

LFAI & Data

Webinar -

Trusted AI Principles – Tools and Techniques

[Trusted AI Committee - Principles Working Group](#) (where you will find the slides and materials)

Register <https://tinyurl.com/trustedAI>

27 October 2021

 LFAI & DATA



Trusted AI Principles - Tools and Techniques

Join us at 10am US Eastern October 27, 2021 to meet with Souad Ouali Chair of the Trusted AI Principles Working Group at the LF-AI & Data & members of the Working Group - Hear about tools and techniques for the RREPEATS Principles



Register : <https://tinyurl.com/trustedAI>



Souad Ouali, Head of interoperators relationships Orange
Co-Chair Trusted AI Committee, LF-AI



Layla Li, Co-founder & CEO,
KOSA



Sarah Luger, Leading strategic AI/ML/NLP startups & technologies engagement, Orange



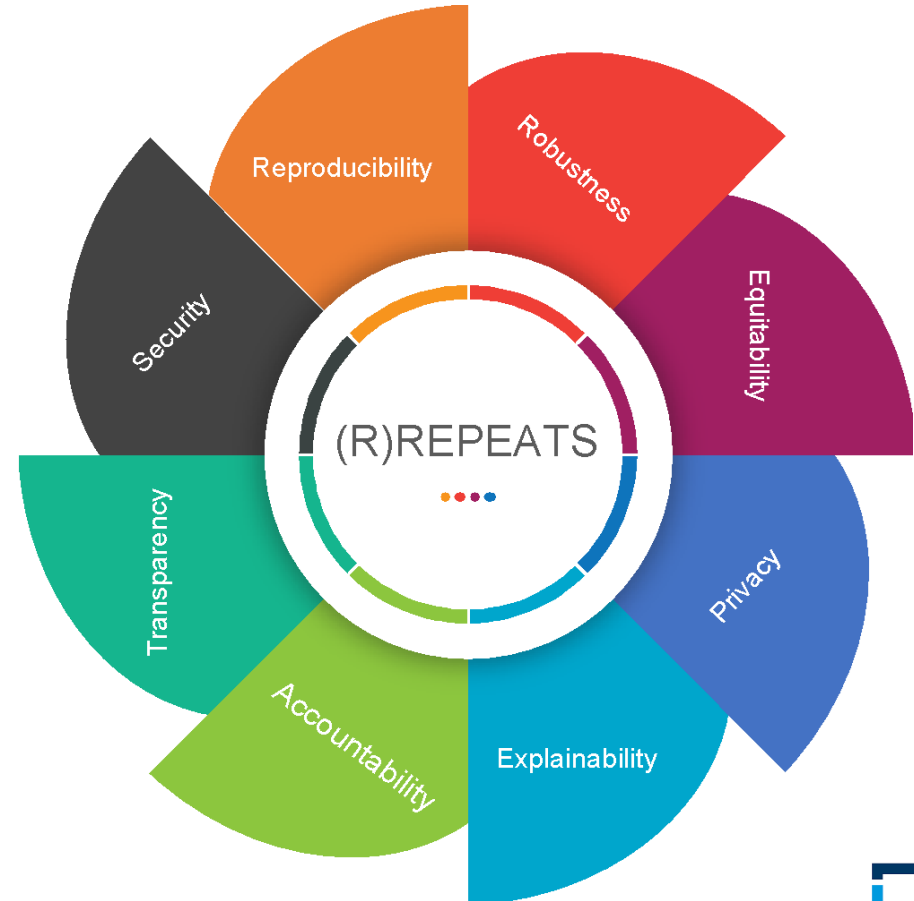
Sri Krishnamurthy, CEO,
QuantUniversity



Animesh Singh, Distinguished Engineer, IBM
Co-Chair Trusted AI Committee, LF-AI



François Jézéquel, Director of Business Development,
Orange Fab



Agenda

- Opening & intro to the session, Souad Ouali, Orange
- The RREPEATS Principles Overview, Layla Li, KOSA; Sarah Luger, Orange
- Operationalizing Trusted AI in Finance using the QuSandbox, Sri Krishnamurthy, QuantUniversity
- Trusted AI Tools (AI Fairness, AI Explainability, Adversarial Robustness etc) and RREPEATS, Animesh Singh, IBM
- Emerging DataOps activities at the LF-AI, Trusted AI and RREPEATS, Animesh Singh, IBM
- Summary & Call to Action, François Jézéquel, Orange



Session Host: Souad Ouali
Head of interoperators relationships Orange - Conseil /
Responsable de relations inter opérateurs chez Orange - Conseil

LFAI & Data

The Trusted AI Principles - Tools and Techniques

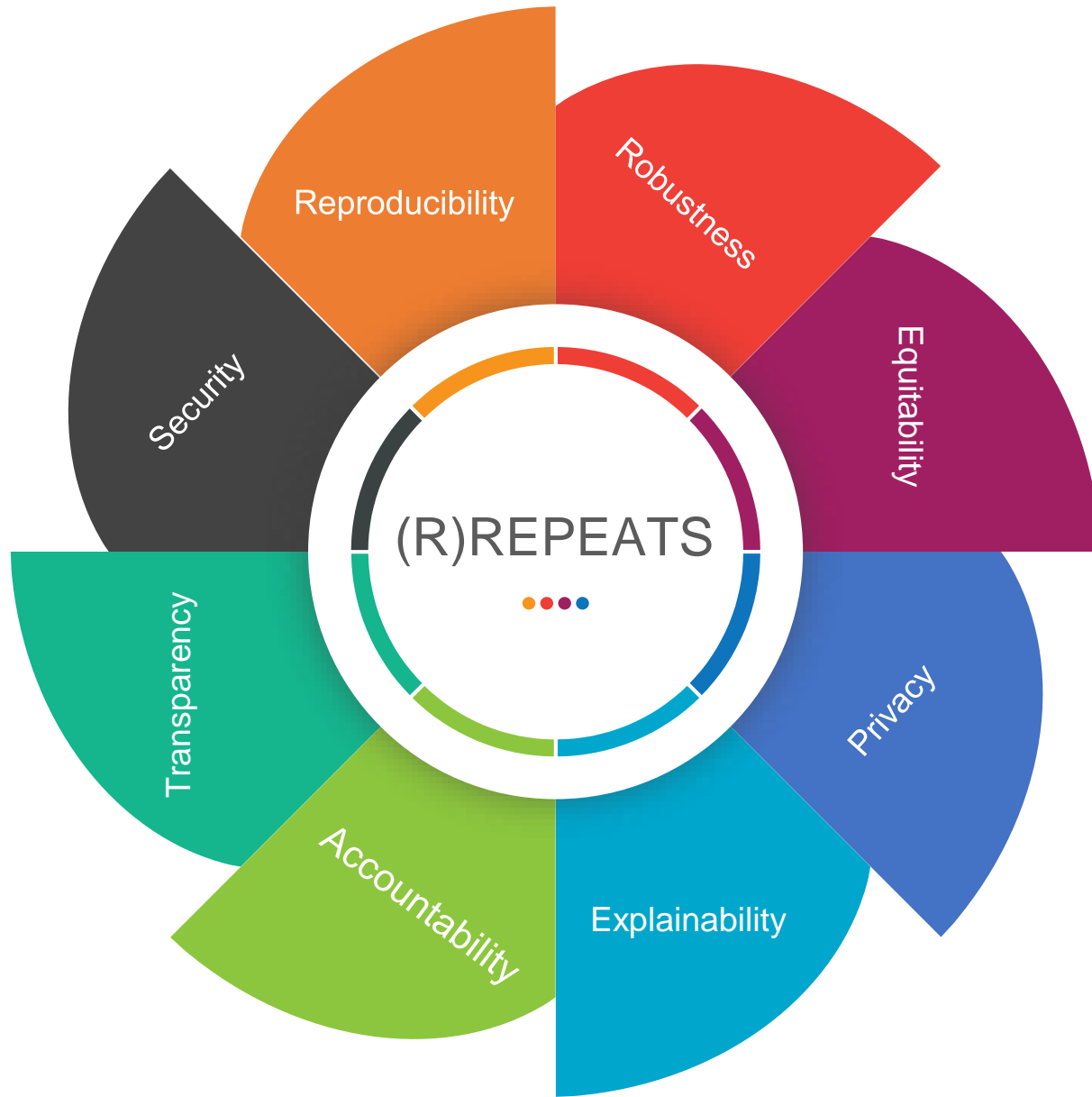
The RREPEATS Principles Overview



Layla Li, Co-founder &
CEO, KOSA



Sarah Luger, Leading strategic
AI/ML/NLP startups & technologies
engagement, Orange

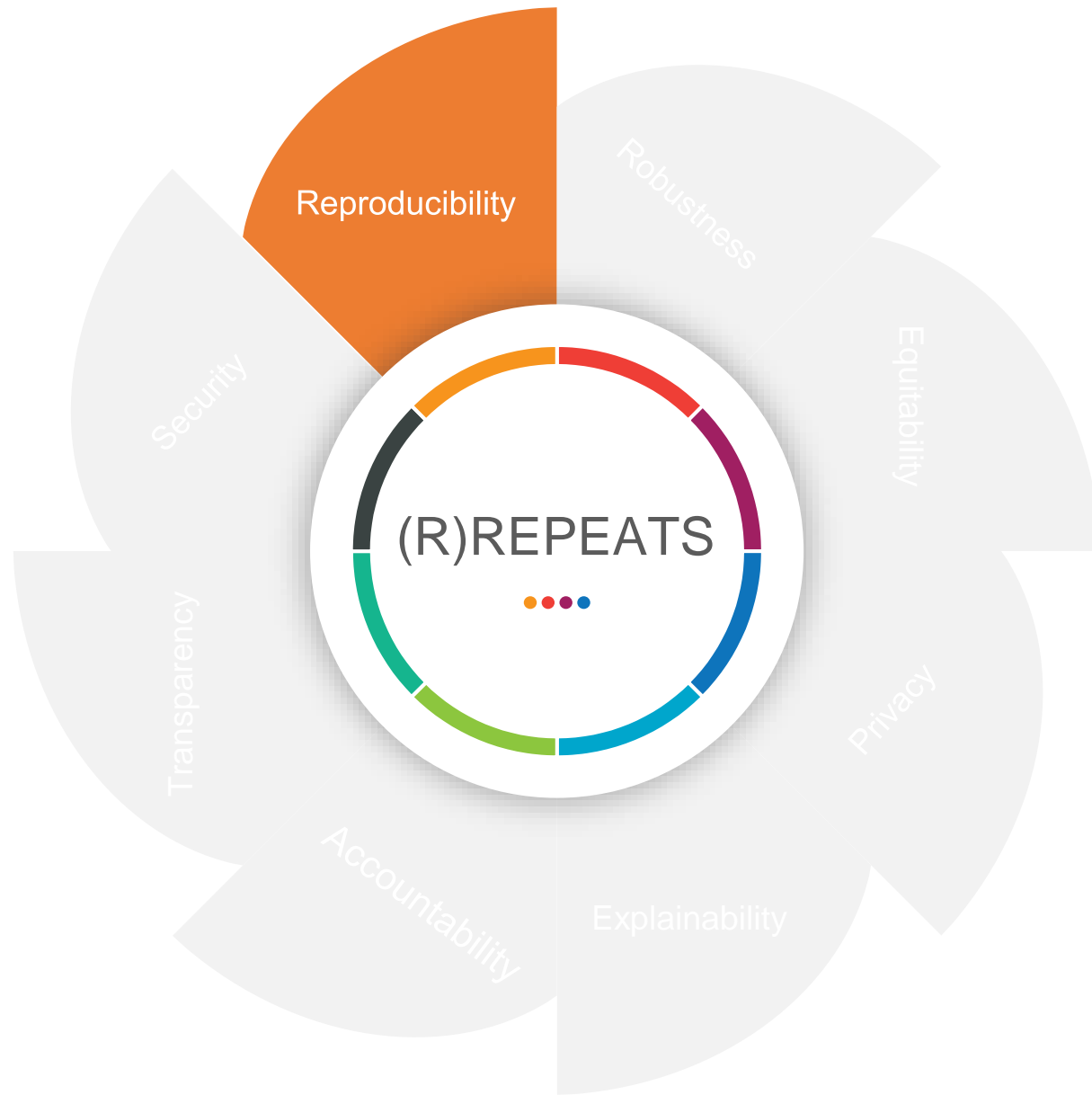


The 8 LFAI Principles for Trusted AI – (R)REPEATS

The principles are of equal importance and value.

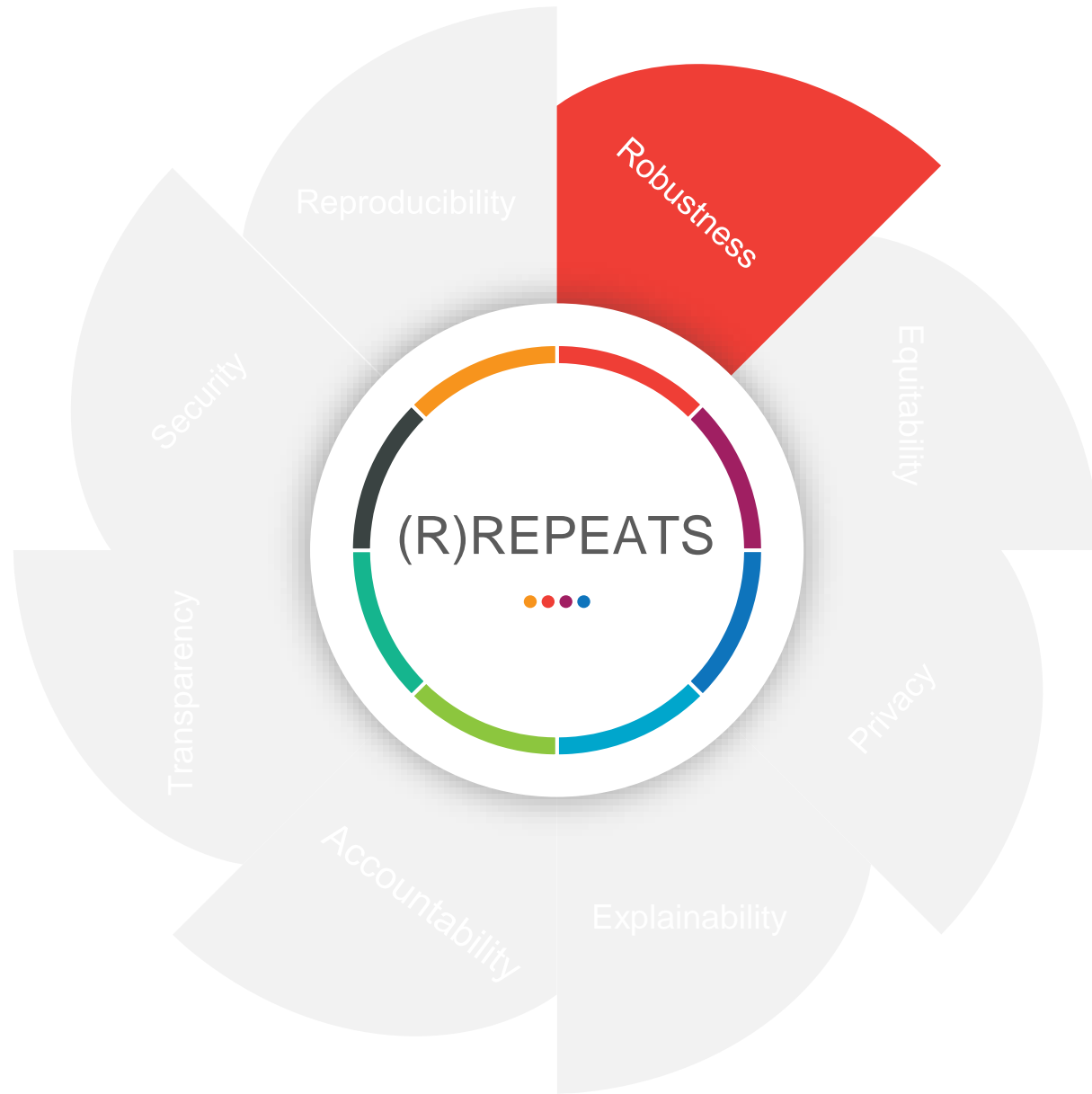
No principle is of higher priority than another.

The principles are related to each other.



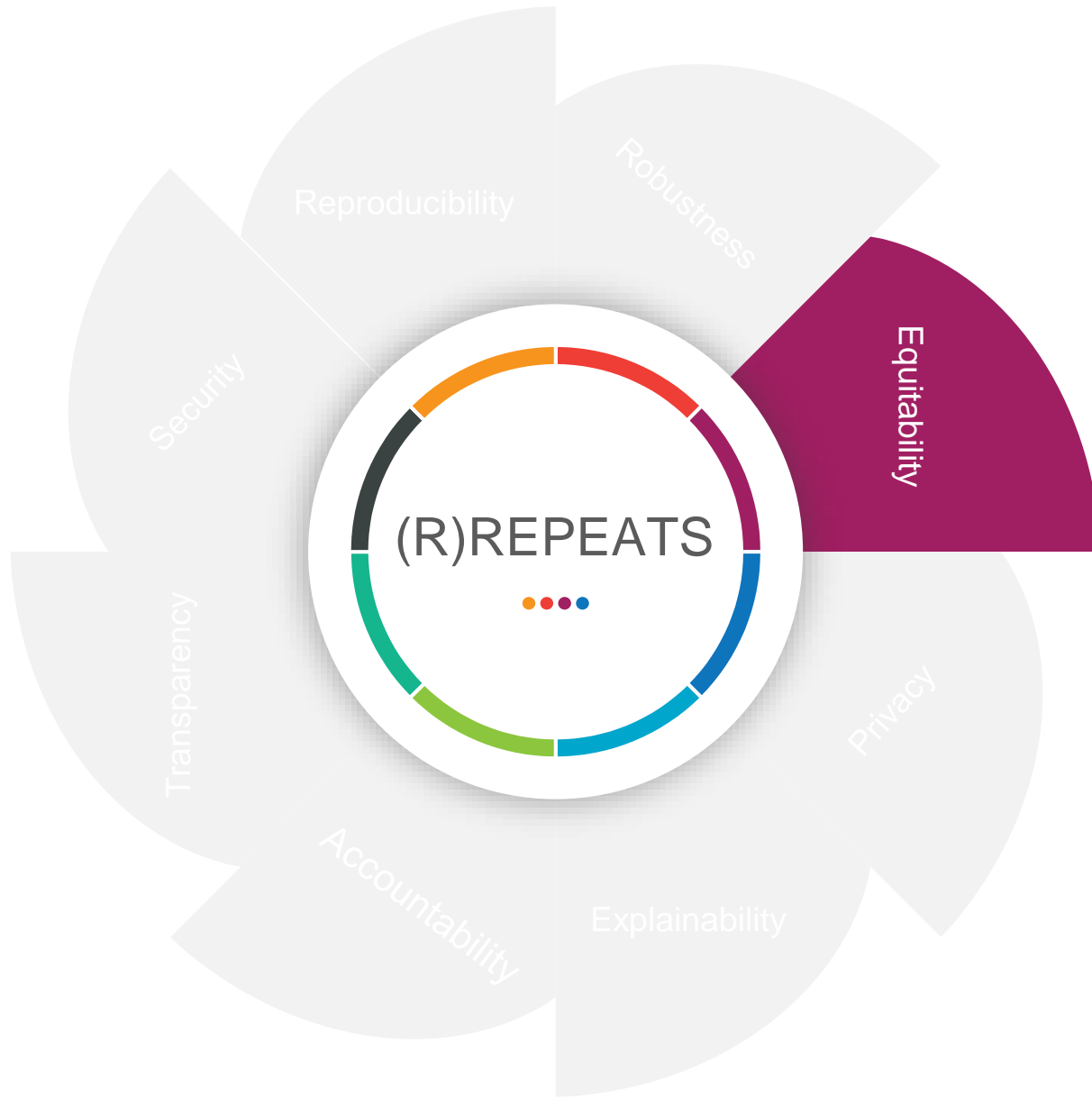
Reproducibility

Adhering to this principle will ensure the reliability of the results or experiences produced by any AI.



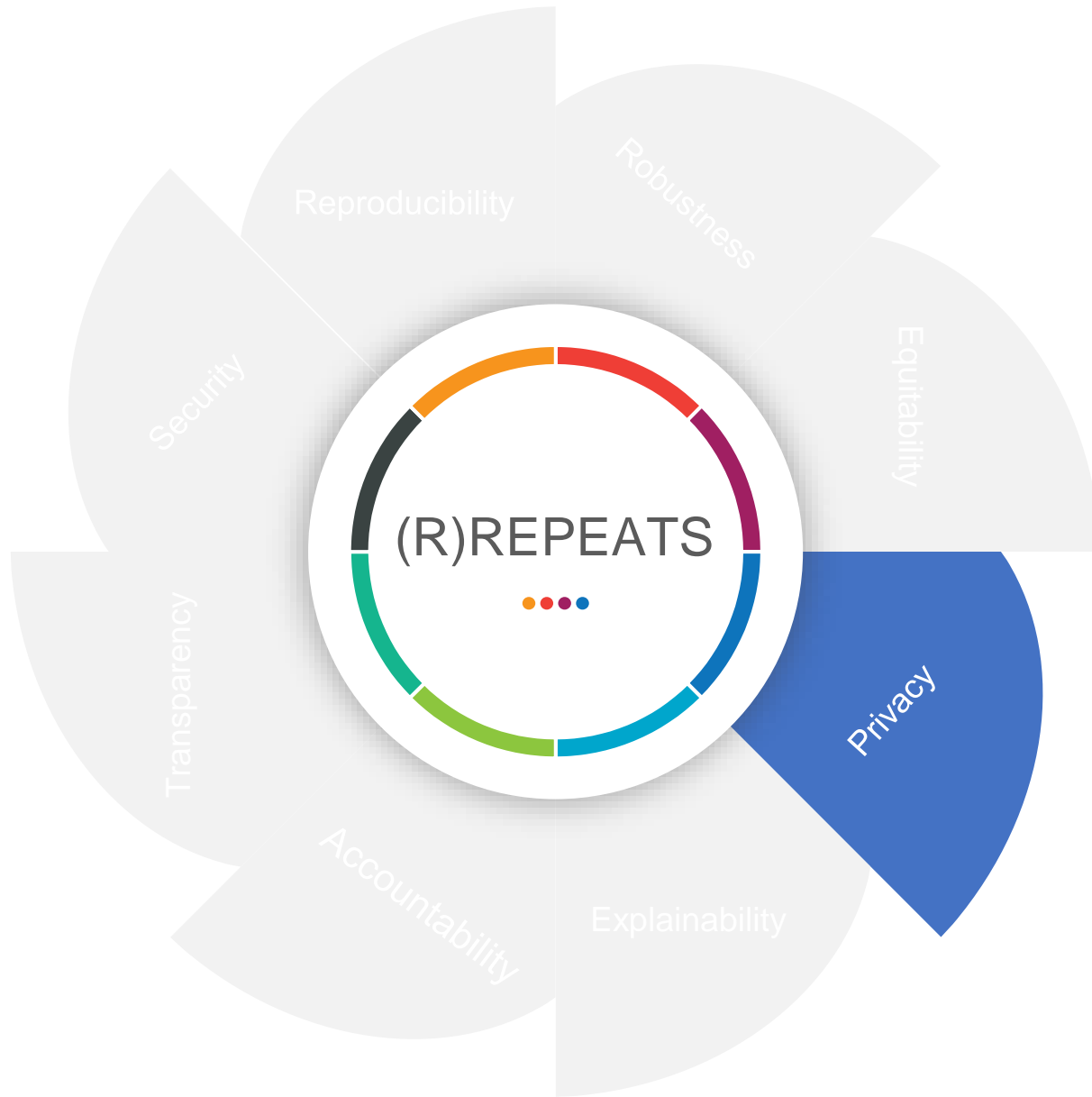
Robustness

Ensure stability, resilience, and performance of the systems in changing ecosystems.



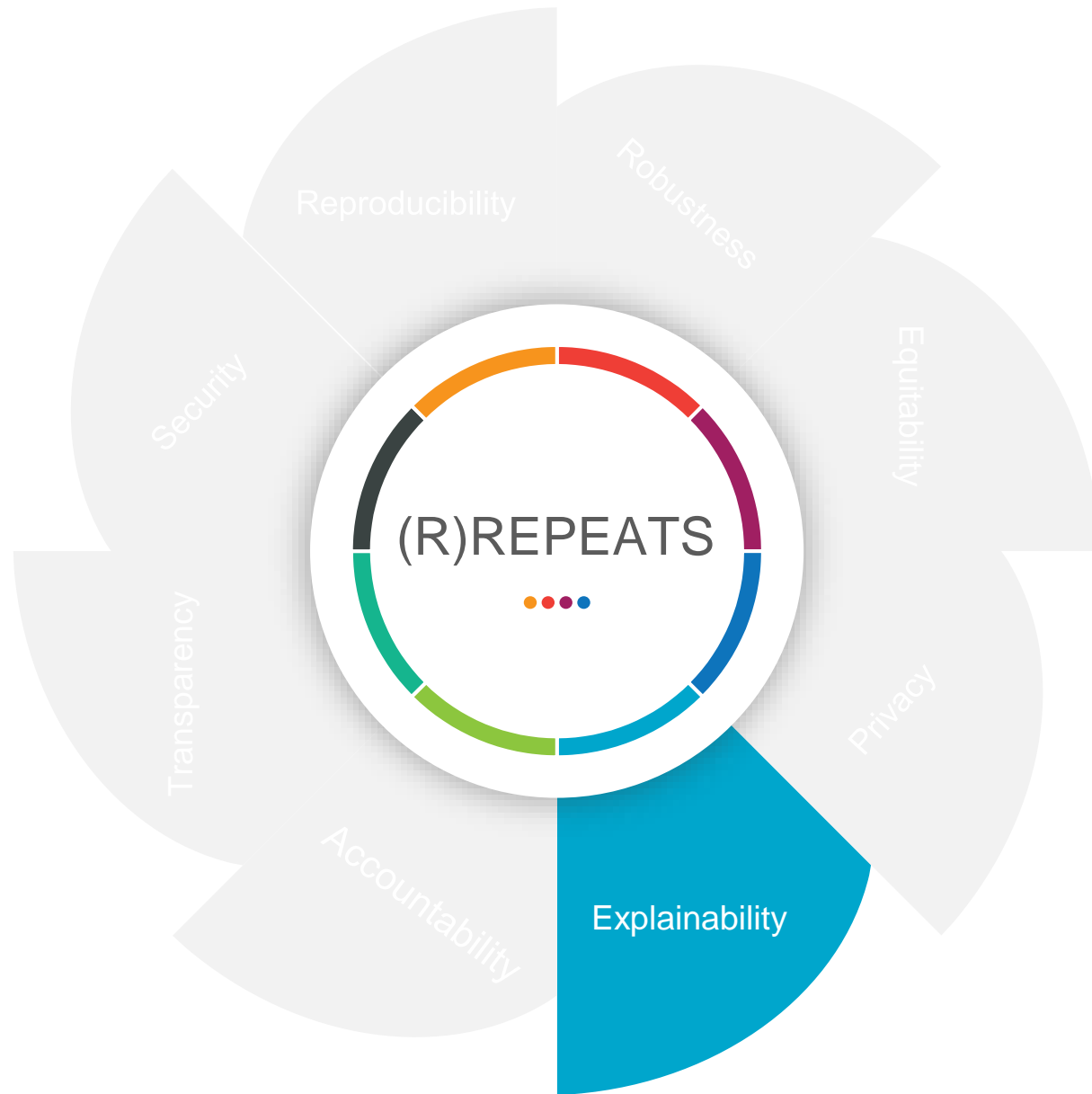
Equitability

Avoid intended or unintended bias
and unfairness that would
inadvertently cause harm



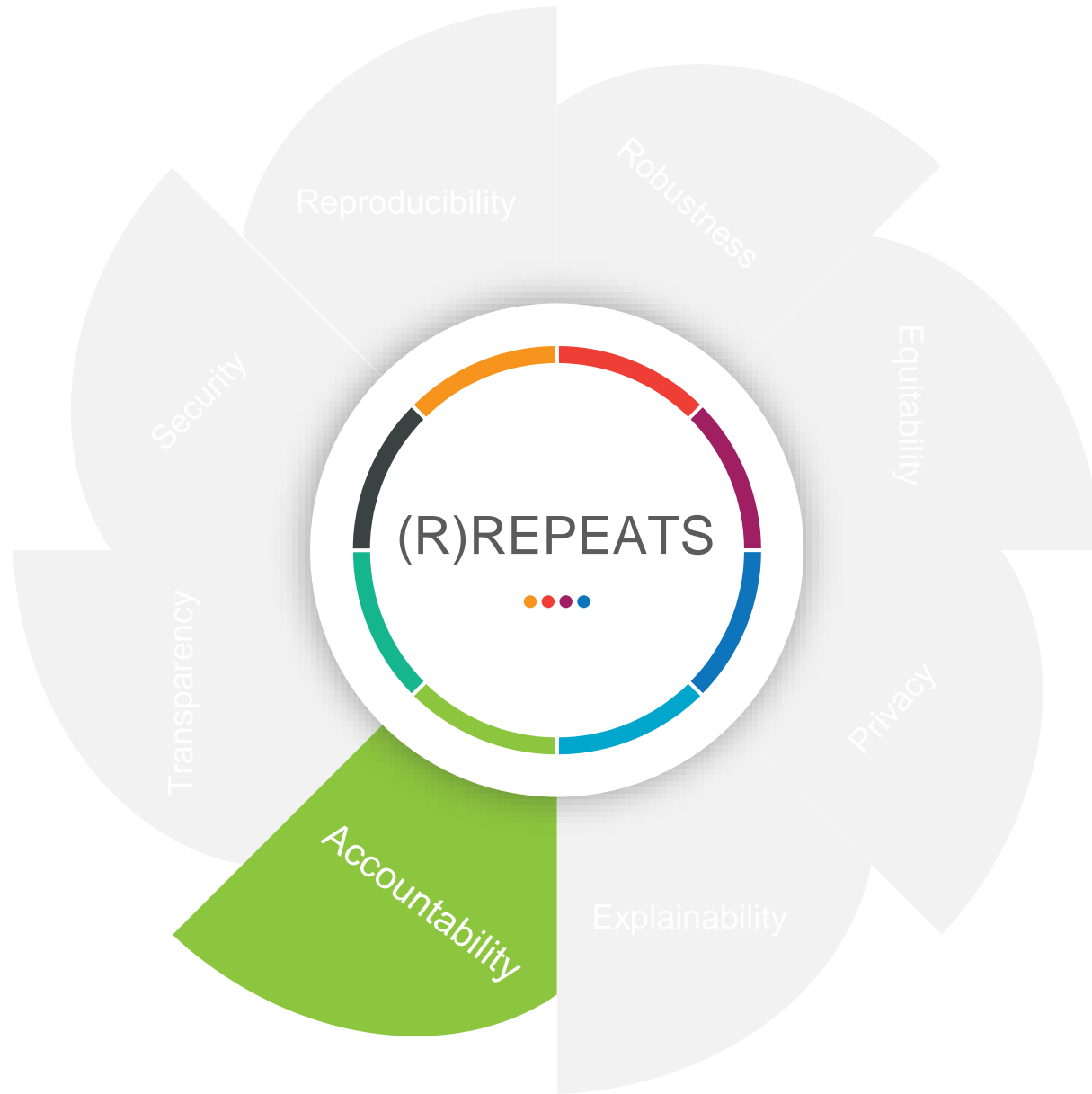
Privacy

Guarantee privacy and data protection throughout a system's entire lifecycle.



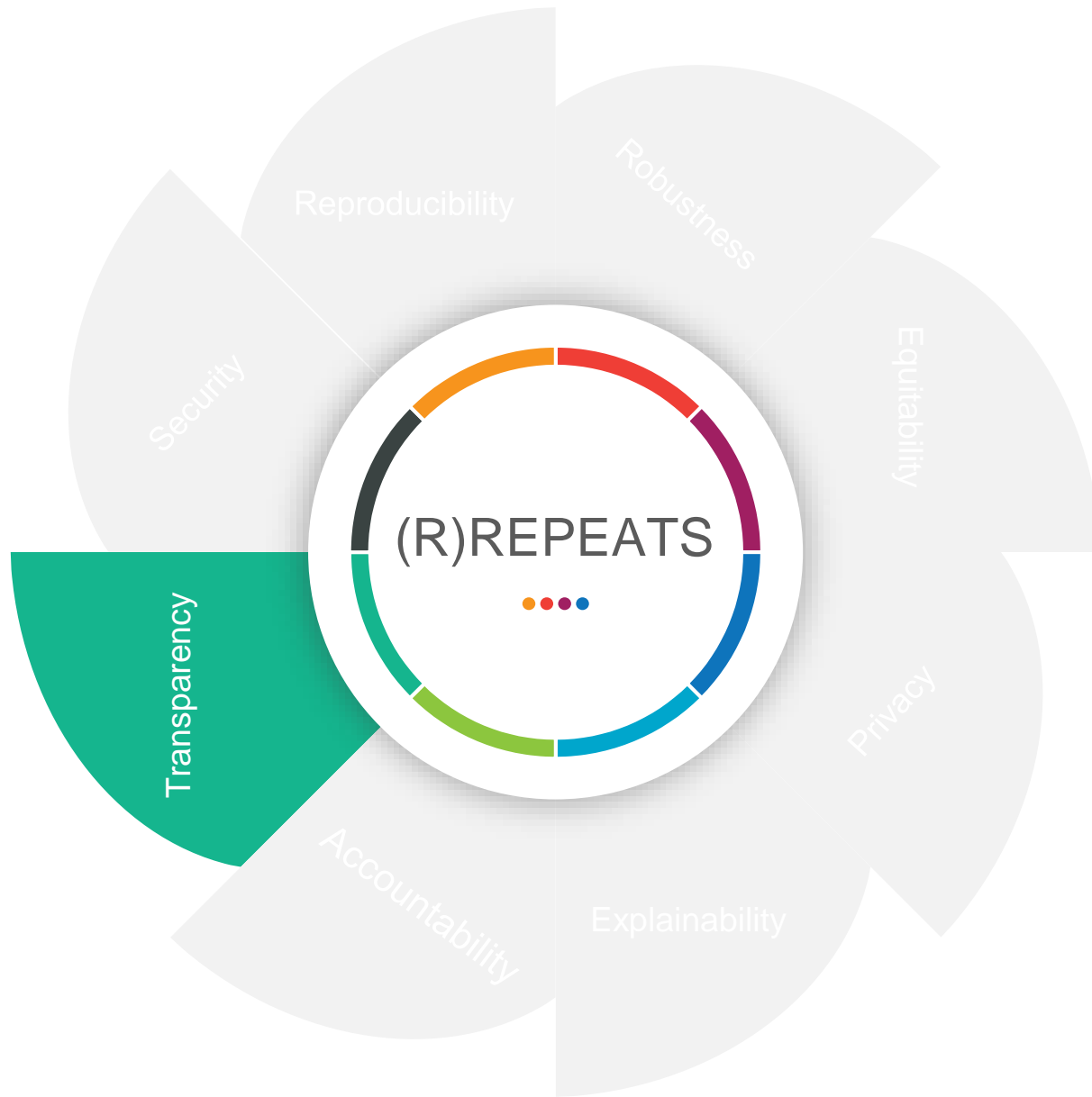
Explainability

Explain how AI “blackbox” make decisions in transparent manners



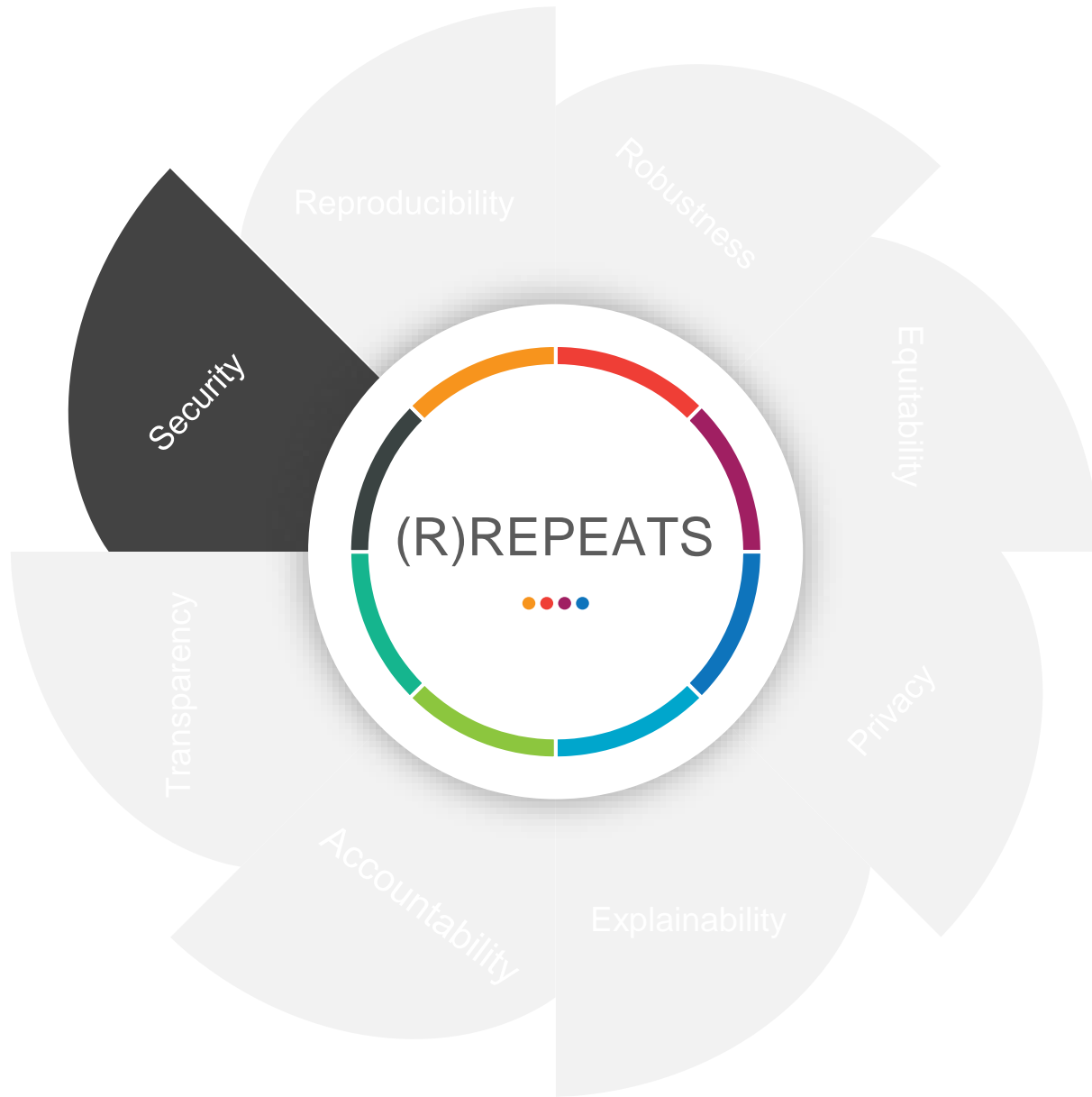
Accountability

“Humans-in-the-loop” to take responsibility and plan for actions of AI



Transparency

Users should be informed of when they interact with AI and understand AI-based outcomes



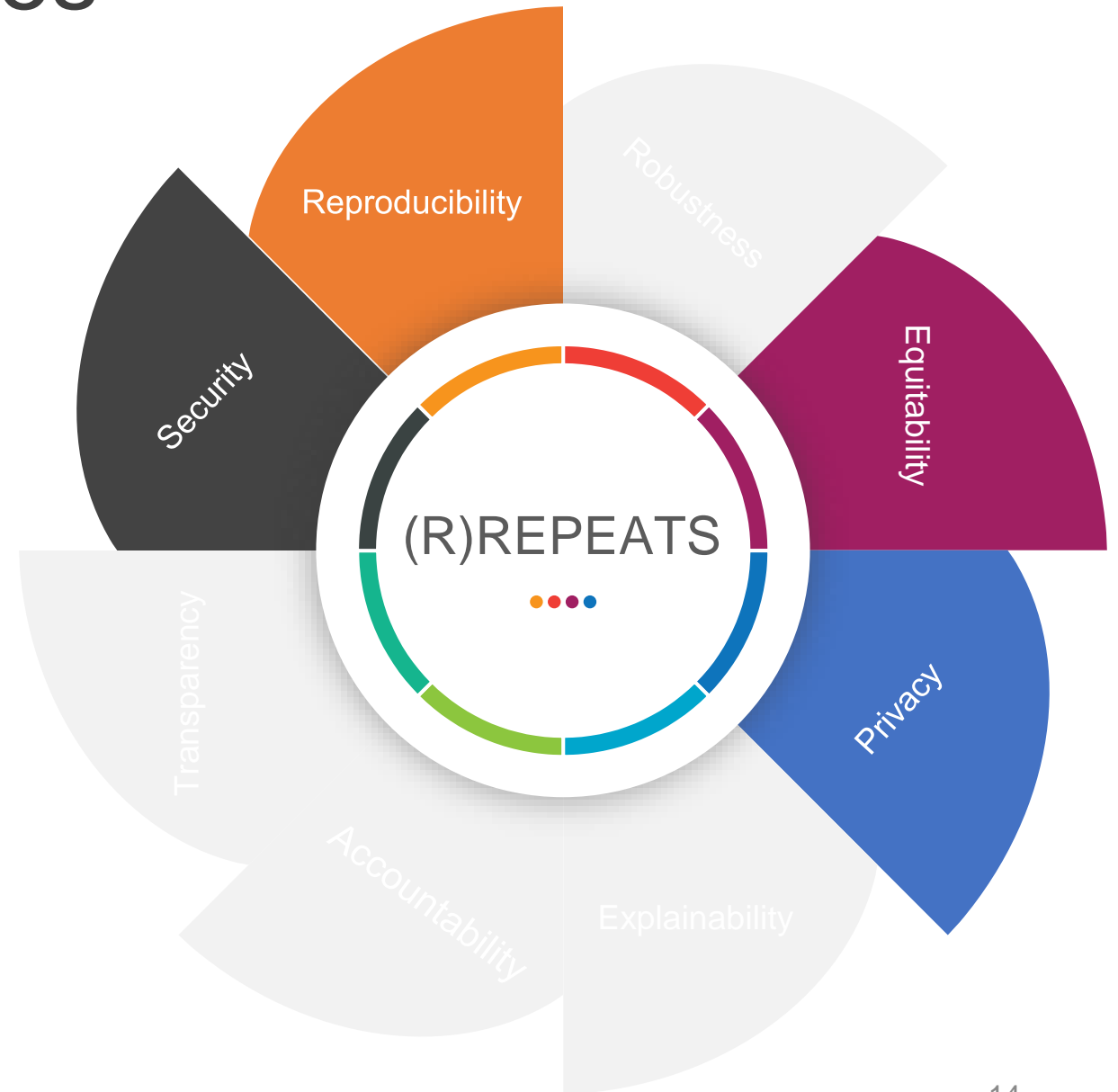
Security

Safety of AI should be tested and assured across the entire lifecycle

The Trusted AI Principles - Case in Point

[Uber sued by drivers over automated “robot-firing”.](#)

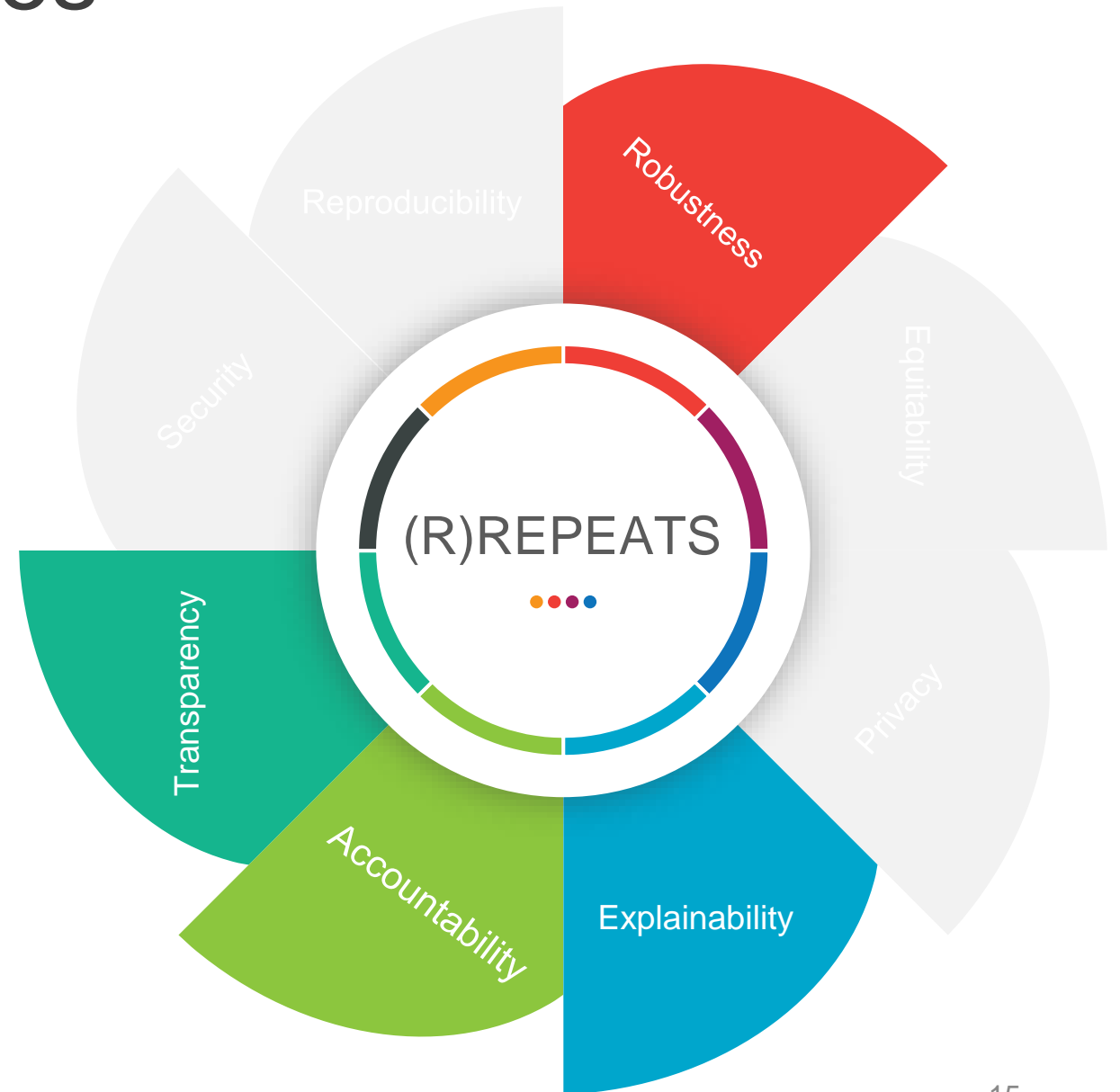
The UK court ruled in favor of the drivers who have allegedly been wrongly accused of fraudulent activity by the company’s algorithms and immediately had their accounts terminated without a right of appeal.



The Trusted AI Principles - Case in Point

[Billions were cut off from Facebook, Instagram, WhatsApp for hours.](#)

Changes to its underlying internet infrastructure that coordinates the traffic between its data centers caused one of the biggest tech companies to disappear from the internet for hours.



LFAI & Data

The Trusted AI Principles – Tools and Techniques

Operationalizing Trusted AI in Finance using the QuSandbox

 LFAI & DATA



Sri Krishnamurthy, QuantUniversity



QuantUniversity, LLC

www.quantuniversity.com

LF AI
& DATA

Operationalizing Trustworthy AI in Finance with the QuSandbox

Presented By:

Sri Krishnamurthy, CFA, CAP

sri@quantuniversity.com

www.quantuniversity.com

October 27th 2021

LF AI Trustworthy AI principles
Meet

QuantUniversity

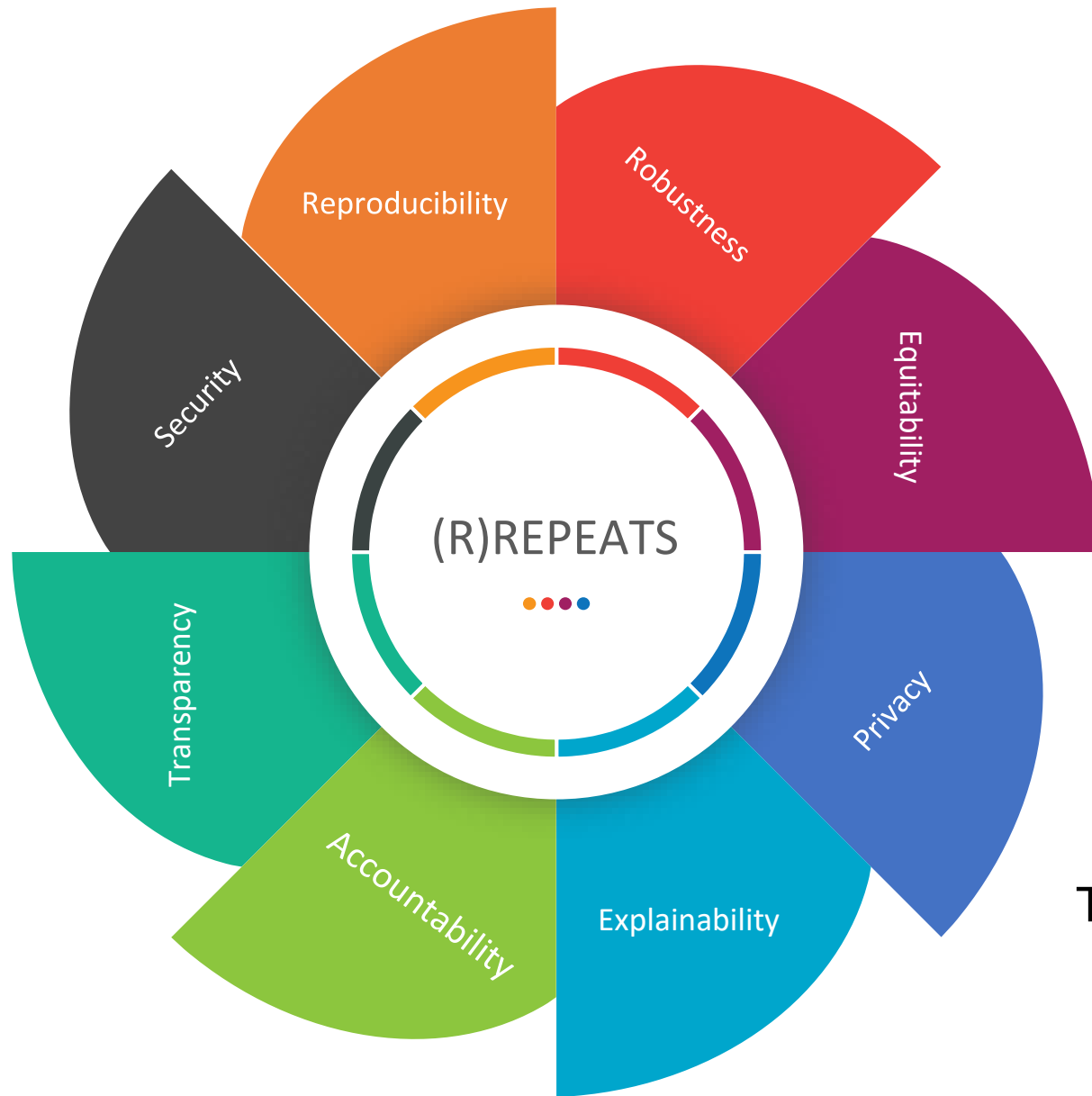
- Boston-based Data Science, Quant Finance and Machine Learning AI Risk Advisory
- Specialties include Algorithmic audits, Model Risk Management and AI project enablement
- Training programs for more than 1000 students in Quantitative methods, Data Science, Machine Learning and AI Risk Management
- Building QuSandbox, a platform for AI and Machine Learning Governance and Risk Management
- Associate Member of the LFAI since 2021



QuantUniversity, LLC

The logo features the text "QuSandbox" in a white, sans-serif font. The "Qu" is positioned to the left of "Sandbox", and a white square outline is placed behind the "u" in "Qu".

QuSandbox



The 8 LFAI Principles for Trusted AI – (R)REPEATS

The principles are of equal importance and value.
No principle is of higher priority than another.
The principles are related to each other.

The Implementation GAP!

“Recognizing this, actors across industry, government and civil society have rolled out an expanding array of ethical principles to guide the development and use of AI – over 175 to date.

*While the explosive growth in AI ethics guidelines is welcome, it has created an **implementation gap** – it is easier to define the ethical standards a system should meet than to design and deploy a system to meet them”*

<https://www.weforum.org/projects/global-ai-action-alliance>





QuantUniversity, LLC
www.quantuniversity.com

Trustworthy AI – Our Approach



Request DEMO at
info@qusandbox.com

EDUCATION
QUACADEMY

EXPERIMENTATION
ON
QUTOOLBOX

ENABLEMENT
QUSANDBOX

QuantUniversity Course Catalog

- Just Enough Python for Data Science**
Understand the core Python constructs needed to build scalable data science and machine learning applications.
- Machine Learning and AI for Financial Professionals**
Learn how to build pragmatic AI and ML applications with case studies in finance.
- Model Risk Management for Machine Learning Models**
Address the key model risk management and validation challenges when deploying data science and machine learning models in the enterprise.
- The Fintech Bootcamp: The 8 Facets of FinTech**
The FinTech Bootcamp: The 8 Facets of FinTech
- Algorithmic Auditing**
Algorithmic Auditing
- Risk & ML Models: Stress, Scenario Testing & Evaluation**
RISK & ML MODELS: STRESS, TESTING & EVALUATION

Qutoolbox interface showing various service cards:

- MATLAB (Version 1.0.0) - Run MATLAB on QuSandbox
- Microsoft (Version 1.0.0) - Try Microsoft APIs
- Jupyter (Version 1.0.0) - Jupyter Notebook Data Science Stack
- Google (Version 1.0.0) - Try Google APIs
- NVIDIA RAPIDS (Version 1.0.0) - Run RAPIDS on QuSandbox
- Julia (Version 1.0.0)
- AWS (Version 1.0.0)

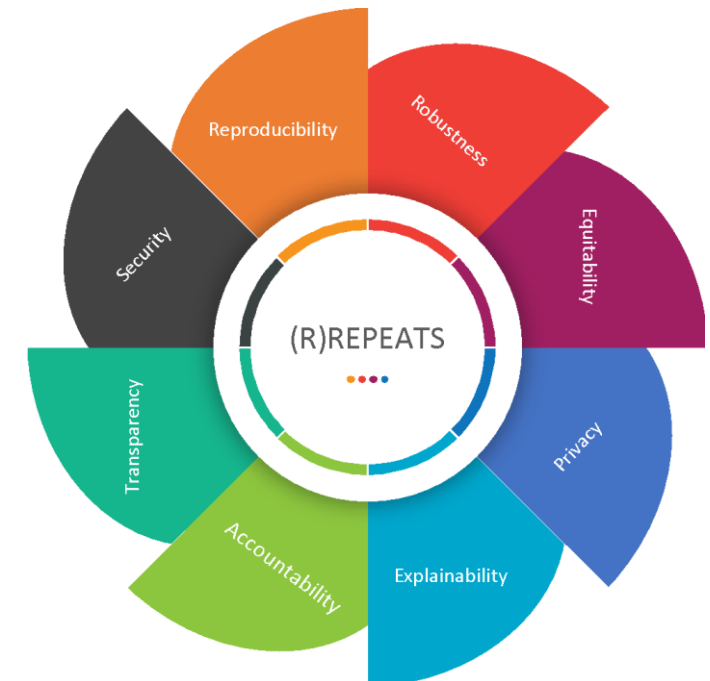
QuSandbox Projects dashboard showing a grid of project cards:

- Summary (Summary card)
- Audit Checklist (Audit Checklist card)
- Data (Data card)
- Model (Model card)
- Environment (Environment card)
- Pipeline (Pipeline card)
- Explainability (Explainability card)
- Fairness (Fairness and Bias)
- Findings (Findings card)
- Recommendations (Recommendations)
- Report (Report card)

Project Details:

- Project Name: M1_Sixam
- Project Description: This model predicts whether breast cancer is benign or malignant based on image measurements.
- Project ID: 0a371e8d31544703a3a8f8c8d8a23d6
- Experiment Name: M1_Sixam Experiment
- Experiment Description:
- Experiment ID: 0880cc037694288be1144f25c5186d

How QuSandbox addresses the LFAI principles



Search



ALL

CREATE PROJECT

ML-fairness-gym



Version: 1.0.0

Exploring Long-Term Impacts of MLS



Fairlearn



Version: 1.0.0

Assess fairness and mitigate unfairness



AI Fairness 360



Version: 1.0.0

Detect and mitigate bias in ML



Audit-AI



Version: 1.0.0

Help implements fairness-aware machine learning algorithms



Sri Krishnamurthy

QuProjects

QuToolBox

Data

Explore

Data Processing

Modeling Tools

Models

Explain

Fairness and Bias

Security

Report

Case studies

QuAcademy



QuantUniversity Course Catalog



Just Enough Python for Data Science

Understand the core Python constructs needed to build scalable data science and machine learning applications



Machine Learning and AI for Financial Professionals

Learn how to build pragmatic AI and ML applications with case studies in finance



Model Risk Management For Machine Learning Models

Address the key model risk management and validation challenges when deploying data science and machine learning models in the enterprise



The FinTech Bootcamp: The 8 Facets of FinTech



Algorithmic Auditing



RISK & ML MODELS: STRESS, TESTING & EVALUATION



QuSynthesize_GAN

The QuSandbox Synthesize API aims at generating synthetic data maintaining the features of the original dataset to solve kinds of data problems. This API focuses on a specific use case that is generating synthetic VIX data using GAN, which could be easily implemented on other stock like datasets.


QuSynthesize 0.2 OAS3

[/openapi.json](#)

The synthetic data generation API, presented by Quant University

Authorize 

test Tests for API access 

dataset Operations on the synthesized datasets 

POST

/gan/simulation Simulation

Privacy 

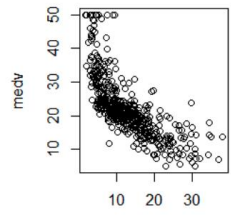
Synthetic Data Hub

The Market Place for Synthetic Data
for your AI and Machine Learning Applications



Anonymity and Privacy

Leverage anonymized datasets for training your machine learning models



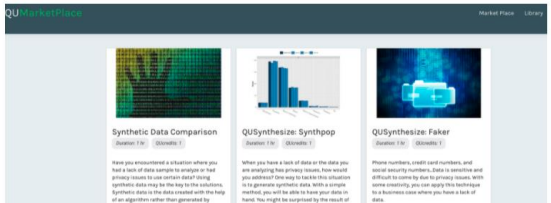
Data Augmentation

Supplement traditional datasets with new and varied datasets



Robust and Tested APIs

Curated and Tested using QUSandbox. Data spec sheets available for datasets



- Synthetic data Hub: A synthetic data marketplace to enable testing, benchmarking and replicability.
- Rich visualization and automated reporting and scorecards to quantify and monitor AI risk.
- Testing framework to run stress, scenario tests for ML models ; accelerated by GPUs



**SYNTHETIC
DATA HUB**

WWW.SYNTHETIC DATA HUB.COM

QuSandbox

- QuToolBox
- Data
- Explore
- Data Processing
- Modeling Tools
- Models
- Explain**
- Fairness and Bias
- Report
- Case studies

⋮ **Explain** ⓘ

Search



CREATE PROJECT

UQ360

Version: 1.0.0

QuSandbox

Adversarial Robustness Toolbox

ART

Version: 1.7

QuSandbox

Adversarial Robustness Toolbox

H2O

Version: 1.0.0

QuSandbox

Run H2O on QuSandbox

AIX-360

Version: 1.0

QuSandbox

AIX-360 Explainability Toolkit

Shapash

Version: 1.0

QuSandbox

Make machine learning interpretable and understandable by everyone

Manifold

Version: 1.0.0

QuSandbox

Model-agnostic visual debugging tool





Projects



Summary

Summary card



Audit Checklist

Audit Checklist card



Data

Data card



Model

Model card



Environment

Environment card



Pipeline

Pipeline card



Explainability

Explainability card



Fairness

Fairness and Bias



Findings

Findings card



Recommendations

Recommendations



Report

Report card

Project Name: ML - Sklearn









Project Description: This model predicts whether breast cancer is benign or malignant based on image measurements.

Project ID: 0d371a9d315447d3af8e9c8adaac23e6













Experiment Name: ML - SKLearn Experiment

Experiment Description:

Experiment ID: 59b00d287b69428b8e6144df25c51d6d

-  Sri Krishnamurthy
-  QuProfile
-  Projects
-  QuApiVault
-  Log Out
-  QuToolBox
-  QuModelStudio
-  QuAcademy

 **Projects**

 Summary Summary card 	 Audit Checklist Audit Checklist card	 Data Data card	 Model Model card
 Environment Environment card	 Pipeline Pipeline card	 Explainability Explainability card	 Fairness Fairness and Bias
 Findings Findings card	 Recommendations Recommendations	 Report Report card	

Project Name: ML - Sklearn

Project Description: This model predicts whether breast cancer is benign or malignant based on image measurements.

Project ID: 0d371a9d315447d3af8e9c8adaac23e6

Experiment Name: ML - SKLearn Experiment

Experiment Description:

Experiment ID: 59b00d287b69428b8e6144df25c51d6d

Search



ALL

CREATE PROJECT

MintNV Version: 1.0

QuSandbox

📄 🗑

CleverHans Version: 1.0

QuSandbox

📄 🗑

Counterfit Version: 1.0

QuSandbox

📄 🗑

AutoAttack Version: 1.0.0

QuSandbox

Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks

📄 🗑

Adversarial Robustness Toolbox Version: 1.0

QuSandbox

📄 🗑

TextAttack Version: 1.0.0

QuSandbox

📄 🗑

- Sri Krishnamurthy
- QuProjects
- QuToolBox
- Data
- Explore
- Data Processing
- Modeling Tools
- Models
- Explain
- Fairness and Bias
- Security**
- Report
- Case studies
- QuAcademy

MATLAB

Version: 1.0.0



Run MATLAB on QuSandbox



Microsoft

Version: 1.0.0



Try Microsoft APIs



Jupyter

Version: 1.0.0



Jupyter Notebook Data Science Stack



Google

Version: 1.0.0



Try Google APIs



NVIDIA-RAPIDS

Version: 1.0.0



Run RAPIDS on QuSandbox



Julia

Version: 1.0.0



AWS

Version: 1.0.0




edgar_pipeline 



Scraping-Environment

Scraping-Model

SCRAPING-BLOCK

ADD ENTITY 


Stage 1



Data-Processing-Environment

Data-Processing-Model

DATA-PROCESSING-BLOCK

ADD ENTITY 


Stage 2



Sentiment-Analysis-Environment

Sentiment-Analysis-Model

SENTIMENT-ANALYSIS-BLOCK

ADD ENTITY 


Stage 3



API-Comparison-Environment

API-Comparison-Model

API-COMPARISON

ADD ENTITY 


Stage 4



MATLAB-Analysis-Environment

MATLAB-Analysis-Model

MATLAB-ANALYSIS-BLOCK

ADD ENTITY 

Reproducibility

Logs

11/10/20 12:47 PM : fetched pipeline

PROJECT

EXPERIMENT

TESTPLAN

Project Name: ONNX Benchmarking
Project Description:
Project Brief Description:
Project ID: c7c7efc41c21447093fef96ddoc72c59
Project Version:

Model Lifecycle: ^

Summary
Summary Board

Environment
Environment Board

Data
Data Board

Model
Model Board

Explainability
Explainability Board

Fairness
Fairness Board

Deployment
Deployment Board

Monitoring
Monitoring Board

Testing: ^

Test
Test Board

StressTests
StressTests Board

ScenarioTests
ScenarioTests Board

WhatifAnalysis
WhatifAnalysis Board

ML Security Review : v

Algorithmic Assessment: v

TEST

REPORTS

NOTES

ISSUES

Test Plan

- Sri Krishnamurthy
- QuProjects
- QuToolBox
- Data
- Explore
- Data Processing
- Modeling Tools
- Models
- Explain
- Fairness and Bias
- Security
- Report
- Case studies
- QuAcademy

QuSandbox

Sri Krishnamurthy

QuProjects

QuToolBox

Data

Explore

Data Processing

Modeling Tools

Models

Explain

Fairness and Bias

QUReport Score Report credit risk

Id: 8cf00cf70ed9499a9b483362ad58eb4c
Version: 1.1
Date: 2021-05-14

Information
 Experiment: 8cf00cf70ed9499a9b483362ad58eb4c
 Owner: Sri Krishnamurthy
 Contact: info@qusandbox.com
 References: The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction

Final Score



QuSandbox

Sri Krishnamurthy

QuProjects

QuToolBox

QuAcademy

Findings Board

Recommendations Board

TESTPLAN

REPORTS

NOTES

ISSUES

Reports









Name *

Version *












Test Plan

- Document the test plan for the report
- How is the model tested (Full, sampled, regression tested)
- Comment on the testing infrastructure (CI/CD, test suites)



-  Sri Krishnamurthy
-  QuProfile
-  Projects
-  QuApiVault
-  Log Out
-  QuToolBox
-  QuModelStudio
-  QuAcademy

Projects

-  **Summary**
Summary card
-  **Audit Checklist**
Audit Checklist card
-  **Data**
Data card
-  **Model**
Model card
-  **Environment**
Environment card
-  **Pipeline**
Pipeline card
-  **Explainability**
Explainability card
-  **Fairness**
Fairness and Bias
-  **Findings**
Findings card
-  **Recommendations**
Recommendations
-  **Report**
Report card

Project Name: ML - Sklearn

Project Description: This model predicts whether breast cancer is benign or malignant based on image measurements.

Project ID: 0d371a9d315447d3af8e9c8adaac23e6

Experiment Name: ML - SKLearn Experiment

Experiment Description:

Experiment ID: 59b00d287b69428b8e6144df25c51d6d

Request DEMO at
info@qusandbox.com

Speaker bio



Sri Krishnamurthy
Founder and CEO
QuantUniversity



- AI advisory focused on AI Risk, Governance and enablement
- Prior Experience at MathWorks, Citigroup and Endeca and 25+ financial services and energy customers.
- Columnist for the [Wilmott Magazine](#)
- Author of forthcoming book [“Pragmatic AI and ML in Finance”](#)
- Teaches AI/ML and Fintech Related topics in the MS and MBA programs at [Northeastern University, Boston](#)
- Reviewer: Journal of Asset Management



QuantUniversity, LLC

www.quantuniversity.com

Thank you!

Contact

Sri Krishnamurthy, CFA, CAP
Founder and CEO
QuantUniversity LLC.

LinkedIn [srikrishnamurthy](#)

www.QuantUniversity.com



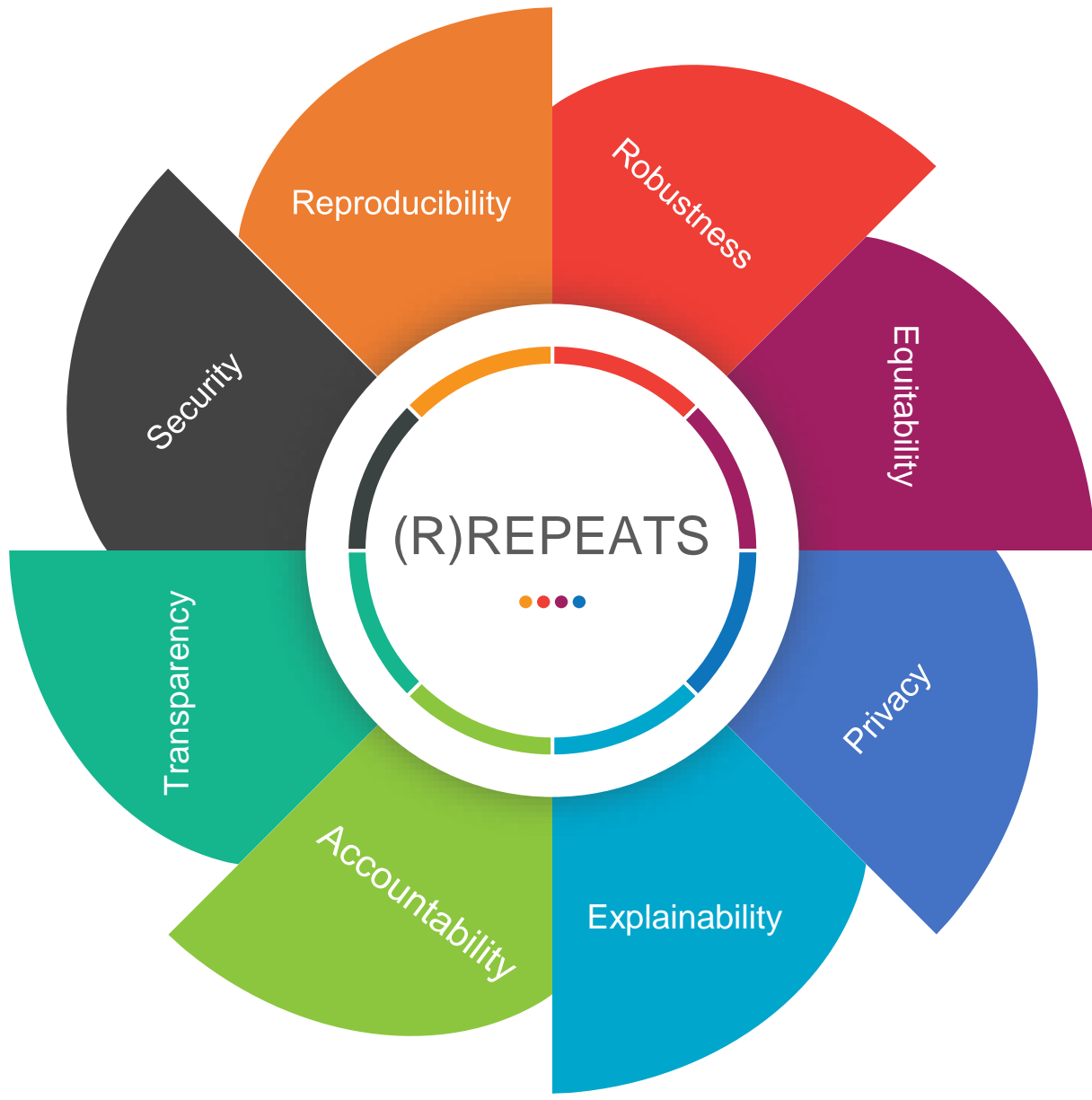
LFAI & Data

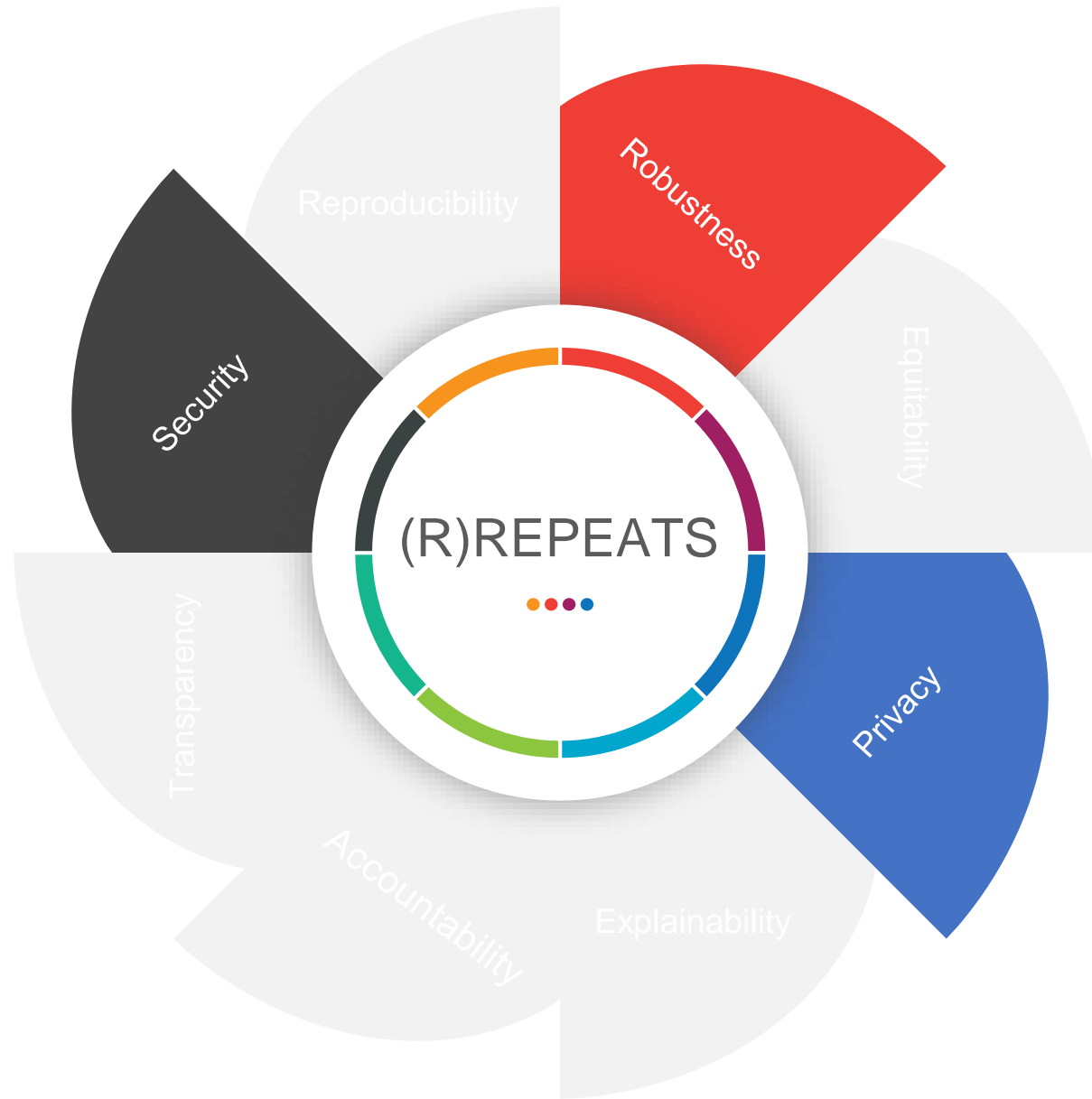
Trusted AI Principles – Tools and Techniques

Trusted AI Tools (AI Fairness, AI Explainability, Adversarial Robustness etc) and RREPEATS
Emerging DataOps activities at the LF-AI, Trusted AI and RREPEATS,
Animesh Singh, IBM



Animesh Singh,
Watson CTO, Open
Technology,
Distinguished Engineer,
IBM

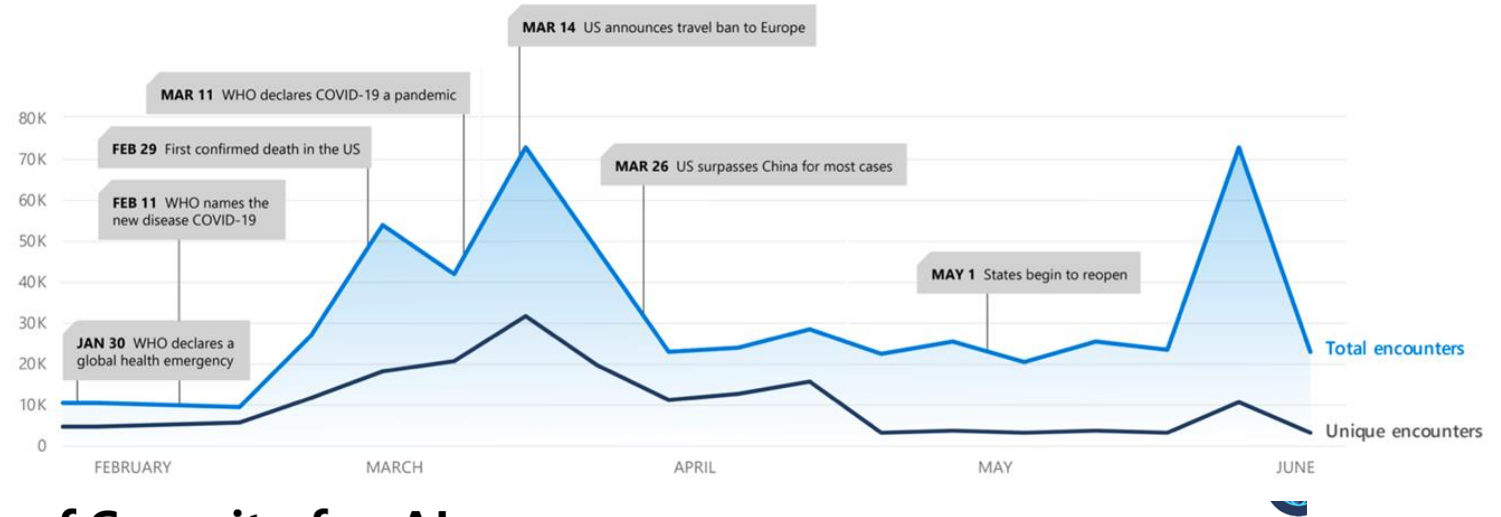




Trusted AI – Increased focus on AI Security

Cybercrime follows the issues of the day Malware encounters align with news headlines

COVID-themed attacks: United States



State of Security for AI



Awareness of risk is low



Low AI security understanding



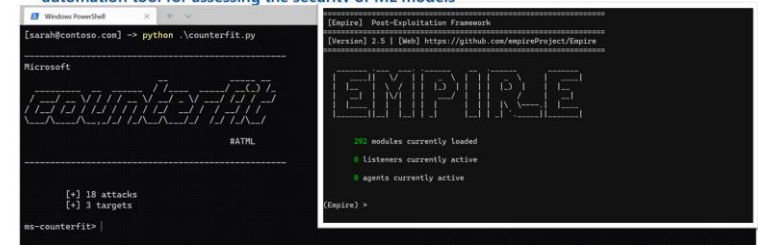
Security posture is close to zero

Gartner research

600+ executives say security/privacy is top blocker to using AI
“Machine Learning presents a new attack surface and increases security risks.... Application leaders must anticipate and prepare to mitigate risks of data corruption, model theft, and adversarial examples.”

Counterfit

automation tool for assessing the security of ML models





Impact of GDPR on AI

Some AI models will be legally classified as personal data

**PHILOSOPHICAL TRANSACTIONS
OF THE ROYAL SOCIETY A**
MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES

Open Access
Check for updates
View PDF

Tools Share

Cite this article

Section

Abstract

1. Introduction

2. European data protection law and machine learning

Research article

Algorithms that remember: model inversion attacks and data protection law

Michael Veale, Reuben Binns and Lilian Edwards

Published: 15 October 2018 | <https://doi.org/10.1098/rsta.2018.0083>

Abstract

Many individuals are concerned about the governance of machine learning systems and the prevention of algorithmic harms. The EU's recent General Data Protection Regulation (GDPR) has been seen as a core tool for achieving better governance of this area. While the GDPR does apply to the use of models in some limited situations, most of its provisions relate to the governance of personal data, while models have traditionally been seen as intellectual property. We present recent work from the information security literature around 'model inversion' and 'membership inference' attacks, which indicates that the process of turning training data into machine-learned systems is not one way,

<https://royalsocietypublishing.org/doi/full/10.1098/rsta.2018.0083>

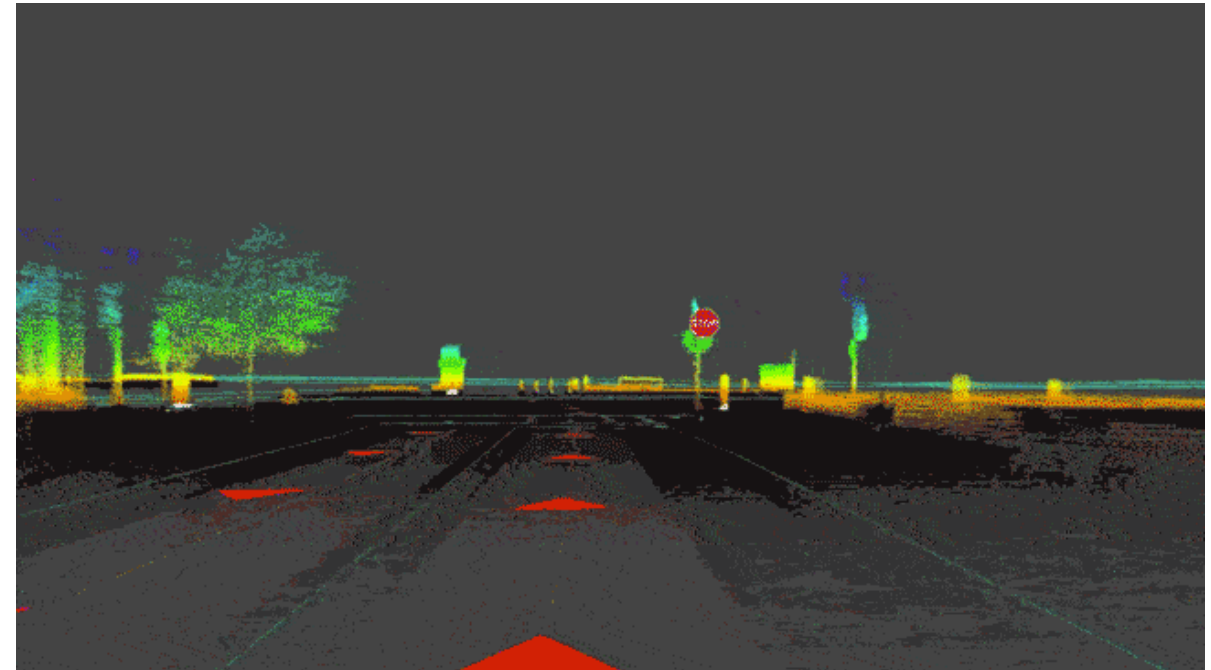
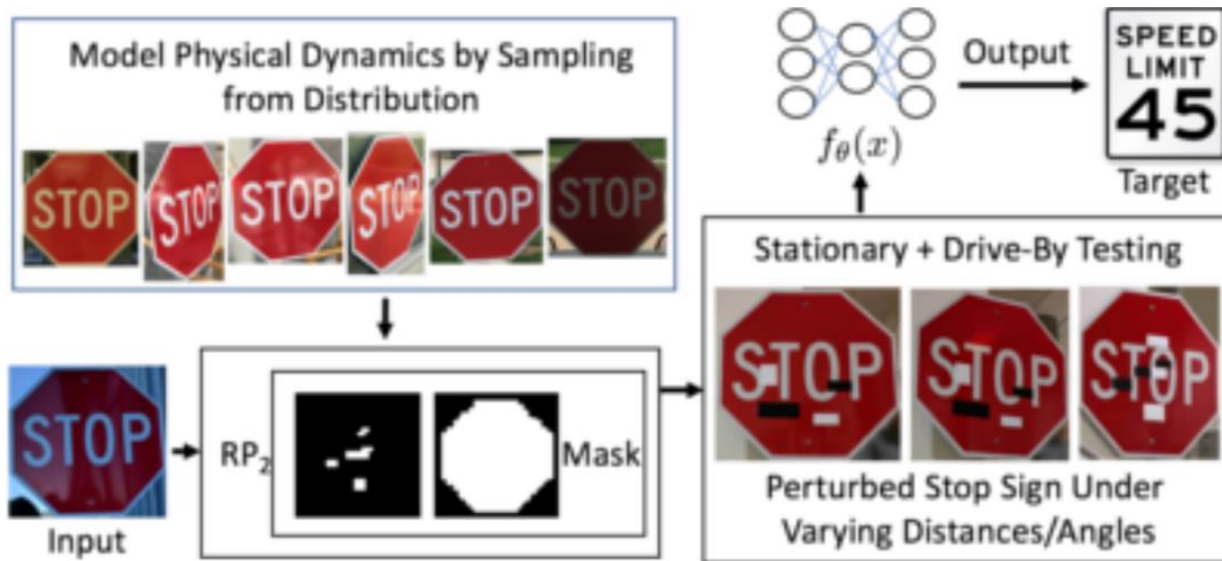
GDPR mostly relates to the governance of personal data.

This paper presents recent work from the information security literature around 'model inversion' and 'membership inference' attacks, indicating that the process of turning training data into machine-learned systems is not one way, and demonstrate how this could lead some models to be legally classified as personal data.

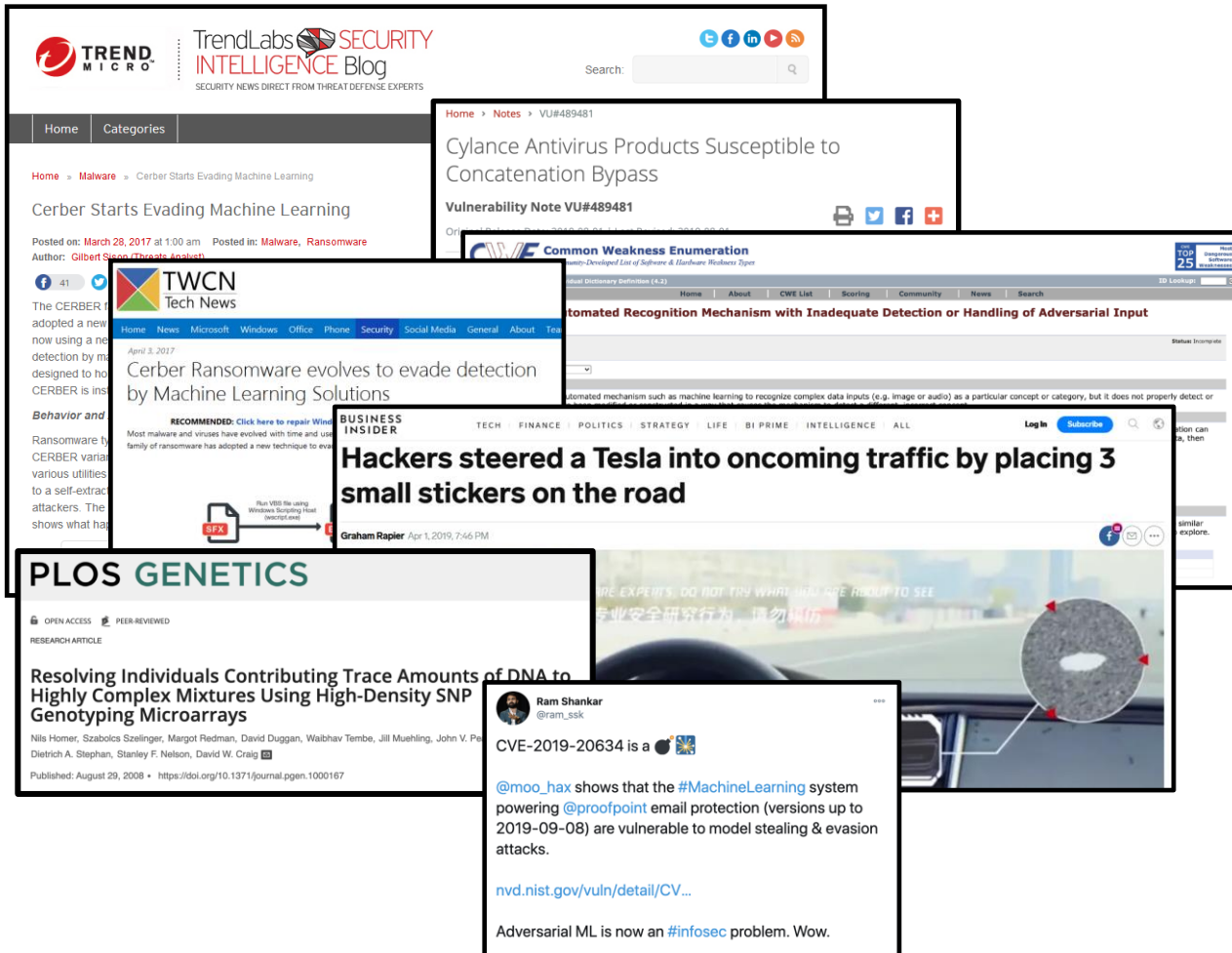
They also explore the different rights and obligations this would trigger.

Adversarial Threats to AI

<https://arxiv.org/pdf/1707.08945.pdf>



Real-world Adversarial Exploits

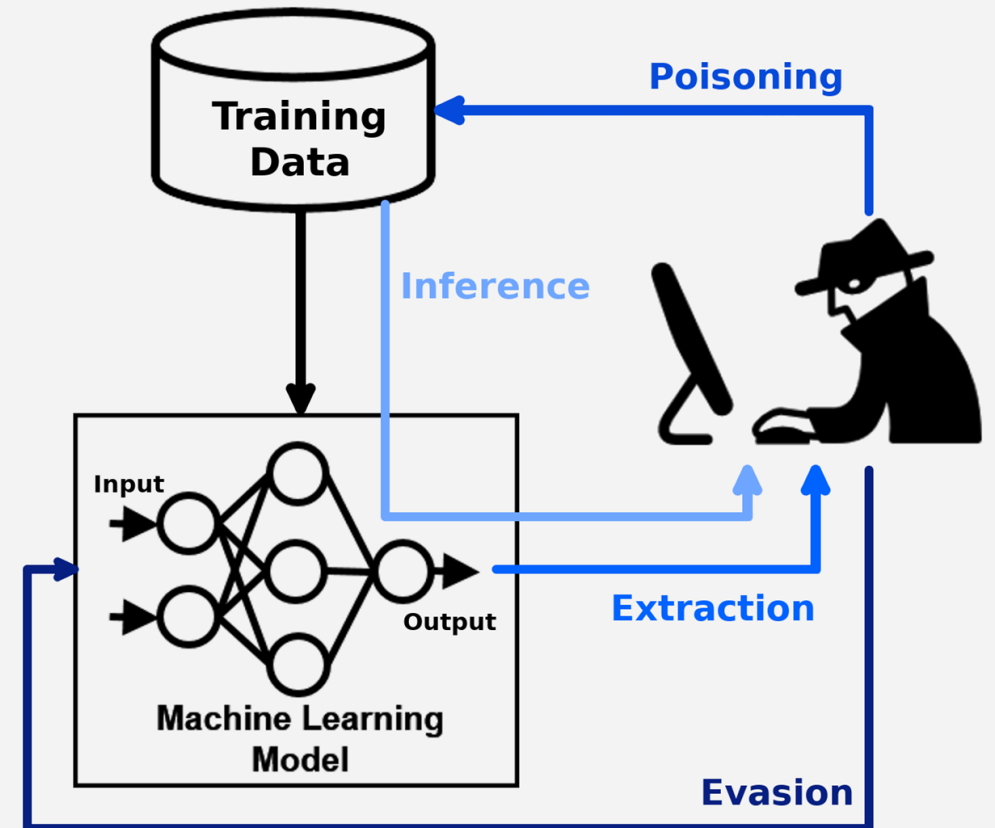


- Evasion of classification in antivirus products
 - Undetected ransomware installs and encrypts your computer
- Real-world adversarial patches for evasion attacks on cars
 - Losing control of autonomous vehicles leads to damages and injury
- Extraction of classification models to stage evasion attack against email protection system
 - Bypassing email security systems increases chances of phishing attacks
- Leaking sensitive private information
 - Revealing a person health condition via membership inference on health-related models

Adversarial Threats to Machine Learning

Adversarial threats against machine learning models and applications have a wide variety of attack vectors.

- **Evasion:** Modifying input to influence model
- **Poisoning:** Modify training data to add backdoor
- **Extraction:** Steal a proprietary model
- **Inference:** Learn information on private data



Adversarial Robustness Toolbox (ART)

ART is a Python library for machine learning security

 TensorFlow  Keras

 PYTORCH  mxnet

 scikit-learn GPy

 dmlc XGBoost

LightGBM

 CatBoost



Adversarial
Robustness
Toolbox

- github.com/Trusted-AI/adversarial-robustness-toolbox
- provide tools to developers and researcher
- Evaluating, Defending, Certifying and Verifying of machine learning models and applications
- **All Tasks:** Classification, Object Detection, Generation, Encoding, Certification, etc.
- **All Frameworks:** TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy
- **All Data:** images, tables, audio, video, etc.
- Contributions and feedback are very welcome!



ART Community – Contributors and Tools

ART Adopters and Contributors

- IBM
- Microsoft
- Troj AI
- Two Six Labs, LLC
- Kyushu University
- Intel Corporation
- University of Chicago
- The MITRE Corporation
- General Motors Company
- AGH University of Science and Technology
- Rensselaer Polytechnic Institute (RPI)
- IMT Atlantique



Adversarial
Robustness
Toolbox

2.5K GitHub Stars

150K Downloads

8K+ Commits

Armory

- Adversarial Robustness Evaluation Test Bed
- Run evaluations with ART locally or scaled in the cloud using Docker containers
- github.com/twosixlabs/armory

Counterfit

- Command line tool to simplify running evaluations with ART in terminals
- github.com/Azure/counterfit

ai-privacy-toolkit

- Tools for privacy and compliance of AI models
- End-to-end privacy evaluation and mitigation of privacy risks
- github.com/IBM/ai-privacy-toolkit



ART Adopter - DARPA

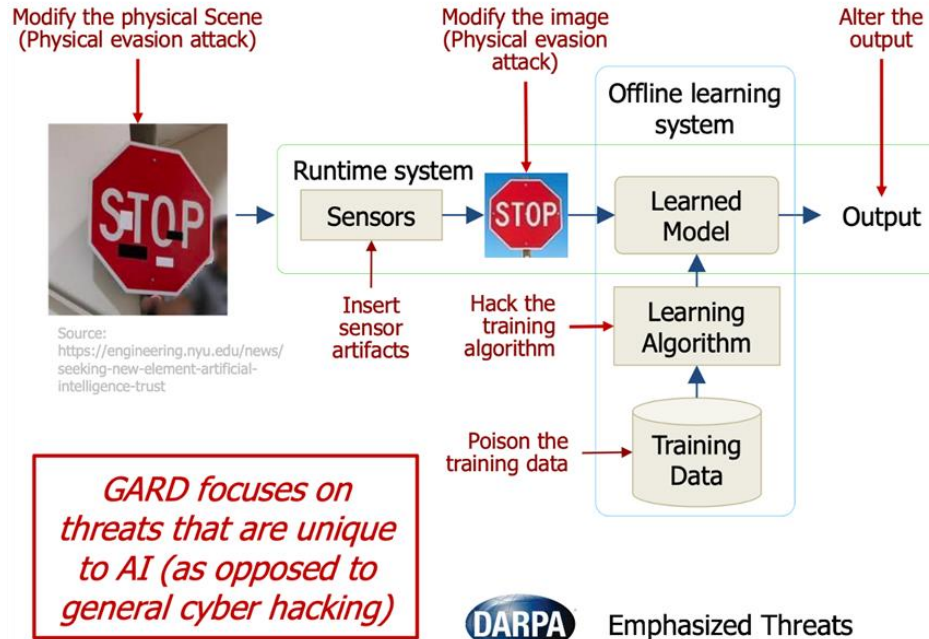
Contributors/Adopters of ART

- IBM
- Microsoft
- Troj AI
- Two Six Labs, LLC
- Kyushu University
- Intel Corporation
- University of Chicago
- The MITRE Corporation
- General Motors Company
- AGH University of Science and Technology
- Rensselaer Polytechnic Institute (RPI)
- IMT Atlantique



Guaranteeing AI Robustness against Deception (GARD)

Develop theoretical foundations, principled defense algorithms, and evaluation frameworks to enable machine learning systems to be robust against deception by an adversary.



Task: Protect AI-enabled systems in context

- Assuming cyber challenges are met
- Against realistic threat models
- In an end-to-end systems context

SoTA Challenges:

- Back-and-forth results with no fundamental advances
- Idealized but unrealistic threat models
- Ad hoc testing methods

Programmatic approach:

- Advance theory of robustness/vulnerability
- Create practical defenses for realistic threat models
 1. Physical evasion attacks (e.g. patch attacks)
 2. Poisoning attacks (at training time)
 3. Digital evasion attacks
- Improved evaluation of robustness
 - Armory testbed
 - Evaluation tutorials



Emphasized Threats

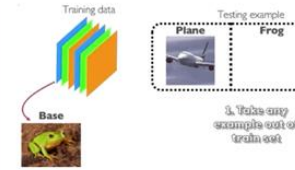
1. Physical Attacks



<https://arxiv.org/pdf/1707.08945.pdf> <https://arxiv.org/pdf/1801.00349.pdf>

- Bad actors have access to the physical world.
- Won't be detected during training or verification.

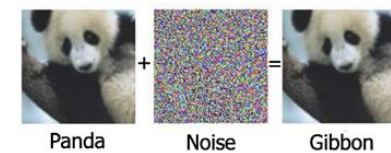
2. Poisoning Attacks



<https://wihuang.com/project/poison-frog/>

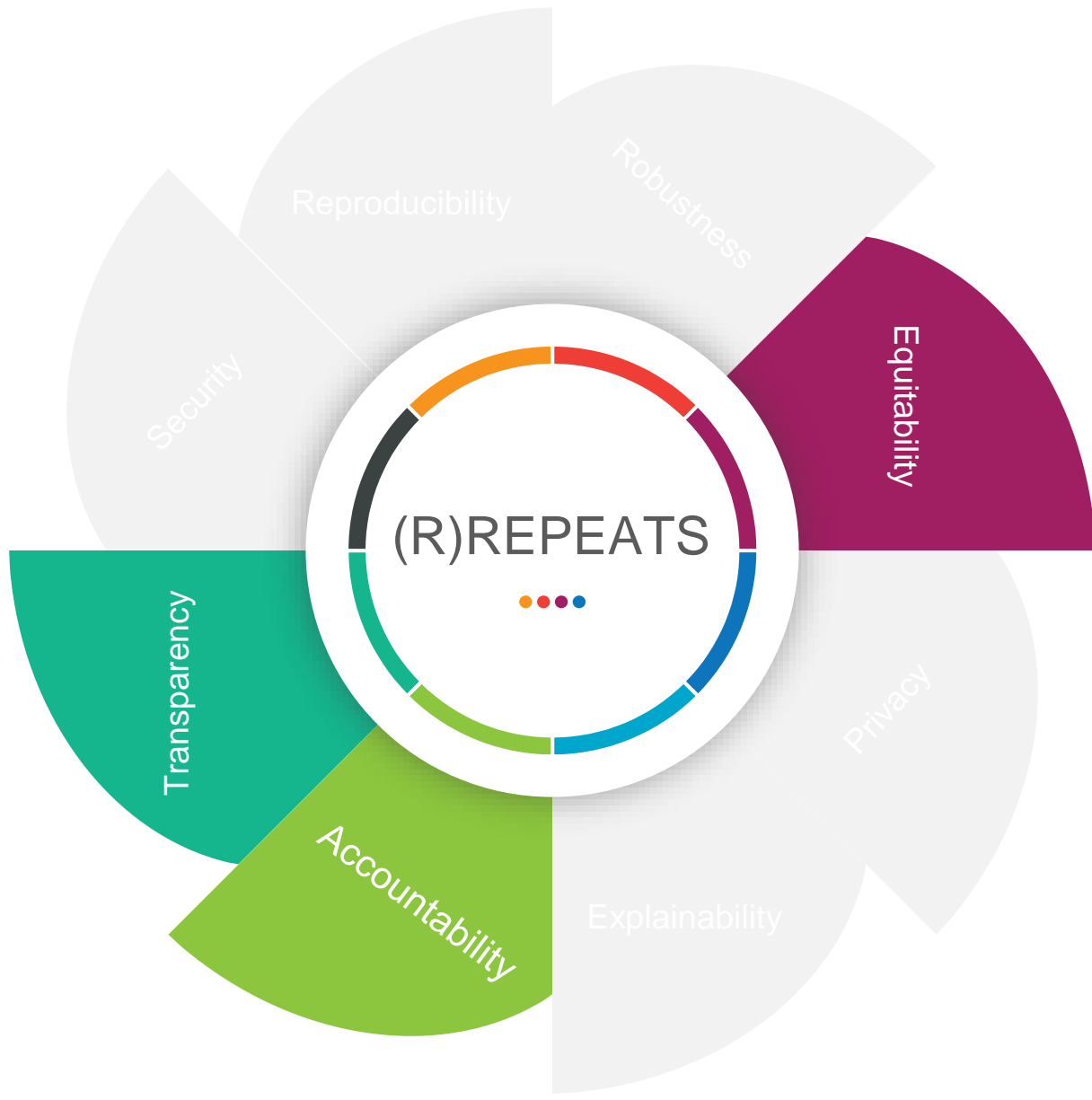
- Training data is expensive; public data is often used.
- Back-doors easy to exploit
- Clean label attacks hard to detect

3. Digital Attacks



<https://openai.com/blog/adversarial-example-research/>

- Hardest attacks to defend against
- Attacker requires access to digital signal
 - Insider attack



Bias in AI Example: Criminal Justice System

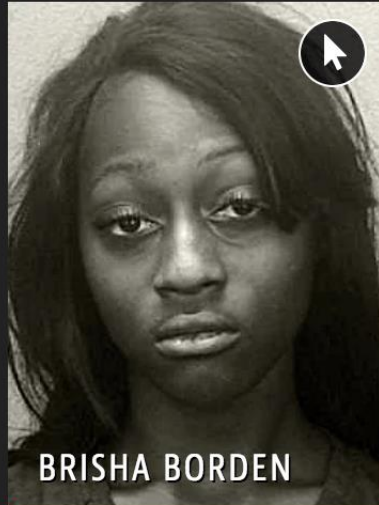
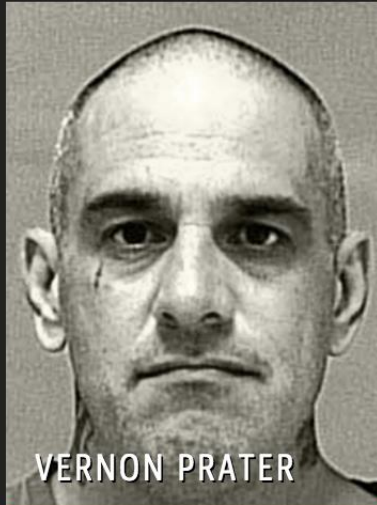
- Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe's COMPAS algorithm.
- Defendants with low risk scores are released on bail.
- It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.



Bias in AI Example: Criminal Justice System

- 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.

Two Petty Theft Arrests



VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

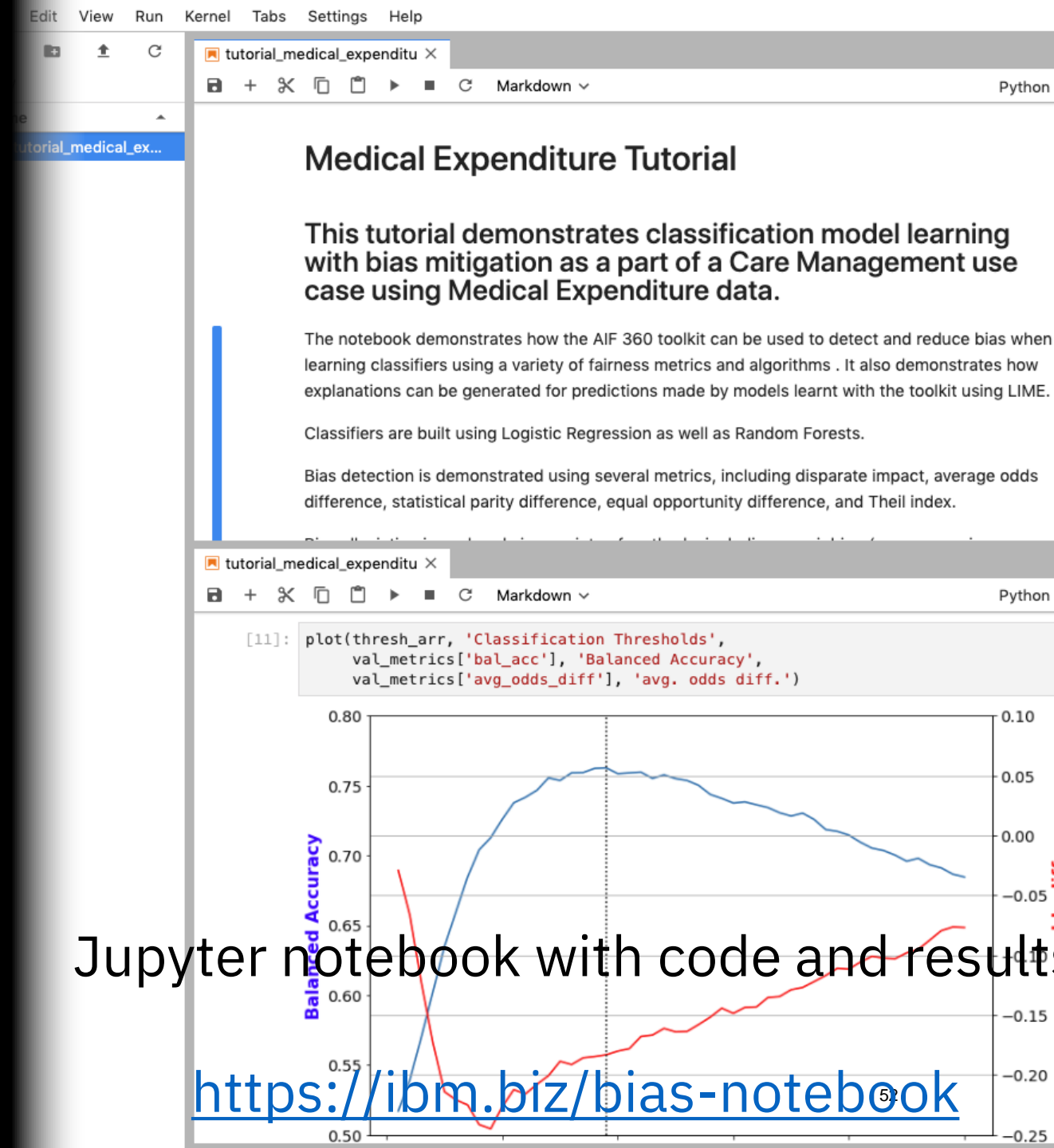
Two Petty Theft Arrests

VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Hypothetical Use Case: Racial Bias in Healthcare

- A healthcare utilization scoring model prioritizes cases for healthcare management.
- Problem: Supplemental care decisions can't be predicated on factors such as race of the patient.
- IBM Researchers used state-of-the-art techniques to measure and reduce racial bias.
- Result: An improved model that is **much more fair** relative to the original model learned from the original data.



Some conservative state about 1.2 m
politicians say Austin is head- has increa

Racial Bias Found in Hospital Algorithm

BY MELANIE EVANS
AND ANNA WILDE MATHEWS

Black patients were less likely than white patients to get extra medical help, despite being sicker, when an algorithm used by a large hospital chose who got the additional attention, according to a new study underscoring the risks as technology gains a foothold in medicine.

Hospitals use the algo-
rithm for

Related Real-World Case
(Wall Street Journal October 25,
2019 Page A3)

Black patients were less likely than white patients to get extra medical help, despite being sicker, when an algorithm used by a large hospital

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

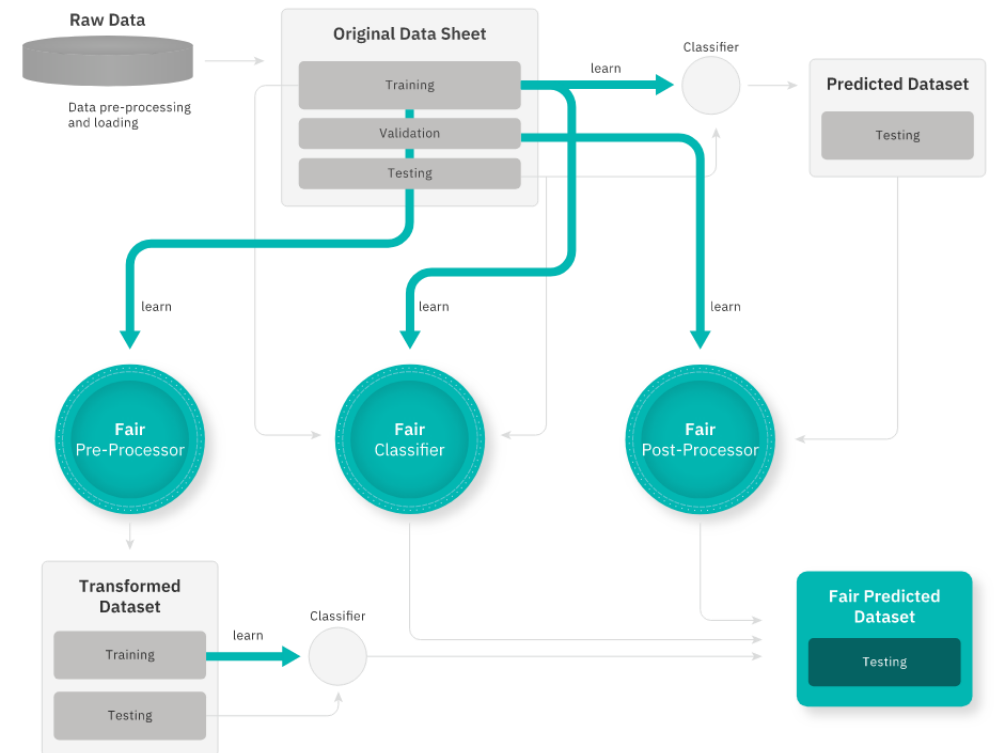
Fairness metrics (70+)

Fairness metric explanations

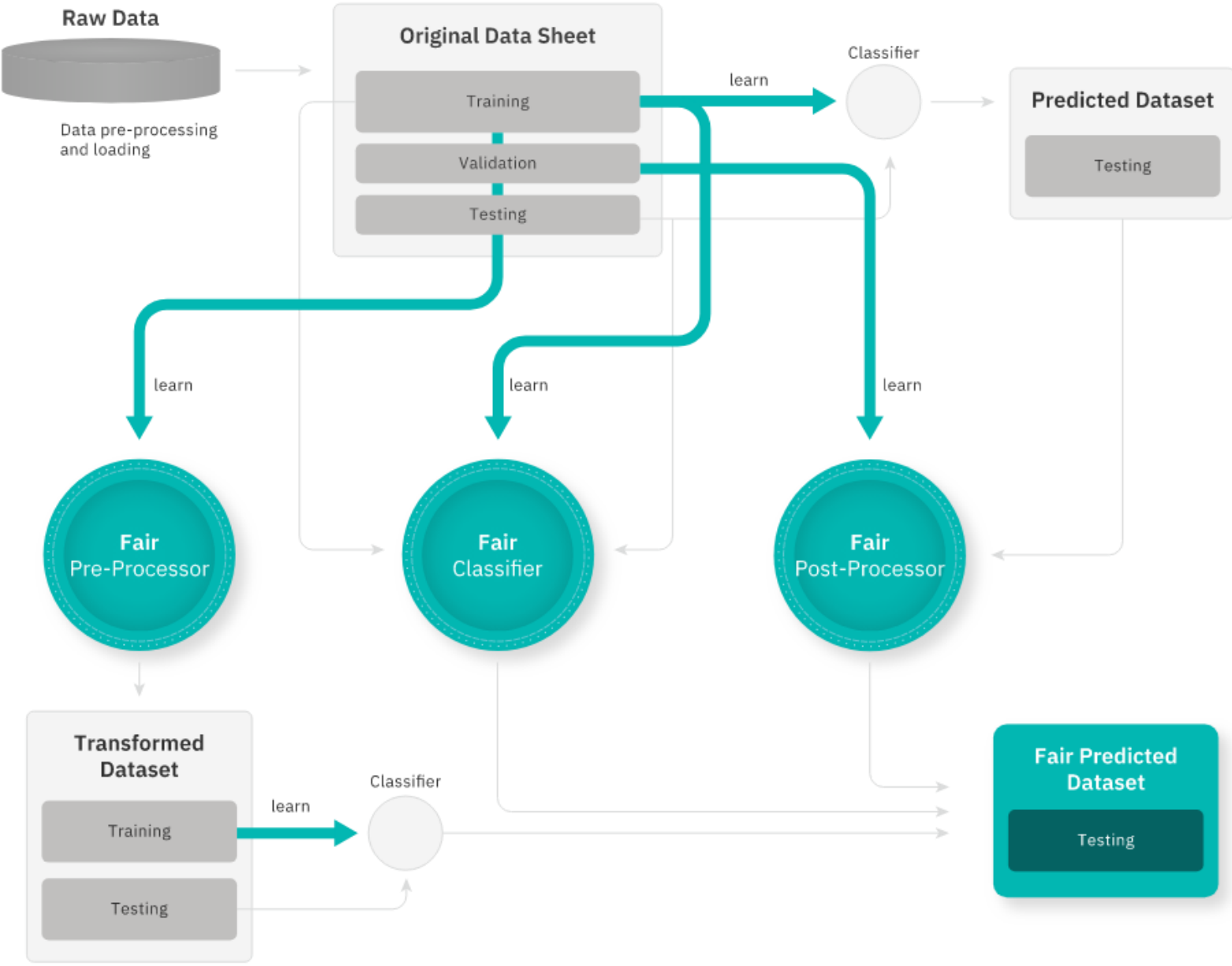
Bias mitigation algorithms (10+)

<http://aif360.mybluemix.net/>

• AIF360



AIF 360 detects for fairness in building and deploying models throughout AI Lifecycle



Metrics (70+)

Statistical Parity Difference

The difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.



Equal Opportunity Difference

The difference of true positive rates between the unprivileged and the privileged groups.



Average Odds Difference

The average difference of false positive rate (false positives/negatives) and true positive rate (true positives/positives) between unprivileged and privileged groups.



Disparate Impact

The ratio of rate of favorable outcome for the unprivileged group to that of the privileged group.



Theil Index

Measures the inequality in benefit allocation for individuals.



Euclidean Distance

The average Euclidean distance between the samples from the two datasets.



Mahalanobis Distance

The average Mahalanobis distance between the samples from the two datasets.



Manhattan Distance

The average Manhattan distance between the samples from the two datasets.



Algorithms (10)

Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.



Reweighting

Use to mitigate bias in training data. Modifies the weights of different training examples.



Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.



Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.



Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.



Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.



Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.



Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.



Equalized Odds Post-processing

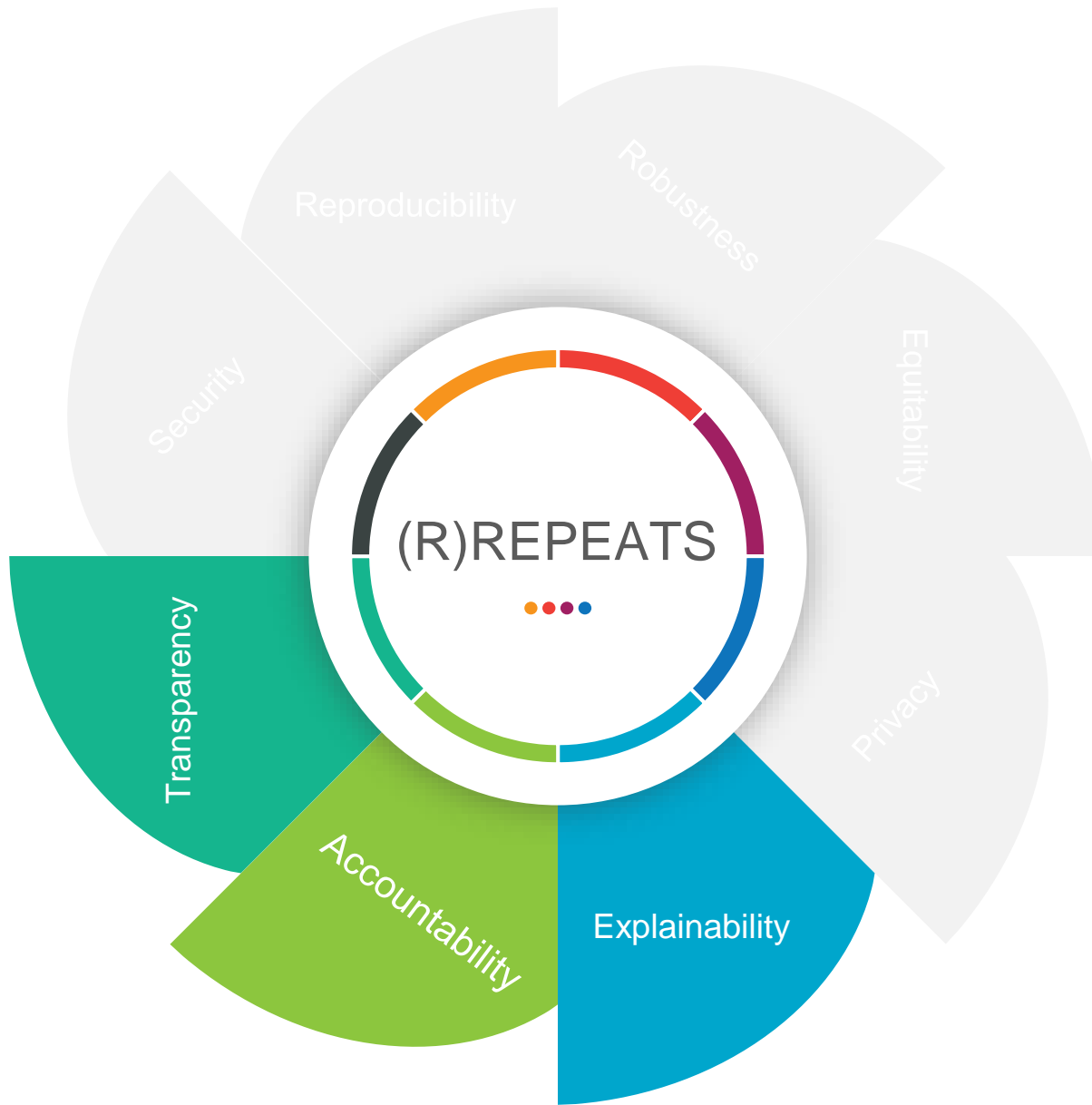
Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.



Meta Fair Classifier

Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.





AI needs to explain its decision, and there are different ways to explain

One explanation does not fit all

Different stakeholders require explanations for different purposes and with different objectives, and explanations will have to be tailored to their needs.

End users/customers (trust)

Doctors: *Why did you recommend this treatment?*

Customers: *Why was my loan denied?*

Teachers: *Why was my teaching evaluated in this way?*

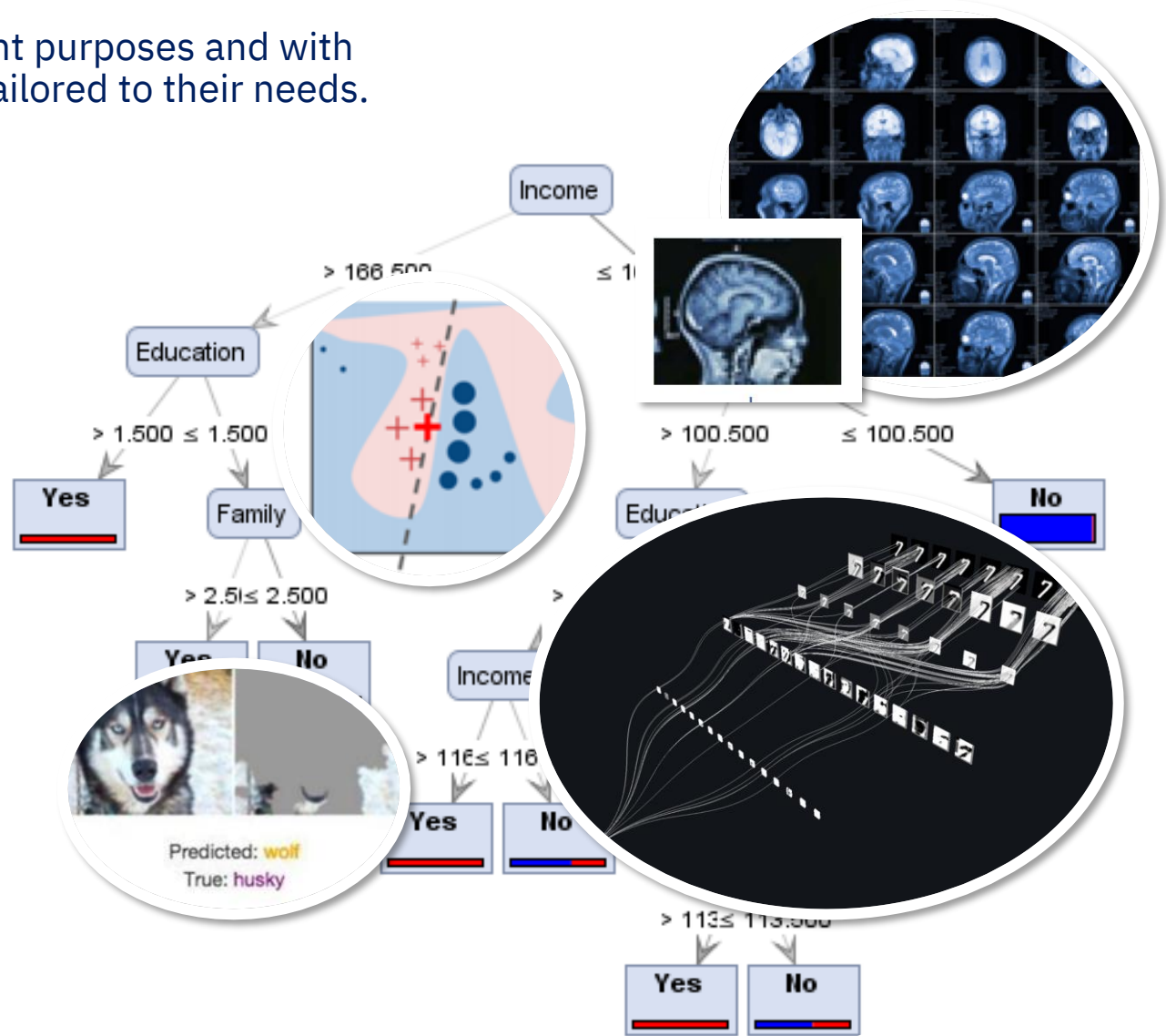
Gov't/regulators (compliance, safety)

Prove to me that you didn't discriminate.

Developers (quality, "debuggability")

Is our system performing well?

How can we improve it?



AI Explainability 360

↳ (AIX360)

<https://github.com/IBM/AIX360>

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

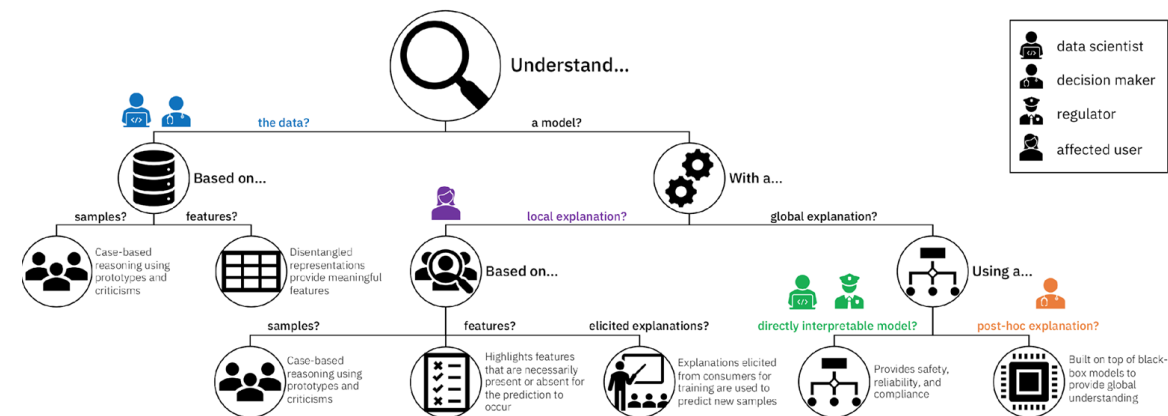
Local post-hoc




Global post-hoc

Directly interpretable

<http://aix360.mybluemix.net>

• AIX360



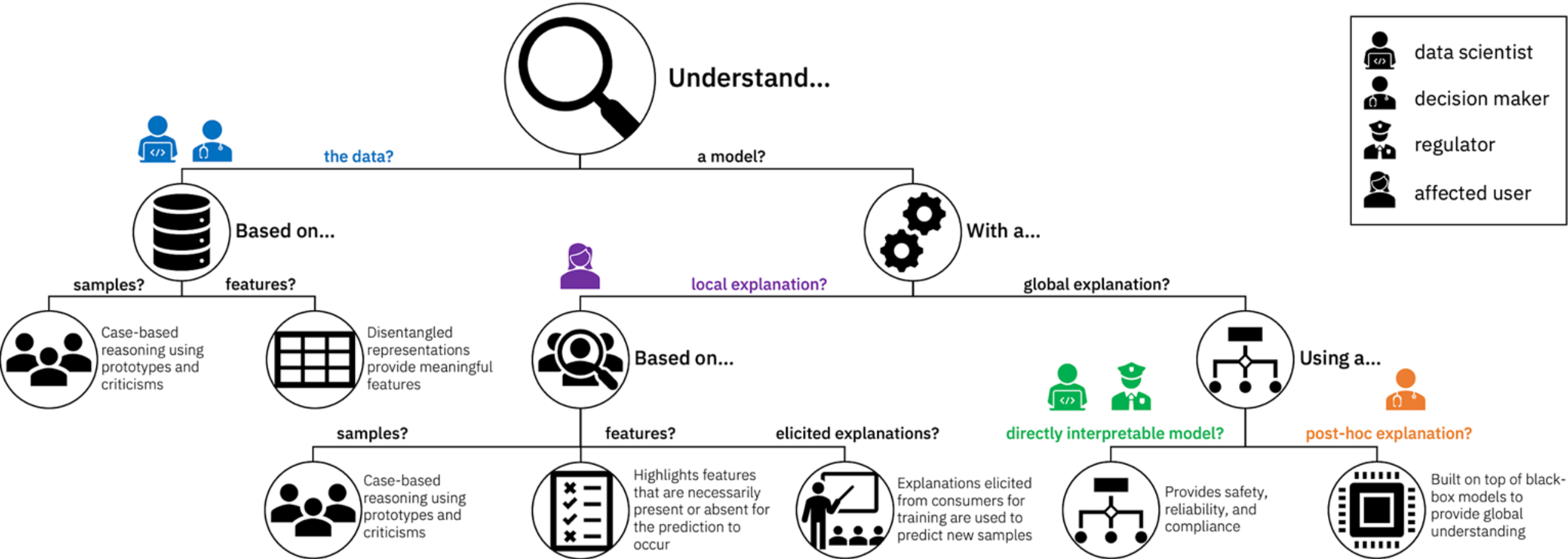
<p>Boolean Decision Rules via Column Generation (Light Edition)</p> <p>Directly learn accurate and interpretable 'or'-of-'and' logical classification rules.</p> <p>→</p>	<p>Generalized Linear Rule Models</p> <p>Directly learn accurate and interpretable weighted combinations of 'and' rules for classification or regression.</p> <p>→</p>	<p>ProfWeight</p> <p>Improve the accuracy of a directly interpretable model such as a decision tree using the confidence profile of a neural network.</p> <p>→</p>	<p>Teaching AI to Explain its Decisions</p> <p>Predict both labels and explanations with a model whose training set contains features, labels, and explanations.</p> <p>→</p>	<p>Contrastive Explanations Method</p> <p>Generate justifications for neural network classifications by highlighting minimally sufficient features, and minimally and critically absent features.</p> <p>→</p>	<p>Contrastive Explanations Method with Monotonic Attribute Functions</p> <p>Contrastive explanations for colored images or images with rich structure.</p> <p>→</p>
<p>Disentangled Inferred Prior VAE</p> <p>Learn disentangled representations for interpreting unlabeled data.</p> <p>→</p>	<p>ProtoDash</p> <p>Select prototypical examples from a dataset.</p> <p>→</p>	<p>AI Explainability 360 - Demo</p> <p> <input checked="" type="radio"/> Data <input type="radio"/> Consumer <input type="radio"/> Explanation </p> <p>Choose a consumer type</p> <ul style="list-style-type: none"> <input type="radio"/>  Data Scientist must ensure the model works appropriately before deployment <input checked="" type="radio"/>  Loan Officer needs to assess the model's prediction and make the final judgement <input type="radio"/>  Bank Customer wants to understand the reason for the application result 			

AI Explainability 360

The most comprehensive **open source** toolkit for explaining ML models and data:

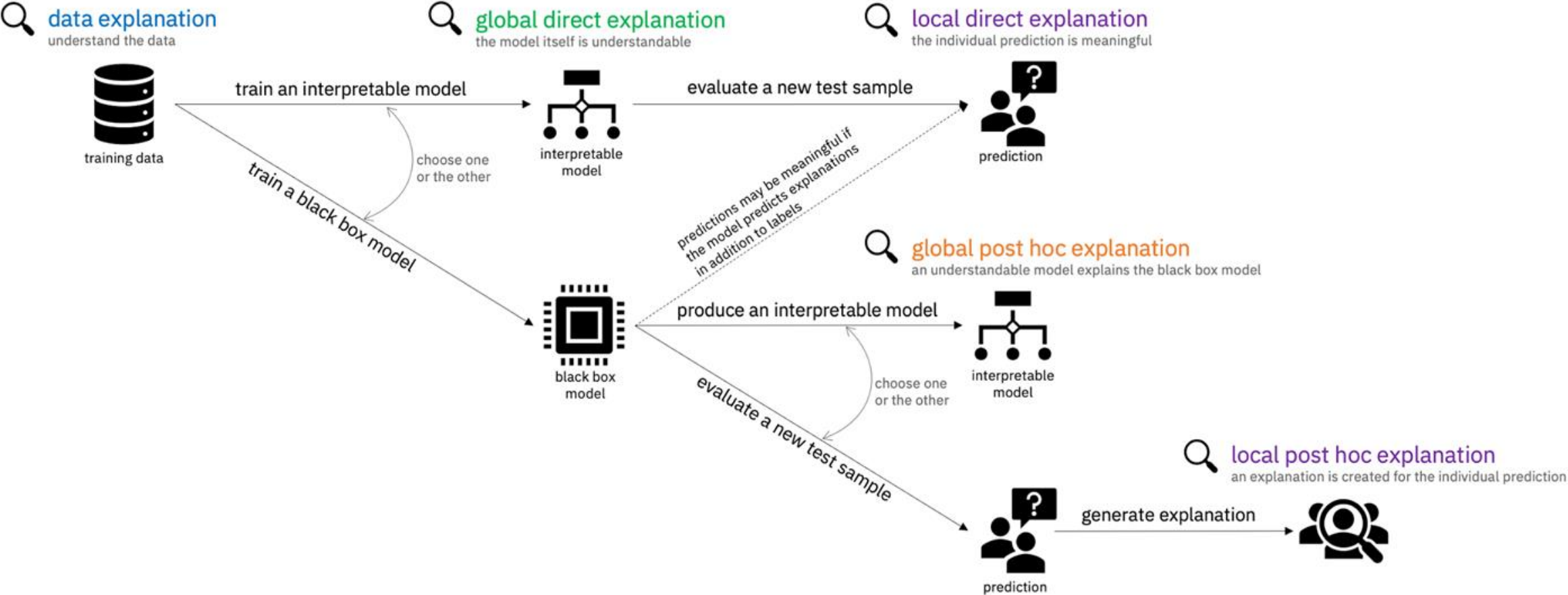
- 8 innovated algorithms from IBM Research + 2 from scientific community
- 13 tutorial notebooks covering use cases in finance, healthcare, lifestyle, retention, etc.

AIX360: Multiple dimensions of explainability



	data scientist
	decision maker
	regulator
	affected user

AIX360: Multiple dimensions of explainability



LFAI Trusted AI Projects Update



AI Fairness 360



AI Explainability 360



Adversarial Robustness Toolbox

<https://ai-fairness-360.org/>

<https://ai-explainability-360.org/>

<https://adversarial-robustness-toolbox.org/>

ART	Very strong growth and adoption.	V1.6.2 released.	2.3K GitHub Stars, 150K Downloads, 8K+ Commits
AIF360:	Strong growth trajectory	V0.4.0 released	1.4K GitHub Stars, 330+Commits
AIX360	Used extensively within enterprises	V0.2.1 released	~850 GitHub Stars, 10K downloads/month, 240+Commits

LFAI & Data

Trusted AI and MLOps

LFAI & DATA



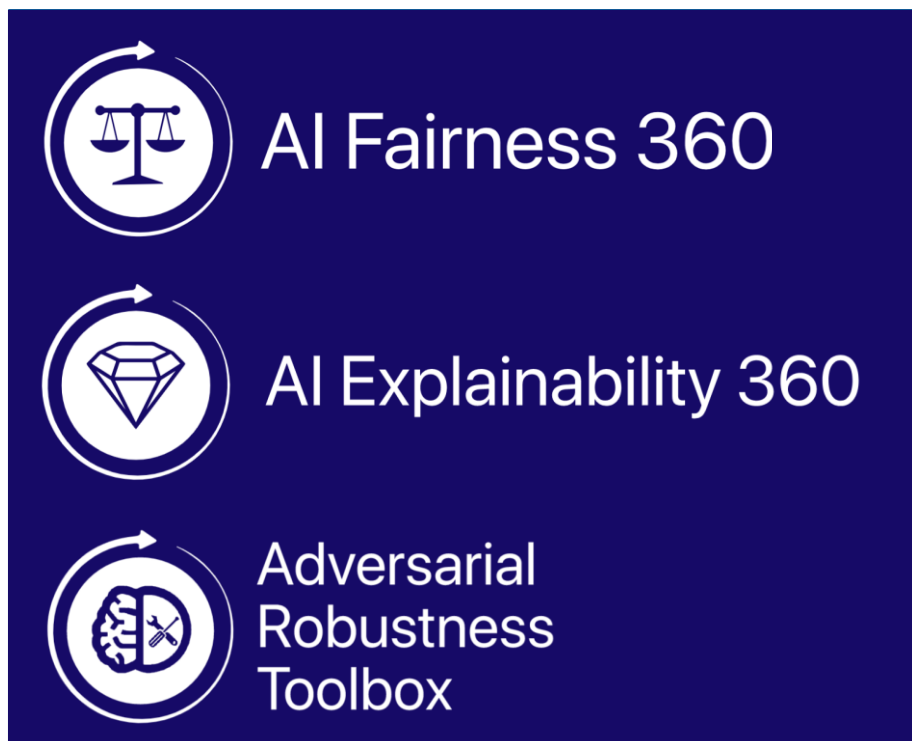
Animesh Singh,
Watson CTO, Open
Technology,
Distinguished Engineer,
IBM

Machine Learning eXchange (MLX): Data and AI Assets Catalog and Execution Engine

The screenshot displays the MLX web interface. On the left, there is a dark navigation sidebar with icons and labels for 'MLX', 'Datasets', 'Models', 'Pipelines', 'Components', 'Notebooks', and 'KFServices'. The main content area is divided into four sections: 'Pipelines', 'Datasets', 'Notebooks', and 'Models'. Each section has a title, a subtitle, and two buttons: 'VIEW ALL [CATEGORY]' and 'REGISTER A [CATEGORY]'. Below the 'Models' section, there is a grid of eight model cards. Each card includes a title, a brief description, a category label, and an orange arrow icon.

Model Name	Description	Category
MAX Human Pose Estimator	IBM Model Asset eXchange(MAX) model that detects humans in an image and estimate the pose for each person.	Human Pose Estimation
MAX Image Caption Generator	IBM Model Asset eXchange(MAX) model that generates captions from a fixed vocabulary describing the contents of images in the COCO dataset.	Image-To-Text Translation
MAX Image Resolution Enhancer	IBM Model Asset eXchange(MAX) model that upscales an image by a factor of 4, while generating photo-realistic details.	Super-Resolution
MAX Object Detector	IBM Model Asset eXchange(MAX) model that localizes and identifies multiple objects in a single image.	Object detection in images
MAX Optical Character Recognition	IBM Model Asset eXchange(MAX) model that identifies text in an image.	Optical Character Recognition
MAX Question Answering	IBM Model Asset eXchange(MAX) model that answers questions on a given corpus of text.	Question and Answer
MAX Recommender System	IBM Model Asset eXchange(MAX) model that generates personalized recommendations.	Recommendations
MAX Text Sentiment Classifier	IBM Model Asset eXchange(MAX) model that detects the sentiment captured in short pieces of text.	Sentiment Analysis

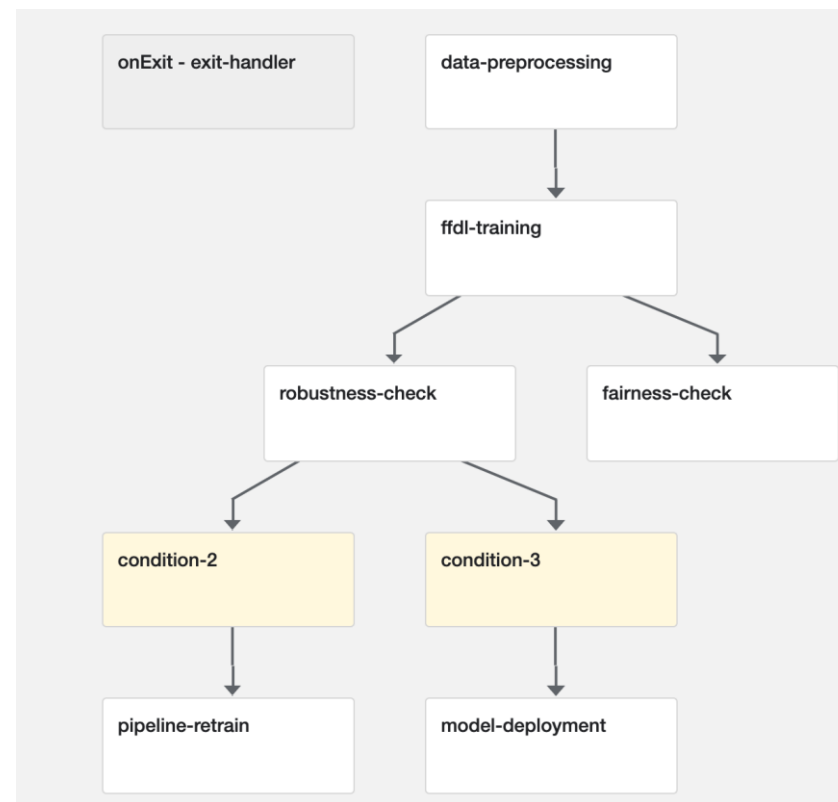
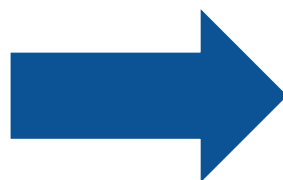
Trusted AI Pipelines available in MLX



<https://ai-fairness-360.org/>

<https://ai-explainability-360.org/>

<https://adversarial-robustness-toolbox.org/>



Trusted AI Pipeline

Machine Learning eXchange

AI & DATA

- Home
- Datasets
- Models
- Pipelines**
- Components
- Notebooks
- Join the Conversation
- Settings

Pipelines

Pipelines for your machine learning workloads.

[VIEW EXPERIMENTS](#) [VIEW ALL PIPELINES](#) [GITHUB](#) [REGISTER A PIPELINE](#) Search

Katib Early Stopping Experiment

An example to launch Katib Experiment with early stopping.

OpenSource

Caching Pipeline

Example pipeline of how to enable/disable caching for individual task.

OpenSource

Exit Handler Pipeline

Downloads a message and prints it.

OpenSource

Flip Coin Example

A conditional pipeline to flip coins based on a random number generator.

OpenSource

Calculation Pipeline

A pipeline that performs arithmetic calculations and displays artifacts.

OpenSource

Nested Pipeline

Download and Get Most Frequent Word in a Sentence.

OpenSource

Trusted AI Pipeline

An example pipeline for trusted AI integration.

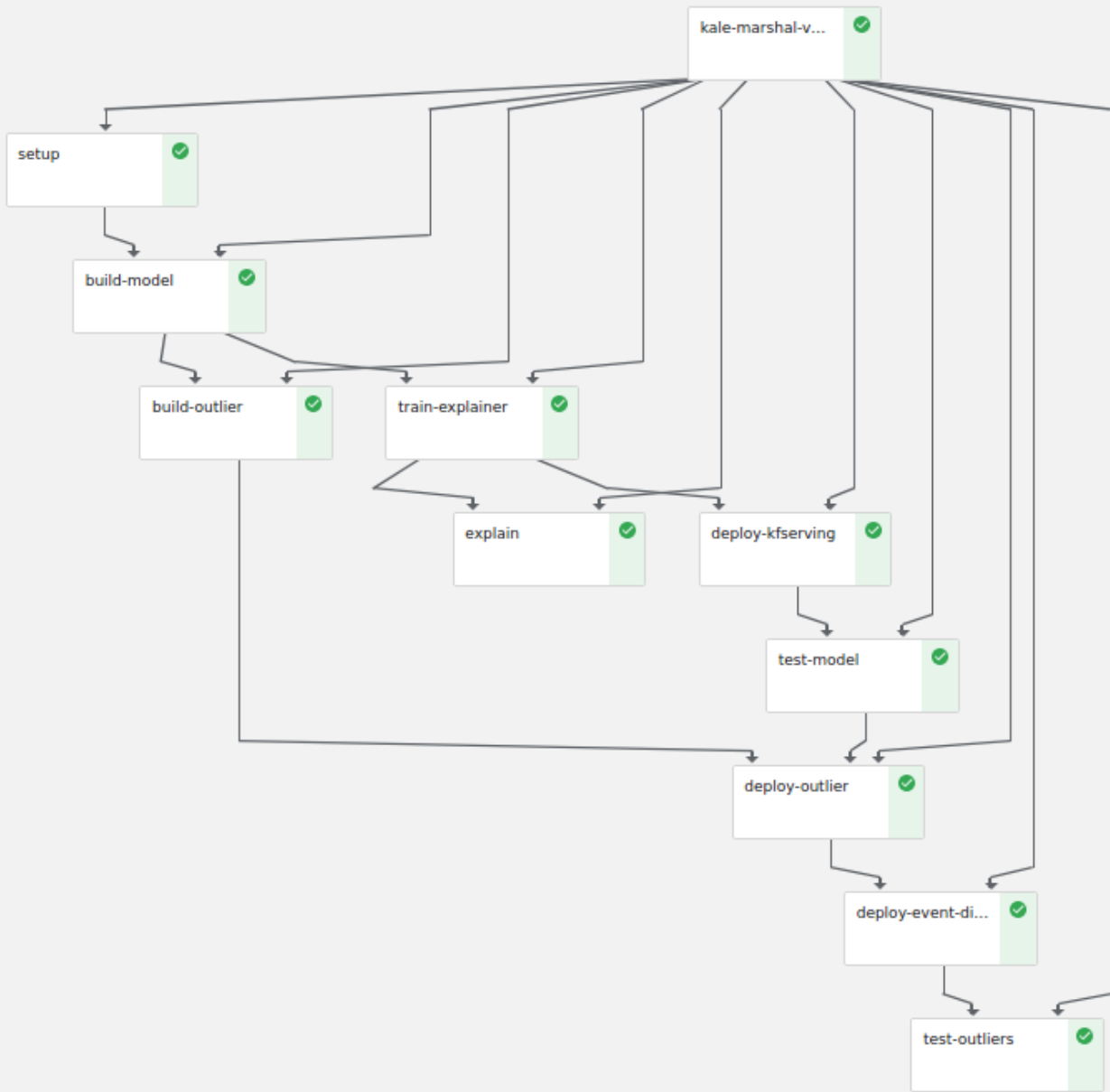
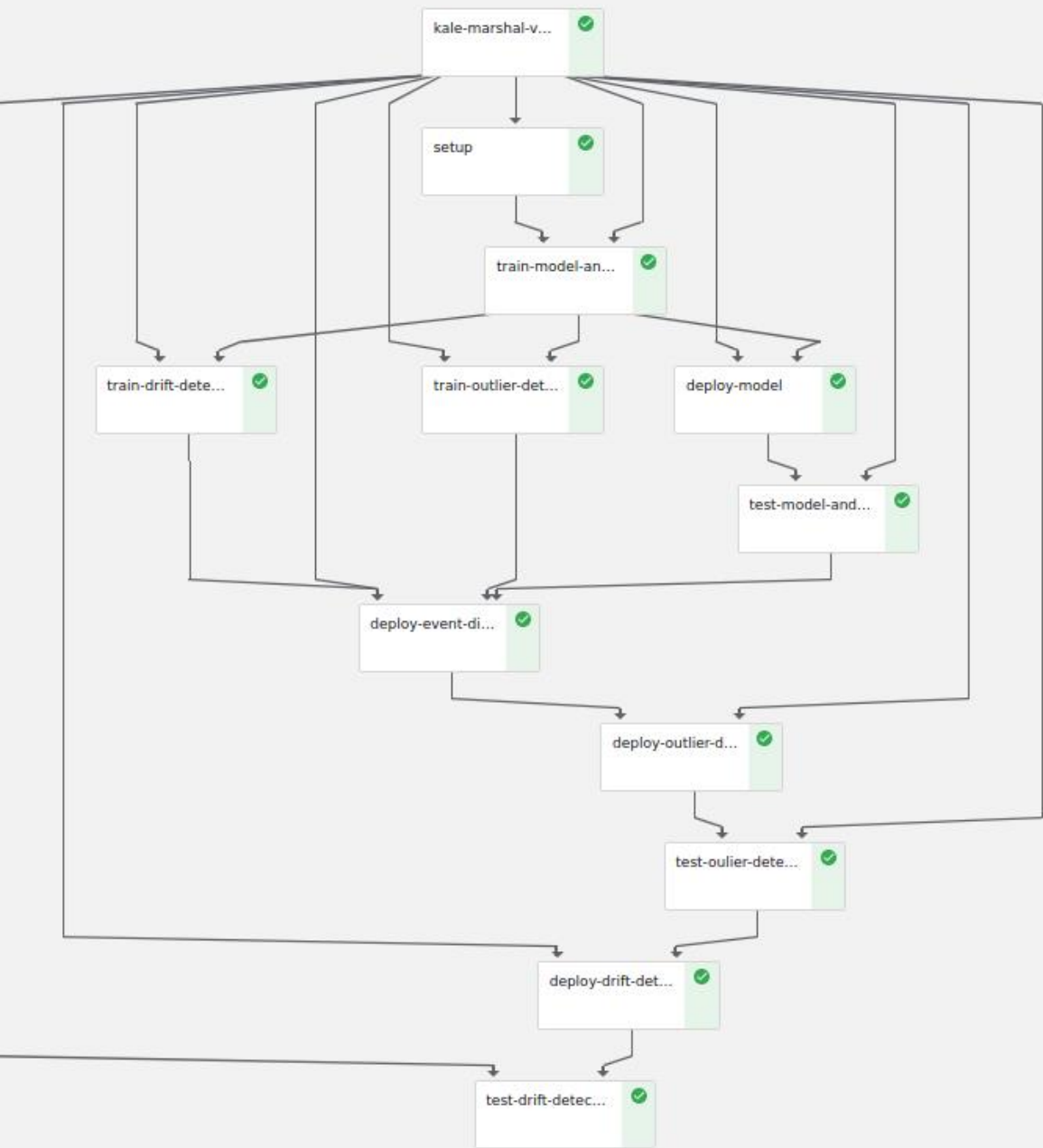
OpenSource

Watson Machine Learning Pipeline

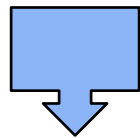
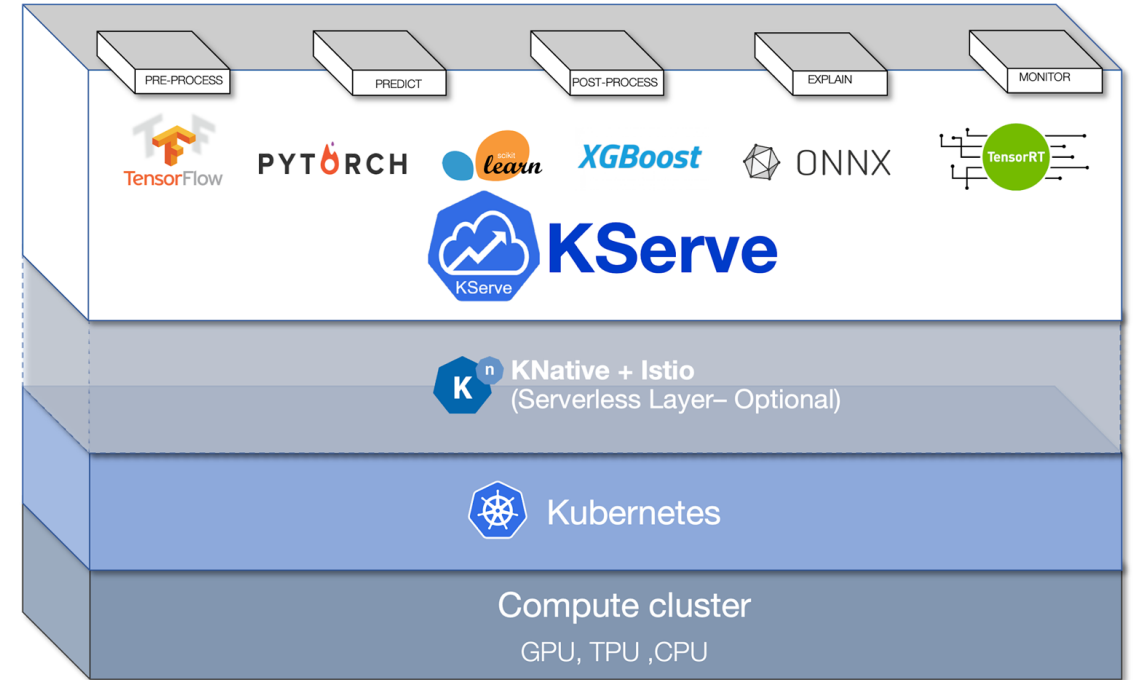
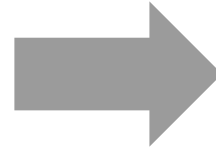
Kubeflow pipelines running on WML performing tensorflow image recognition.

OpenSource

Copyright © 2021 Machine Learning eXchange The Linux Foundation®. All rights reserved. The Linux Foundation has registered trademarks and uses trademarks. For a list of trademarks of The Linux Foundation, please see our [Trademark Usage](#) page. Linux is a registered trademark of Linus Torvalds. [Privacy Policy](#) and [Terms of Use](#).



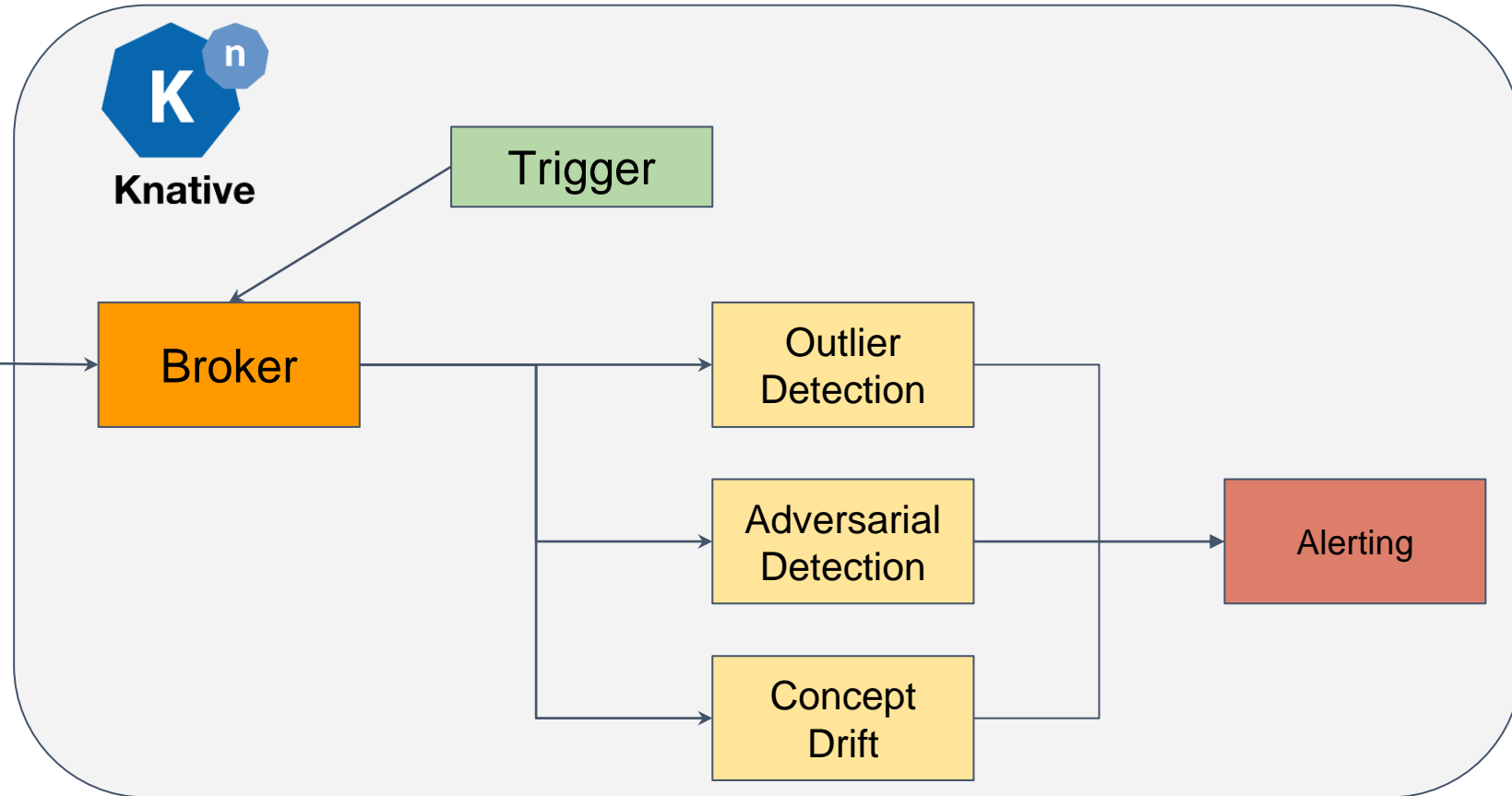
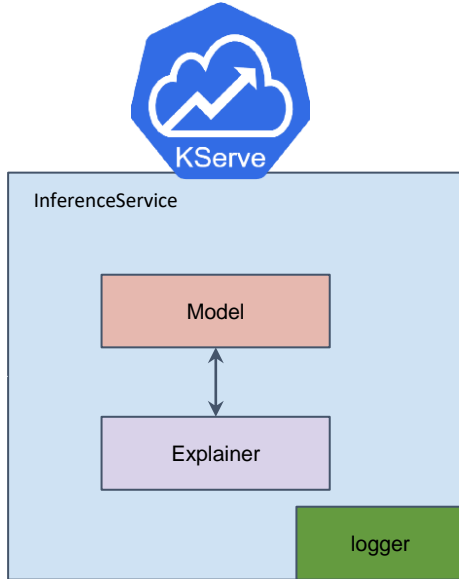
Trusted AI with KServe



QLF AI

<https://ai-fairness-360.org/>
<https://ai-explainability-360.org/>
<https://adversarial-robustness-toolbox.org/>

And more advanced Metrics



```
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>

LFAI & Data

The Trusted AI Principles - Tools and Techniques

Education Opportunities

 LFAI & DATA

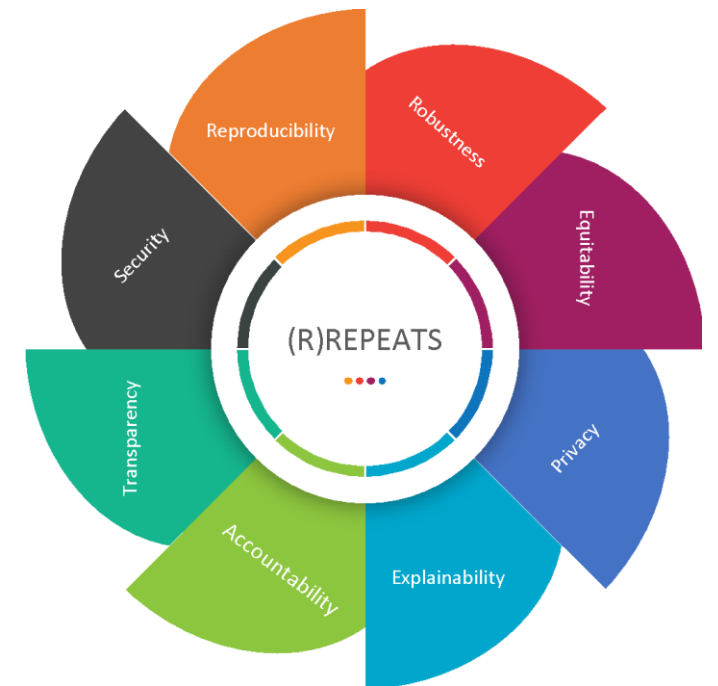


Haluk Demirkan
(University of Washington,
Tacoma)

Education Opportunities

Principles Working Group Team @ LF AI & Data is in a process of developing courses to educate our community

- Currently, LF AI & Data is offering a course “Ethics in AI and Data Science” (<https://www.edx.org/course/ethics-in-ai-and-data-science>) through EdX
- In addition, Principles Working Group Team is developing a series of courses to educate our community about Trusted AI Principles. Please let us know if you would like to join forces to educate our community about importance of Trusted AI.
- Join the mailing list [here](#) and participate in an upcoming meeting! Learn more [here](#)



LFAI & Data

The Trusted AI Principles - Tools and Techniques

Summary

LFAI & DATA



François Jézéquel, Director
of Business Development,
Orange Fab

Thank you

Stay connected with the Trusted AI Committee by joining the mailing list [here](#) and participate in an upcoming meeting!

Learn more [here](#)

References and Resources

- [Trusted AI Committee - Principles Working Group](#) (where you will find the slides and materials)
- [LF-AI] The Trusted-AI Principles document bit.ly/lfai-trustedai-principles
- [LF-AI Blog] [LF AI & Data Announces Principles for Trusted AI](#)
- [LF-AI Webinar] [RREPEATS – An Introduction to the Principles for Trusted AI – Thoughts and Next Steps](#)
- [LF-AI Webinar] [RREPEATS Practical Examples Review](#)
- [ACM] ACM Principles for Algorithmic Transparency and Accountability https://www.acm.org/binaries/content/assets/publicpolicy/2017_usacm_statement_algorithms.pdf
- [EU] Ethics Guidelines for Trustworthy AI - High-Level Expert Group on Artificial Intelligence set up by the European Commission <https://ec.europa.eu/futurium/en/ai-alliance-consultation>
- [EUFeb2020] On Artificial Intelligence -A European approach to excellence and trust https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
- [IEEE] Ethically Aligned Design, IEEE <https://ethicsinaction.ieee.org/>
- [DoD] AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF
- [OECD] Organisation for Economic Co-operation and Development <https://www.oecd.org/going-digital/ai/principles/>
- [SoA] State of the Art: Reproducibility in Artificial Intelligence Odd Erik Gundersen, Sigbjørn Kjensmo, Department of Computer Science Norwegian University of Science and Technology https://www.researchgate.net/publication/326450530_State_of_the_Art_Reproducibility_in_Artificial_Intelligence

Contributions

Principles Working Group Team:

- Souad Ouali (Orange)
- Jeff Cao (Tencent)
- Haluk Demirkan (University of Washington, Tacoma)
- Francois Jezequel (Orange)
- Sri Kishnamurthym (Quant University)
- Layla Li (KOSA)
- Sarah Luger (Orange)
- Susan Malaika (IBM)
- Alka Roy (The Responsible Innovation Project/ex-AT&T)
- Alejandro Saucedo (The Institute for Ethical AI / Seldon)
- Animesh Singh (IBM)
- Marta Ziosi (AI for People)