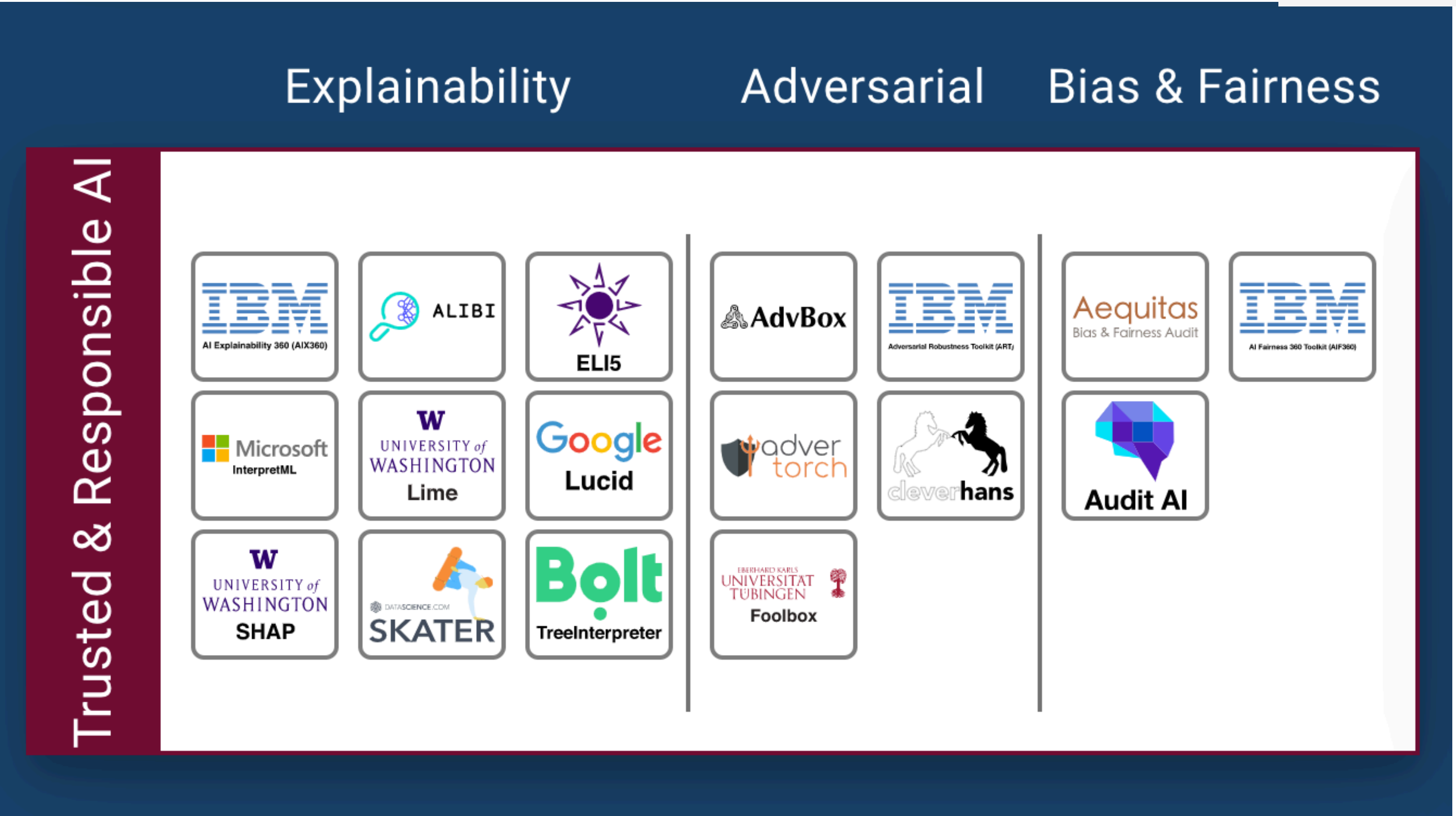
amdocs

MAIEI

LF AI | TRUSTED AI

Trusted AI Updates

LF AI Landscape


Explainability
Adversarial
Bias & Fairness
Trusted & Responsible AI
IBM AI Explainability 360 (AIX360)
ALIBI
ELI5
AdvBox
IBM Adversarial Robustness Toolkit (ART)
Aequitas Bias & Fairness Audit
IBM AI Fairness 360 Toolkit (AIF360)
Microsoft InterpretML
W UNIVERSITY of WASHINGTON Lime
Google Lucid
advertorch
cleverhans
Audit AI
W UNIVERSITY of WASHINGTON SHAP
DATASCIENCE.COM SKATER
Bolt TreeInterpreter
EBERHARD KARLS UNIVERSITÄT TÜBINGEN Foolbox

Trusted AI Updates

Technical Working Group:

Topics presented and discussed

AI Fairness 360: MLOps Section created

Kubeflow Pipelines Integration
https://github.com/IBM/AIF360/tree/master/mlops/kubeflow

Apache Nifi Integration:
https://github.com/IBM/AIF360/tree/master/mlops/nifi

SKLearn API support for AIF360https://github.com/IBM/AIF360/tree/master/aif360/sklearn

The AI Fairness 360 R package
https://github.com/IBM/AIF360/tree/master/aif360/aif360-r

AI Factsheets
https://www.ibm.com/blogs/research/2018/08/factsheets-ai/

KFServing Integration
http://bit.ly/kubeflow-trusted-ai

KPMG: Trusted AI in field
https://lists.lfai.foundation/g/trustedai-committee/files/KPMG%20AI%20in%20Control%20May142020.pdf

Adversarial Robustness 360: MLOps Section created

Kubeflow Integration
https://github.com/IBM/adversarial-robustness-toolbox/tree/master/mlops

Principles Working Group:

Materials submitted to Trusted AI Committee from
--- Orange (document draft 4.1 dated 12 August 2019)
--- AT&T (Working Draft Artificial Intelligence Operating Principles Under Development version dated Nov 7, 2019)
--- TenCent (Jeff Cao - Tencent Research Institute - slides)
--- IBM (https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee#TrustedAICommittee-Assets)
--- Institute of Ethical AI https://github.com/EthicalML/awesome-artificial-intelligence-guidelines https://ethical.institute/

Initial PWG Trusted AI documents produced. Tencent, Orange and IBM have signed off. Seeking an AT&T signoff

Orange Responsible AI Presentation
https://lists.lfai.foundation/g/trustedai-committee/files/2020_Responsable_AI_Orange_LFAI.pdf

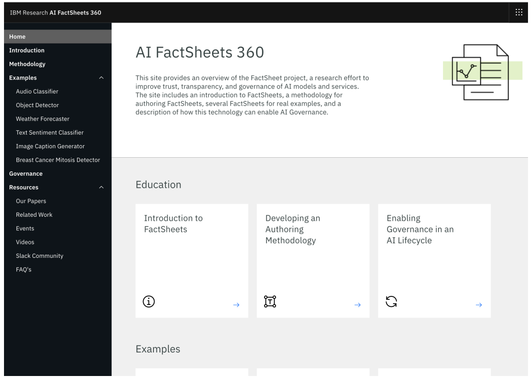
Upcoming Work::

Finalize the LFAI Principle document outlining, among other things-
-- scope - who creates AI - humans or machines/AIs
-- bias in definitions
-- consider how to organize principles - perhaps in a hierarchy
- particular contribution of document: linking principles to implementation; incorporating global principles and thinking, linking to business throug use cases
-- more on correlation to business e.g., how explainability links to business (edited)

Technical Working Group Activities

AI Factsheets 360

https://aifs360.mybluemix.net



IBM Research **AI FactSheets 360**

- **Home**
- **Introduction**
- **Methodology**
- **Examples**
  - Audio Classifier
  - Object Detector
  - Weather Forecaster
  - Text Sentiment Classifier
  - Image Caption Generator
  - Breast Cancer Mitosis Detector
- **Governance**
- **Resources**
  - Our Papers
  - Related Work
  - Events
  - Videos
  - Slack Community
  - FAQ's

AI FactSheets 360

This site provides an overview of the FactSheet project, a research effort to improve trust, transparency, and governance of AI models and services. The site includes an introduction to FactSheets, a methodology for authoring FactSheets, several FactSheets for real examples, and a description of how this technology can enable AI Governance.

Education

Introduction to FactSheets →

Developing an Authoring Methodology →

Enabling Governance in an AI Lifecycle →

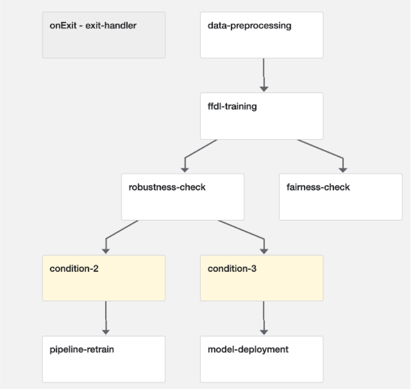
Examples

- AI Governance
- AI Transparency
- Sample factsheets with IBM Model Asset Exchange (MAX) models

AIF 360

- Available now in R (before was only in python)
- Compatibility to Scikit Learn
- Blogs:
- [The AIF360 fairness toolkit is now available for R users]()
- [The AIF360 team adds compatibility with scikit-learn]()
- [IBM continues momentum in AI and trust leadership]()

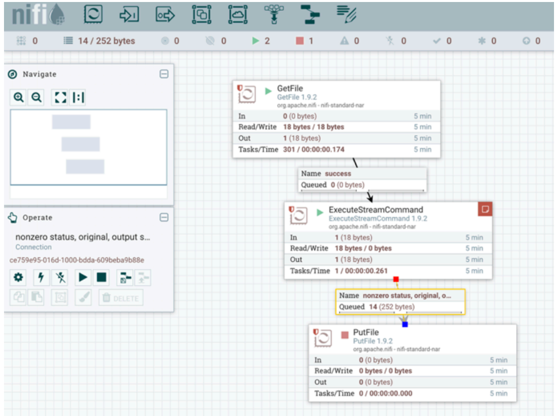
MLOps: Kubeflow Pipelines



- Starter MLOps components for AIF360 and ART
- Use for fairness and adversarial detection with Kubeflow Pipelines

https://github.com/IBM/AIF360/tree/master/mlops

MLOps: Apache Nifi



nonzero status, original, output s...
Connection
ce759e95-016d-1000-bdda-609beba9b88e

GetFile
GetFile 1.9.2
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	18 bytes / 18 bytes	5 min
Out	1 (18 bytes)	5 min
Tasks/Time	301 / 00:00:00.174	5 min

Name success
Queued 0 (0 bytes)

ExecuteStreamCommand
ExecuteStreamCommand 1.9.2
org.apache.nifi - nifi-standard-nar

In	1 (18 bytes)	5 min
Read/Write	18 bytes / 0 bytes	5 min
Out	1 (18 bytes)	5 min
Tasks/Time	1 / 00:00:00.261	5 min

Name nonzero status, original, o...
Queued 14 (252 bytes)

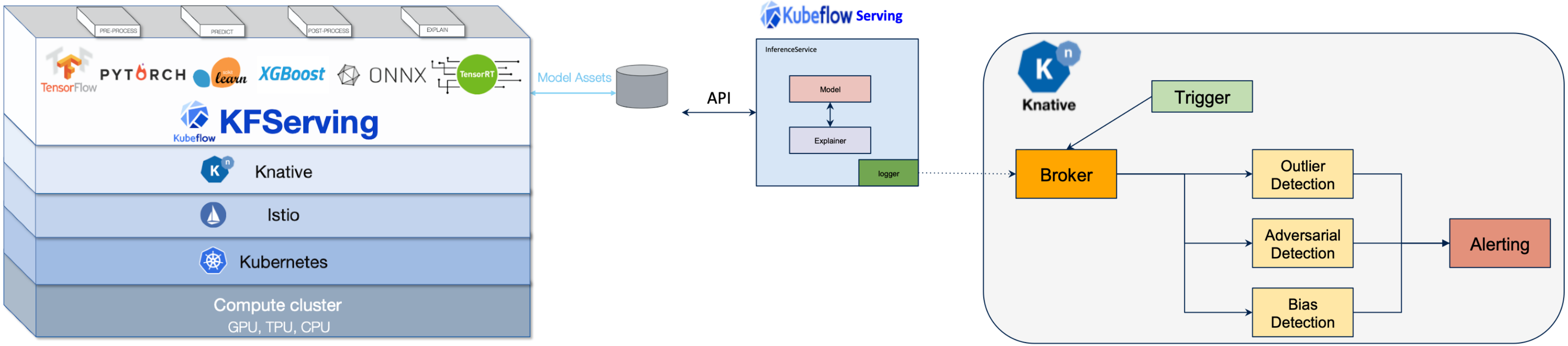
PutFile
PutFile 1.9.2
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

- Apache Nifi processor for AIF360

https://lfai.foundation/tag/apache-nifi/

Payload Logging to enable Trusted AI



KfServing Implementation (alpha):

- Add to any InferenceService Endpoint: Predictor, Explainer, Transformer
- Log Requests, Responses or Both from the Endpoint
- Simple specify a URL to send the payloads
- URL will receive CloudEvents

cloudevents

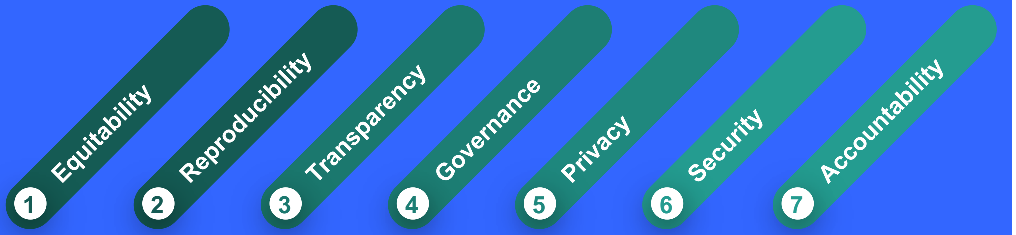
```
POST /event HTTP/1.0
Host: example.com
Content-Type: application/json
ce-specversion: 1.0
ce-type: repo.newItem
ce-source: http://bigco.com/repo
ce-id: 610b6dd4-c85d-417b-b58f-3771e532

<payload>
```

Principles Working Group

bit.ly/trusted-ai

LFAI Trusted AI Principles:

- Identified 7 principles with participating organizations, which were abstracted from existing initiatives, and have been structured to be relevant for any open source project. Evolving further
- Agreed that principles will focus "beyond the algorithms" into the OSS governance process



1 Equitability
2 Reproducibility
3 Transparency
4 Governance
5 Privacy
6 Security
7 Accountability

IBM

AT&T

orange™

ERICSSON

Tech Mahindra

Tencent

HUAWEI

ETHICAL ML INSTITUTE

AI for people

amdocs

MAIEI

LF AI | TRUSTED AI

KPMG’s AI in Control – Framework detail

Our AI in Control framework is aimed at identify the key risks, activities and control points that should be embedded in the governance construct to help ensure the unique risks and implications of AI are appropriately identified, managed, assessed and monitored on-going. Illustrative activities are shown below through the AI lifecycle from strategy through deployment of an AI solution.

	Strategy	Design	Model	Evaluate	Deploy & Evolve
Integrity Understand & Track Lineage Protect Reputation	— Alignment with strategy, business, compliance requirements — Available corporate policies and guidelines	— Use of data sources and inputs — Test for data quality and cleansing — Qualified SME’s involved — Data provenance check	— Evaluate training methodology or procedures — Feature provenance	— Experiment setup and configuration — QC model experiments and evaluation reports — Model accuracy and precision	— Runtime model metrics detection — Continuous training governance and assessment — Implementation control
Explainability Achieve Transparency Gain Confidence	— Adherence to data usage guidelines — Explanatory feature names	— Explainability requirements and schema/template defined	— Check for model metadata including attributes	— Explainability testing and acceptance	— Model improvement/change log — System Documentation — Report output evaluation
Fairness Be Inclusive & Ethical Ensure Appropriate Use	— Published list of allowed features	— Validation and quality check of ground truth — Bias verification, mitigation of ground truth (train, test & eval)	— Features compliance with policies, business requirements and regulations	— Model and Concept drift evaluation — Inclusiveness testing — Model risk scoring	— Setup for continuous monitoring of fairness and accuracy — Escalation process
Resilience Serve & continuously monitor Prevent Attacks	— Model usage guidelines, restrictions, and specifications — Defined model SLA’s — Required skills and support to manage and maintain	— Data usage guidelines; data privacy and protection	— Training data access protection and auditability — Model deployment/serving interoperability	— Use of approved frameworks, runtimes, and API’s — Security vulnerability and adversarial attack testing — Model and Concept drift detection	— Program execution — Model access and ACL — Continuous monitoring, protection & testing *(recalibration, incident response, BCP)* — Usage and feedback data protection — Model breach/Incident response plan
The Value of the Framework	Mapping to business needs Ethics and policy adherence Model measurement metrics	Understanding data lineage Detect imbalances in data Feature Analysis Bias detection & mitigation	Check feature compliance Modeling assumptions Auditability and Logging Hyperparameter changelog	Business Operational indicators Model explainability Evidence profiles Adversarial and security testing Continuous model monitoring	Production readiness Interoperability and serving Continuous protection Monitoring for metrics and drift Model and data governance

© 2020 KPMG LLP, a Delaware limited liability partnership and the U.S. member firm of the KPMG network of independent member firms affiliated with KPMG International Cooperative (“KPMG International”), a Swiss entity. All rights reserved.

Client case studies

Global bank

Challenge

A global bank wanted to drive increased effectiveness of how it was detecting fraud and at a reduced cost. This required the application of intelligent automation to detect more potential fraud cases, while at the same time detect fewer false positives.

Approach

KPMG's multidisciplinary team of forensics, banking and data science specialists built and deployed an AI solution that automated this business area. As this is a regulated business, controls were embedded from the start, e.g. a random forest algorithm was used to help ensure adequate Explainability and rigorous ongoing monitoring was implemented to demonstrate effectiveness.

Benefit

KPMG automated the work that 100 fraud prevention staff were doing previously, and at higher levels of accuracy than the human approach. Five people are now continuously monitoring the solution, and the remaining staff have been deployed elsewhere in the bank.

Major Credit card company

Challenge

With machine learning playing an ever greater role in their business decisions, including credit risk, fraud detection and marketing business functions, the Internal Audit team was unsure about their approach to auditing machine learning models. KPMG was asked to help evaluating their AI audit capabilities, skills, and procedures against leading practice.

Approach

Combining with many teams across the firm, KPMG has developed a solution which helps evaluate Internal Audit's capabilities, skills, and procedures against our AI In Control framework, in addition to providing roving training to internal audit teams in evaluating machine learning risk and controls.

Benefit

The credit card company gains greater comfort over their AI models internally, reducing the risk of failures in their models, which if left unaddressed, could lead to public brand damage and financial loss.

European Capital City

Challenge

A capital city in Europe is using an algorithm to identify, record, allocate and prioritize complaints coming from its citizens. The city wants to ensure that the allocation and prioritization of the complaints is unbiased, and is therefor looking to implement a system of managing risks to overcome these biases.

Approach

KPMG is using its AI In Control method to provide guidance to setup system of controls on the design, implementation and operation of the algorithm.

Benefit

The city will be able to disclose to its citizens that the algorithm is thoroughly controlled and reviewed by an objective party that assesses its design, implementation and operation.

studies

International brewery company

Challenge

An international brewery company has been actively pursuing the added value of robotics in the business and underlying IT landscape. Currently the company is exploring to implement advanced robotics opportunities using machine learning in a controlled way

Our Approach

Setup a control and governance framework for solutions based on robotics and machine learning including detailed risks, controls and tests that help to put advanced robotics/analytics models in production in a controlled manner

Benefit

A ready-to-go control framework on the design, development, test, deployment, management lifecycle of advanced robotic models is in place which can be leverage to put advanced robotics solutions into production in a controlled manner

Online travel agency

Challenge

An online travel agency is running an algorithm to place bids on online advertisements. They have reason to believe it is working very well. Yet they are looking to get additional insights on further improvements, including a risk assessment for future robustness

Our Approach

KPMG is using its AI In Control method to perform an independent review on the design and implementation of the algorithm, including the companies governance model to ensure the continuous operation (robustness) of the algorithm

Benefit

Independent review of an algorithm and its governance and control model, including points of improvement both for performance and for future robustness

Responsible AI at Orange

Under construction regulation in Europe

> *"In my first 100 days in office, I will put forward legislation for a coordinated European approach on the human and ethical implications of artificial intelligence."* – Ursula von der Leyen, European Commission President-elect

February 2020 White paper

- Awaited update of existing Product Liability Directive
- Differentiation beween high-risk AI and other scenarios based on activity sector and specific usage
- Common Risk analysis Framework needed
- prior conformity assessment would be mandatory for high-risk applications, voluntary label for others

Ongoing consultation : Ecosystem of Excellence and Ecosystem of Trust

https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Is Open Source enough?

- Open Source projects run by a single individual or controlled by a single vendor are quite closed in their governance.
- Projects delaying or not allowing outside contributions
- Projects welcoming of outside contributions, but not providing leadership roles to set technical strategy and direction.
- Projects controlled by a single individual or organization present a greater risk and lower the opportunity for collaboration and innovation.

Benevolent Dictator

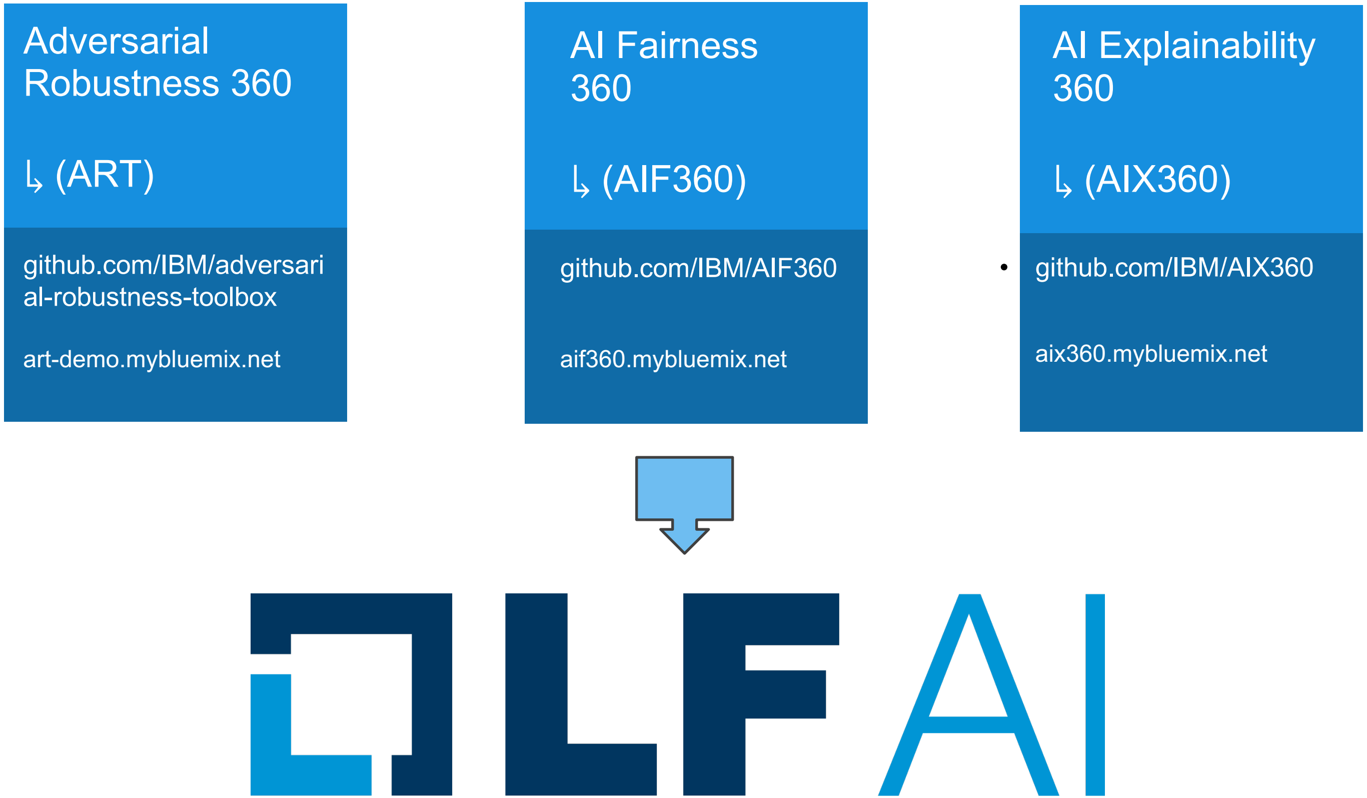
We need Open Governance

We need a neutral foundation that holds the copy rights and associated marks

- Reduces risk of project abandonment.
- Reduces risk of unilateral project license changes
- Eliminates single-vendor control
- Creates a real sense of ownership by the community members
- Gives a safe place to innovate

More details:
https://developer.ibm.com/articles/open-governance-community

Announcing: Moving Trusted AI projects in Open Governance to LFAI



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

Join the mission!

Linux Foundation AI
Trusted AI Committee

bit.ly/trusted-ai

https://github.com/IBM/AIX360

AI Explainability 360

Interpret and explain machine learning models.

https://github.com/IBM/adversarial-robustness-toolbox

Adversarial Robustness 360

Defend against adversarial attacks and make AI systems more secure

https://aif360.mybluemix.net/

AI Fairness 360

Open Source Toolbox to Detect and Mitigate Bias

- *Demos & Tutorials on Industry Use Cases*
- *Comprehensive Toolbox*
- *75+ Fairness metrics*
- *10+ Bias Mitigation Algorithms*
- *Fairness Metric Explanations*

Trusted AI @ IBM

ibm.biz/trusted-ai

(https://www.research.ibm.com/artificial-intelligence/trusted-ai/)