Responsible AI

Brillio POV

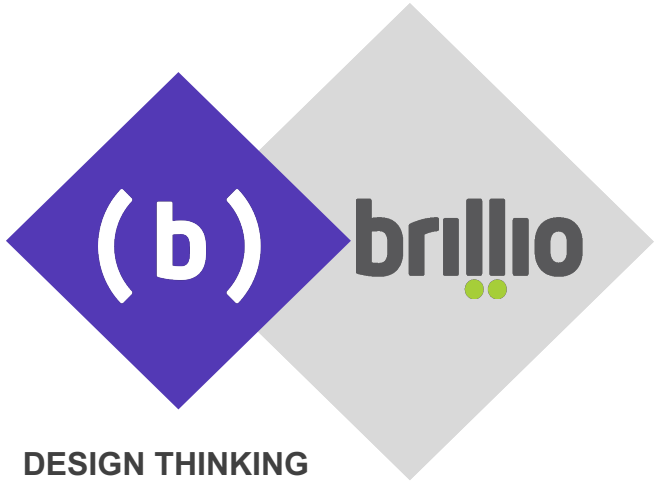
Jan 2023

AGENDA

- Our Digital Transformation Enablers
- Introduction to Responsible AI
- Solution Approach
- Tools Overview
- Engagement Models

Our Digital Transformation ENABLERS

DRIVEN BY PRODUCT MINDSET



DESIGN THINKING

CONTENT COLLABORATION

DESIGN STUDIO

PRODUCT ENGINEERING

- OMNI CHANNEL APPS
- MICROSERVICES/MESH ARCHITECTURE
- MODERN APPS & CONTAINERIZATION
- DEVOPS
- LOW/NO CODE SOLUTION
- COGNITIVE TESTING

CUSTOMER EXP PLATFORMS

- CRM IMPLEMENTATION
- MARKETING/SERVICE CLOUD
- SERVICE BOT
- HYBRID INTEGRATION
- INTELLIGENT SALES & E-COMMERCE

DATA & ANALYTICS

- MASTER DATA MANAGEMENT
- DATA MIGRATION
- DATA LAKE ON CLOUD
- AI/ML
- ANALYTICS AS A SERVICE

DIGITAL INFRASTRUCTURE

- CLOUD STRATEGY & MIGRATION
- DIGITAL OPERATIONS
- ROBOTIC PROCESS AUTOMATION
- MANAGED SERVICES
- ZERO OPS
- SECURITY & COMPLIANCE

ADVANCED TECHNOLOGY GROUP

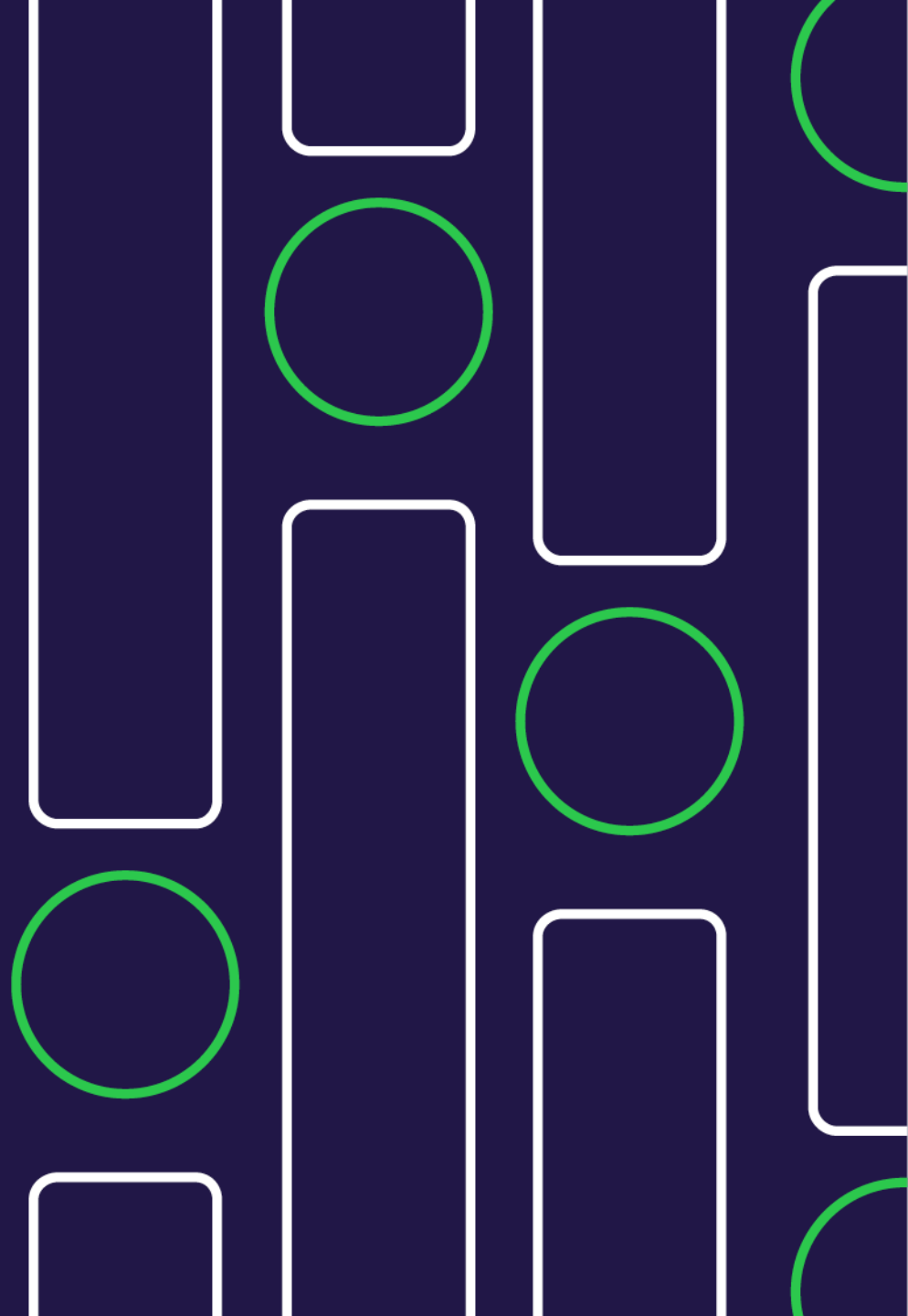
TECH STRATEGY & CONSULTING | TECH LABS | ENTERPRISE ARCHITECTURE | BLOCKCHAIN | EDGE | SERVERLESS COMPUTING

Microsoft apigee salesforce SAP paloalto NETWORKS Digital Asset snowflake

Accelerators: brillio BOLT brillio SMARTEST brillio DEVOPS TOOLKIT brillio one brillio CLIP

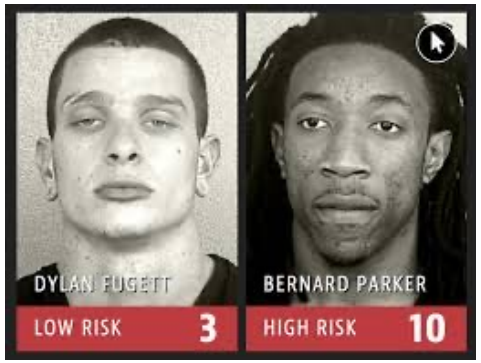
brillio

INTRODUCTION TO RESPONSIBLE AI



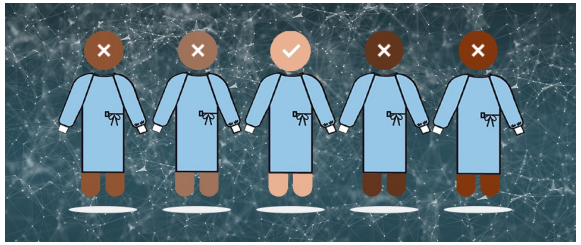
brillio

“The growing use of artificial intelligence in sensitive areas, including hiring, criminal justice, and healthcare, has stirred a debate about bias and fairness”



ProPublica reported that COMPAS is racially biased. According to the analysis, the system predicts that black defendants pose a higher risk of recidivism than they do, and the reverse for white defendants.

In 2015, Amazon realized that their algorithm used for hiring employees was found to be biased against women.



In October 2019, researchers found that an algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care heavily favored white patients over black patients.

brillio

While many of the businesses realize the importance of Responsible AI, they are still at stage 1-2 of maturity

Ignorant

- Unaware about “Responsible AI”
- “Responsible AI” sounds intriguing
  yet seems enigmatic

Exploration

- Aware about “Responsible AI” & its impact on business
- Unsure about the root-cause of introduction of biasness in AI Models
- Heard about some AI Fairness tools(ex- LIME/SHAP) but uncertain about how to use.

Operational

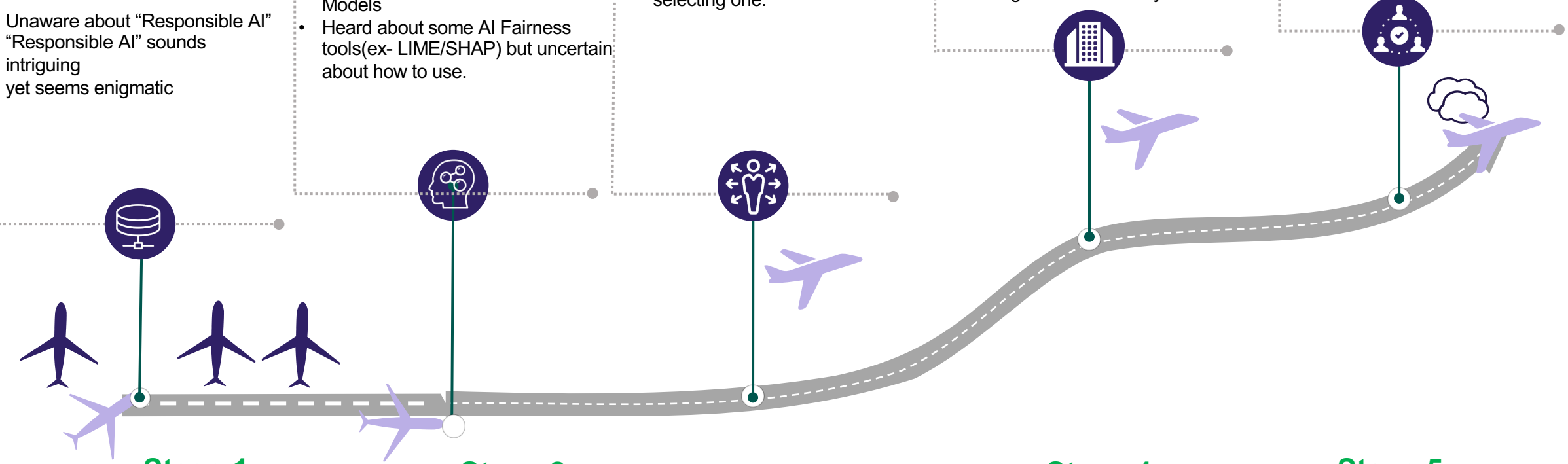
- Aware about “Responsible AI” & its impact on business
- Intermediate understanding of Biasness in AI Models
- Cognizance of AI Fairness tools(XAI) but indecisive in selecting one.

Systemic

- Familiar with “Responsible AI” & its impact on business
- Acquainted with the root-causes of biasness & their remedies in AI Models
- Capability for building fair AI models is at nascent stage
- Unsure about building Governance capability for managing AI Fairness throughout the AI Lifecycle

Transformational

- Well informed about “Responsible AI” & its impact on business
- Possess exhaustive understanding of Data Bias, Algorithm bias & XAI
- Proficient in building trust in AI models (Cloud/Open-Source Tools)
- Dedicated team building Governance capability for managing AI Fairness throughout the lifecycle of AI Models



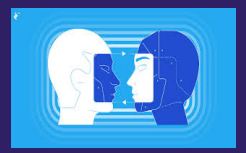
Responsible AI – An Overview

PEOPLE

ACCOUNTABILITY & TRANSPARENCY

TECHNOLOGY

BIAS & FAIRNESS



INTERPRETABILITY & EXPLAINABILITY

ETHICS & REGULATIONS

PROCESS

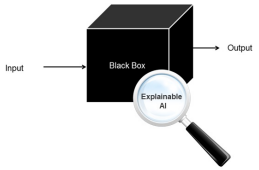
GOVERNANCE

brillio

Responsible AI - Principles

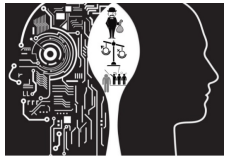
AI FAIRNESS

Justness



Transparency

ETHICAL AI



Governance

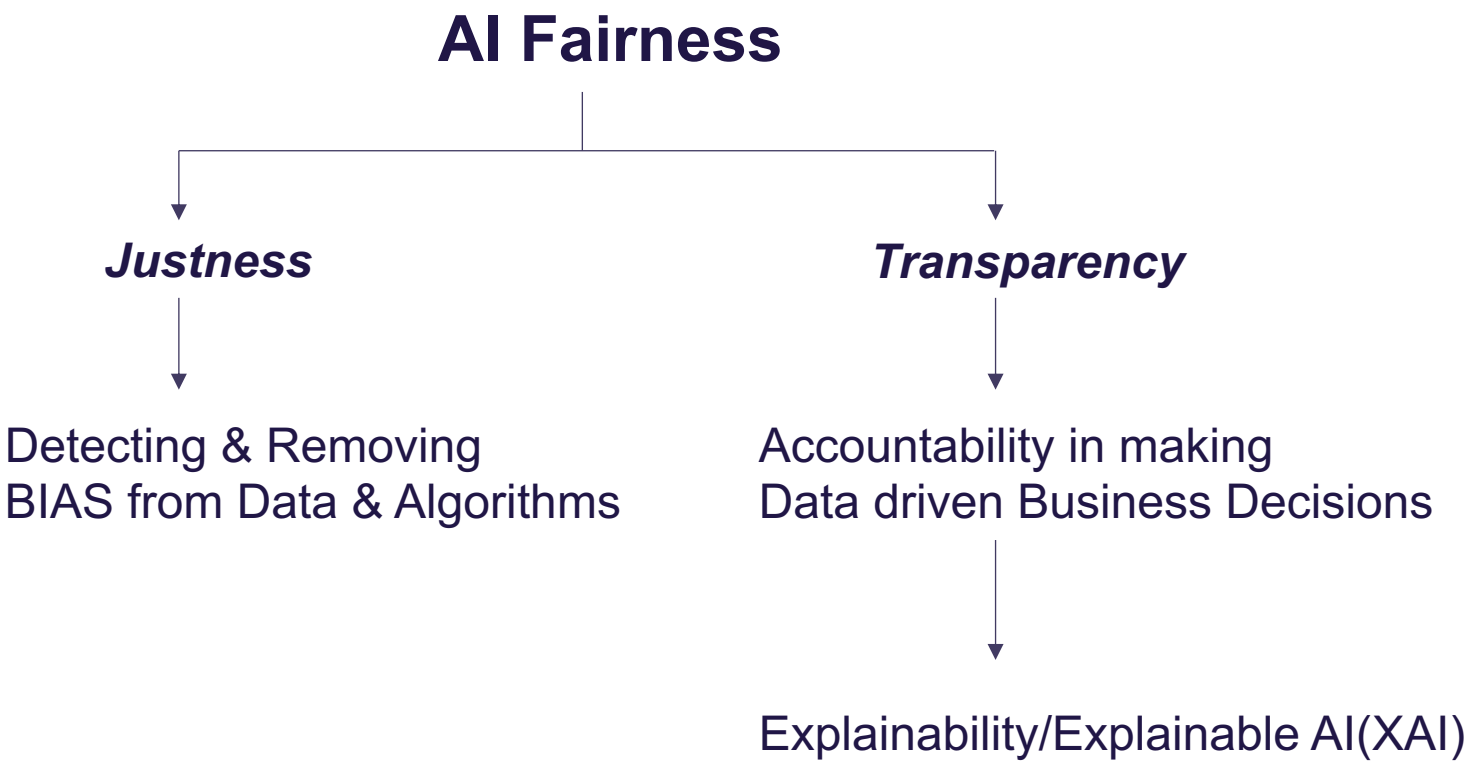
Privacy

brillio

Solution Approach

brillio

AI Fairness - Overview



A formal definition:

" Explainable AI is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models. With it, you can debug and improve model performance, and help others understand your models' behaviour "

AI/ML PROCESS PIPELINE

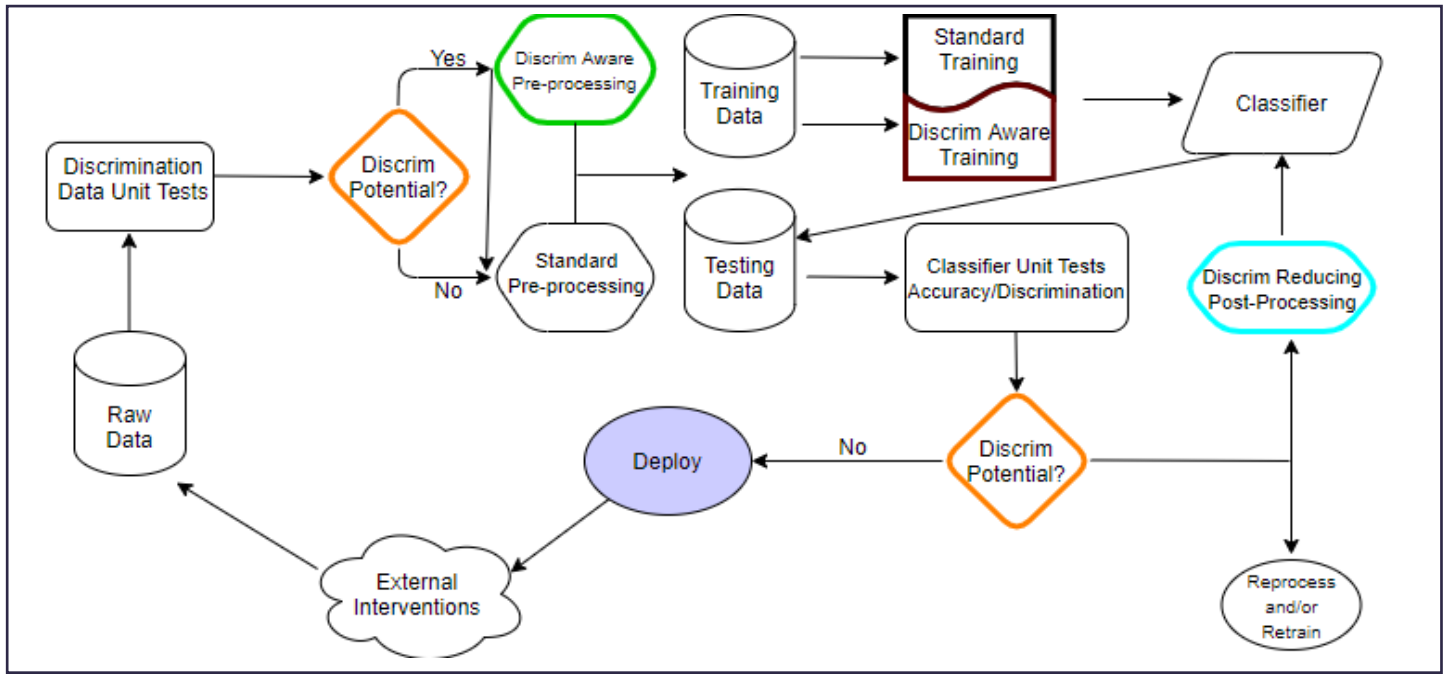
1 Problem Formulation & Data Discovery

2 Data Acquisition

3 Model Build

4 Test, Tune & Validate

5 Deployment



AI Bias – An Overview

"AI bias is an anomaly in the output of machine learning algorithms. These could be due to the prejudiced assumptions made during the algorithm development process or prejudices in the training data"

What causes AI Bias?

Bias in Data

UNBIASED DATASET → Biased Agent → BIASED DATASET

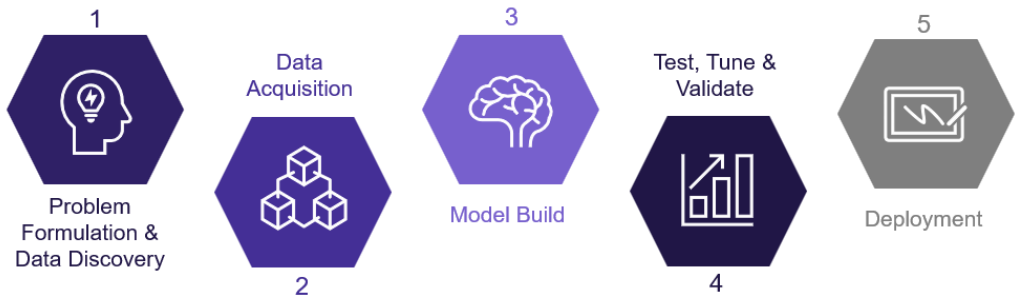
Bias in Algorithm

- COMPLIANCE RISK
- CONTROL RISK
- SECURITY RISK
- PERFORMANCE RISK
- ETHICAL RISK

brillio

Bias In Data

AI/ML Steps



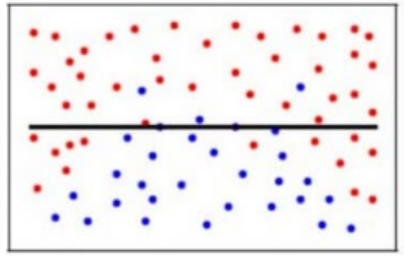
BIAS IN DATA

- **Data Labelling Bias** - Occurs when the annotation process introduces bias during the creation of training data
- **Outcome Proxy Bias -** Occurs when the machine learning task is not specified appropriately. Example : using the cost of a person to a health system is a biased proxy for the person's quality of health
- **Selection Bias –** Occurs when sample is unrepresentative of population. Example: Class Imbalance
- **Bias in Predicted Data –** Occurs during the evaluation phase
- **Bias in Incoming Data –** Occurs after the deployment of the model , primarily because of Data & Concept drift

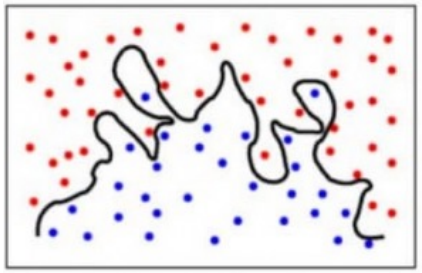
brillio

Bias In Algorithm

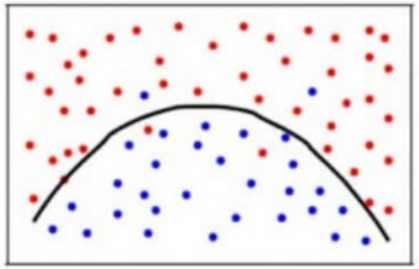
1) Bias Variance Trade-off amongst ML/DL Algorithms



Underfitting(High Bias)



Overfitting(High Variance)



Right-fitting(Low Bias, Low Variance)

2) Inherent Biasness in the results of selected algorithm

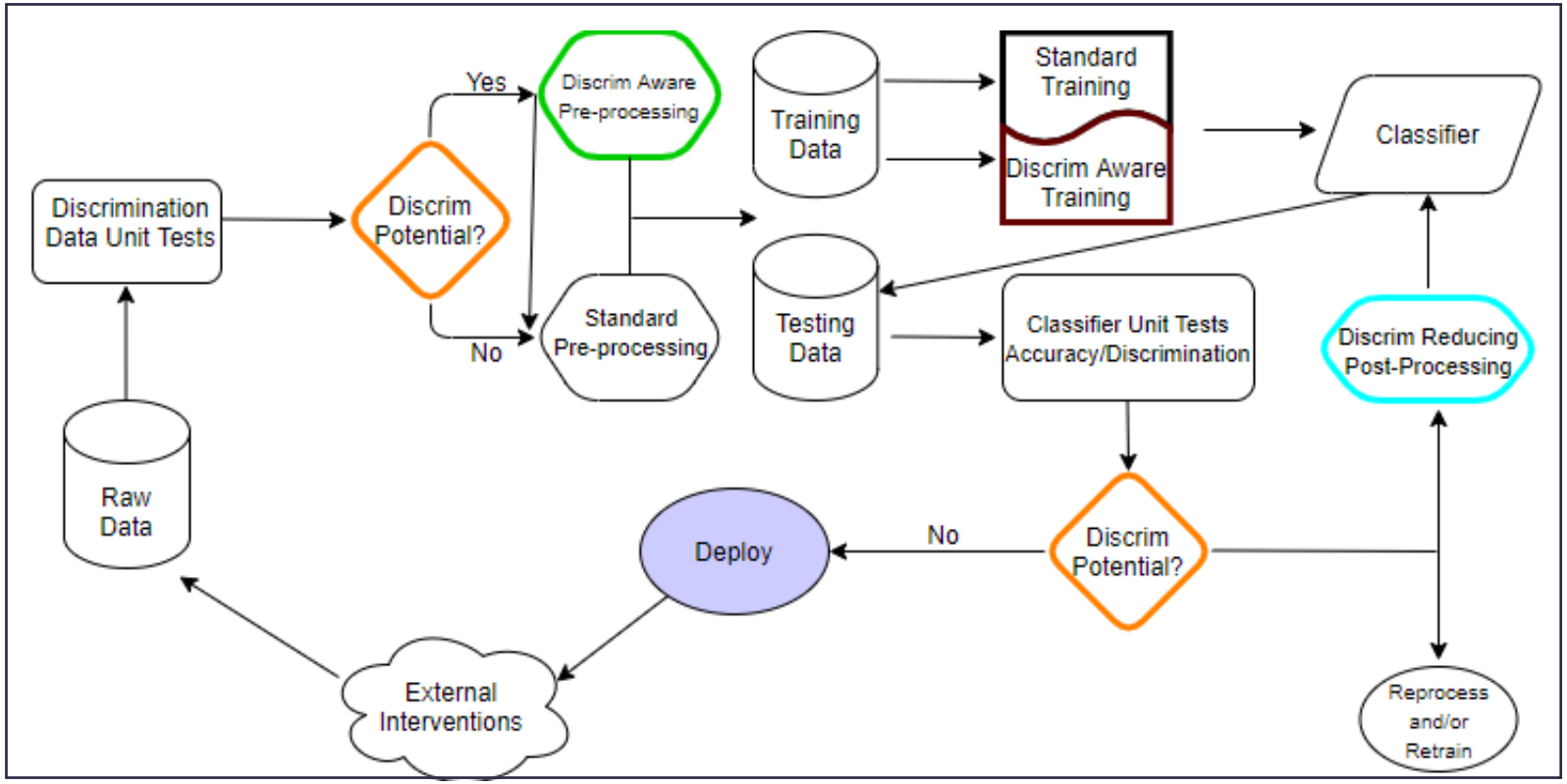
Tackling Bias

BIAS IN DATA

- Data Specific Techniques to remove Bias
- Algorithms to correct Biasness in Training Data (**Pre-Processing**)
- Fairness metrics to remove biasness in predicted data
- Avoiding Biasness after Deployment (Model Monitoring)

BIAS IN ALGORITHM

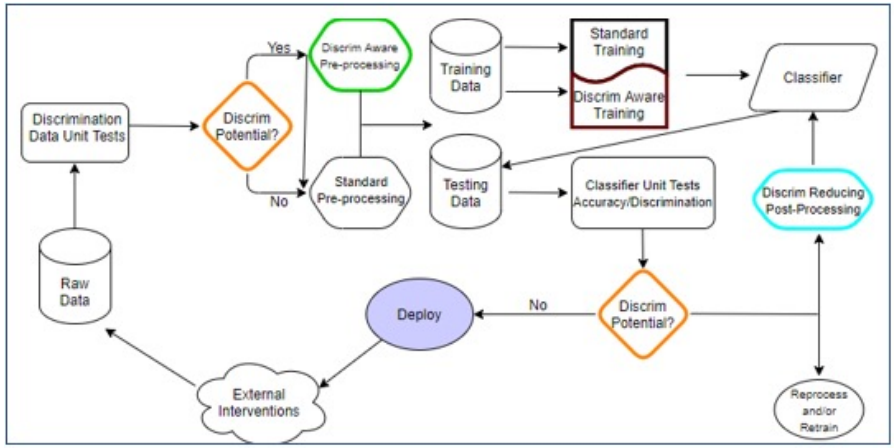
- Model Training Algorithm (Bias-Variance Trade-off)
- Penalizing Cost function during Model Training (tackling class imbalance)
- In-Processing Algorithms
- Bias Mitigation Algorithm to predicted Labels(Post Processing)



Data Bias Correction
Algorithm to correct Data Bias
Algorithmic Bias Correction
Algorithm to correct Data Label Bias

brillio

Brillio's Approach to Tackling Bias in Data



IDENTIFICATION & CORRECTION OF BIAS IN DATA THROUGH AI/ML PIPELINE

Data Specific Techniques to detect Bias

- Equal Parity Check
- Proportional Parity
- Conditional Demographic Disparity in Labels

Algorithms to mitigate Biasness in Training Data (Pre-Processing)

- Re-weighing Pre-Processing
- Optimized Pre-Processing
- Learning Fair Representation
- Disparate Impact Remover
- Oversampling/Under sampling Techniques

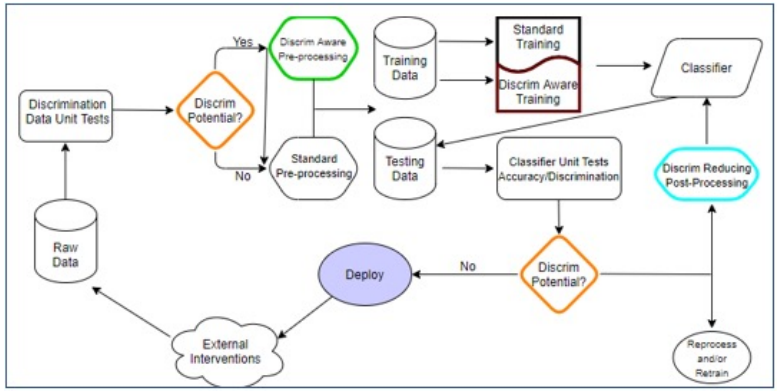
Fairness metrics to detect biasness in predicted data

- Specificity Score
- Sensitivity Score
- Difference in positive predictions in Predicted Labels
- Theil Index
- Equal Opportunity Difference

Avoiding Biasness after Deployment (Model Monitoring)

- Check distribution of Incoming data(IV) vs Benchmark Dataset (Training) & Predicted data(DV) vs Benchmark Labels(Training)
- Techniques to compare distributions: PSI, KS Stats, Histogram Intersection, Z-test, T-test, Control Charts
- Create Alerts based on Data drift & Concept Drift
- Wait for new data to be collected & then Re-calibrate the model & deploy

Brillio's Approach to Tackling Bias in Algorithm



IDENTIFICATION & CORRECTION OF BIAS IN ALGORITHMS THROUGH AI/ML

Model Training-Bias Variance Trade-off

- Regularization/Dropout
- Resampling
- Ensembling
- Removing Irrelevant Features

Penalizing Cost function during Training

- Balancing "class_weights" during Model training to tackle Class Imbalance problem

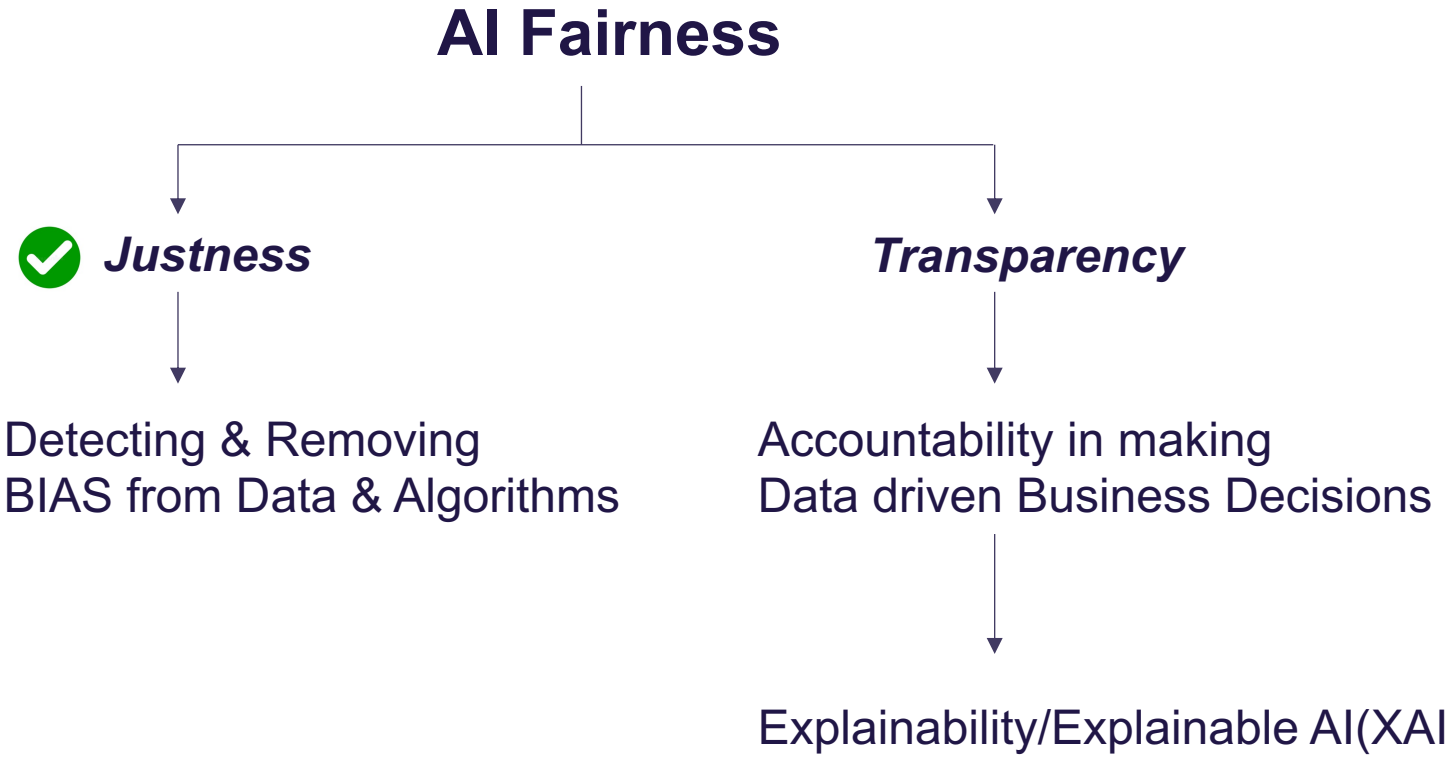
In Processing Algorithms

- Adversarial Debiasing
- Prejudice Remover
- Meta fair Classifier

Post Processing Algorithms

- Equalized Odds
- Calibrated Equalized Odds
- Reject Option Classification

AI Fairness - Overview



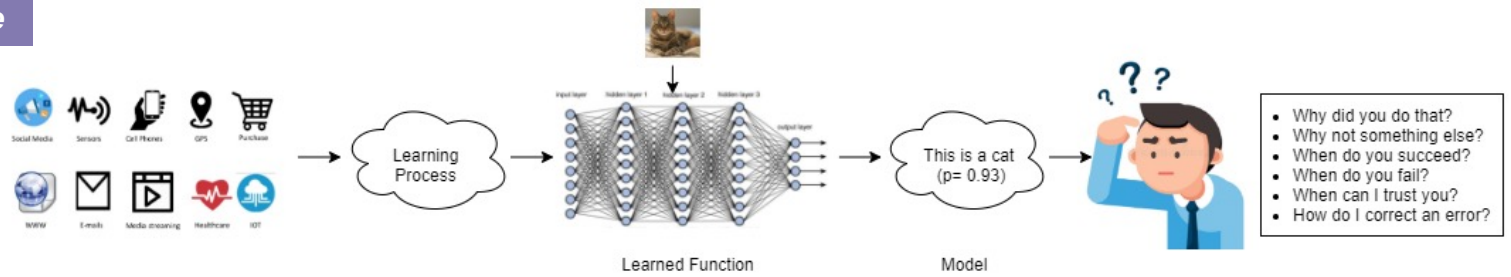
A formal definition:

" Explainable AI is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models. With it, you can debug and improve model performance, and help others understand your models' behaviour "

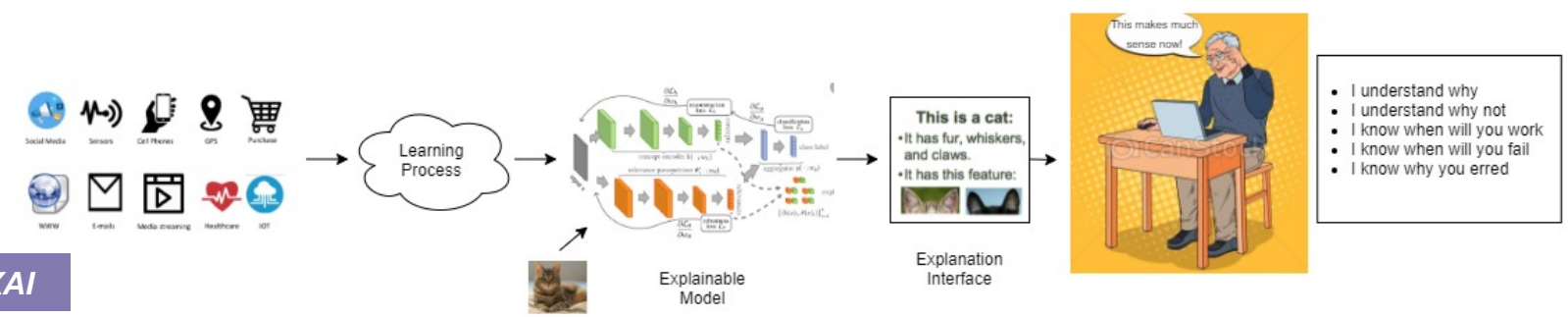
brillio

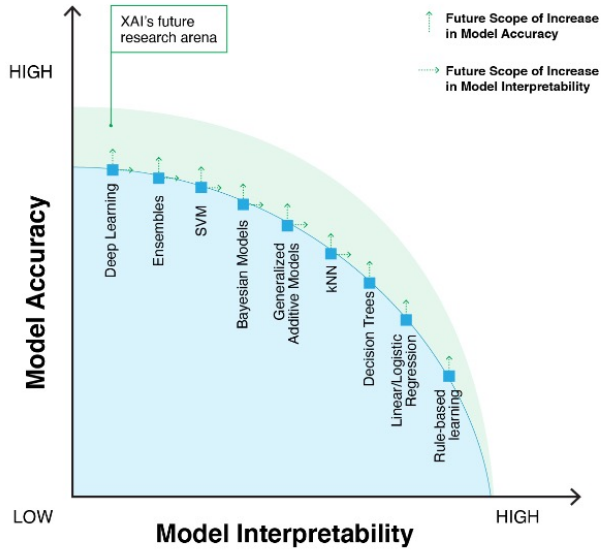
Explainable AI (XAI)

Before

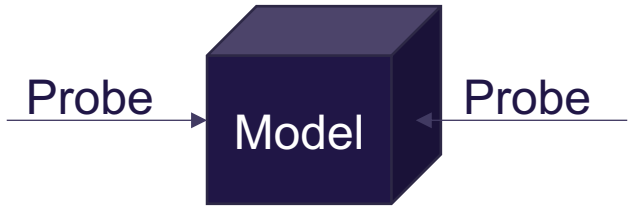


With XAI





Explainable AI (XAI) Tools : Lime and Shap



Explanations

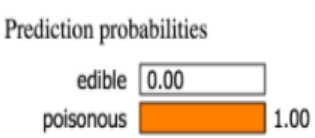
Global (Behaviour of the Model in General)

Local (Behaviour of the Model in a particular case)

LIME

***L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations*

- *Explains why model makes a specific prediction(**Local**)*
- *Applicable for Tabular, Text & Image data*
- *Faster than SHAP*
- *Doesn't explain what in general influenced the prediction*

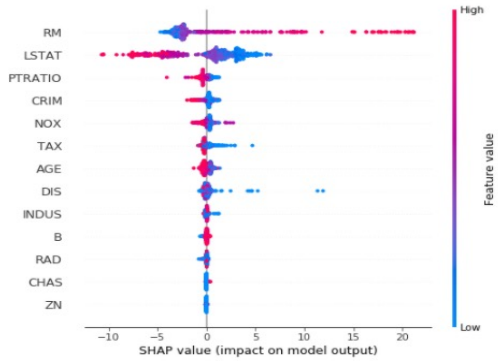


Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

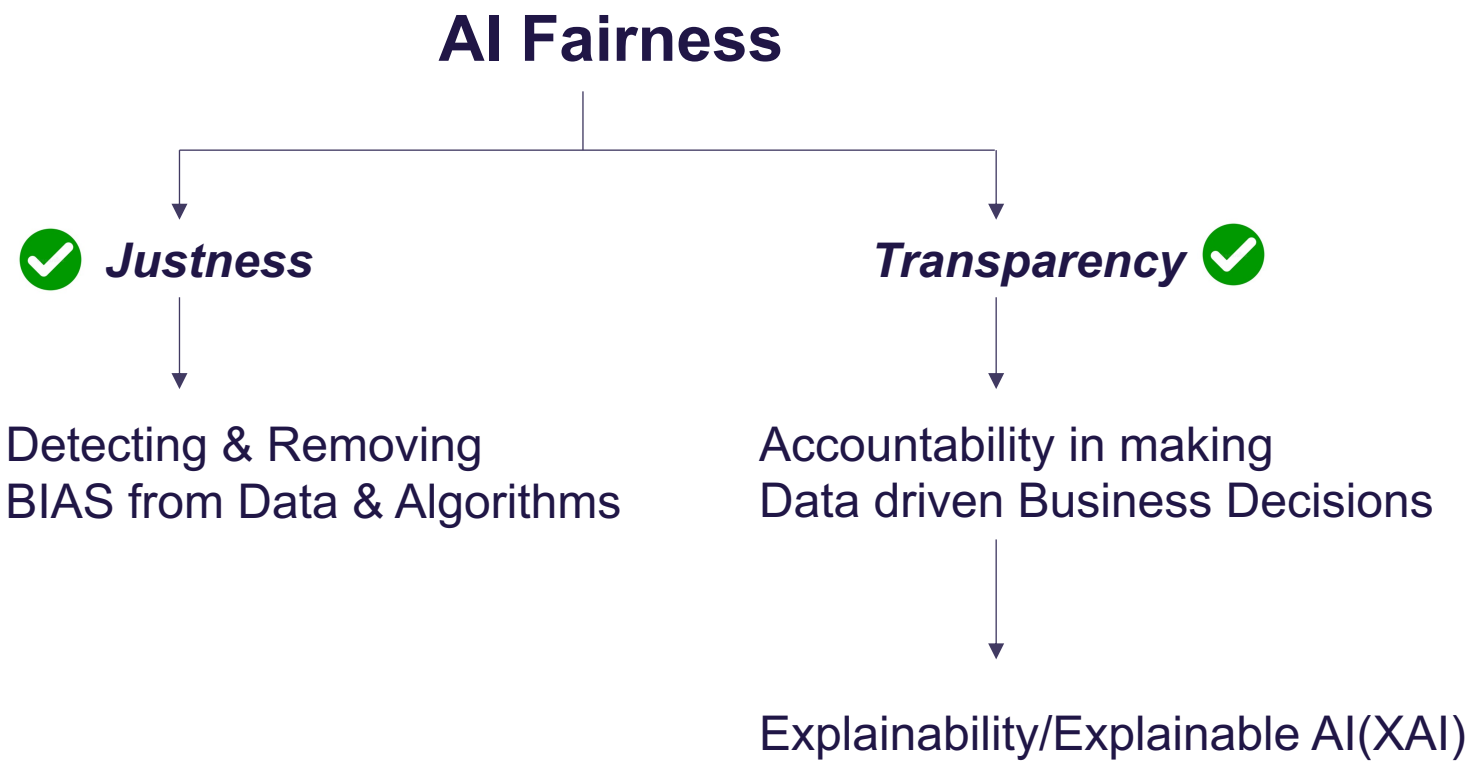
SHAP

***SH**apely **A**dditive ex**P**lanations*

- *Extended version of LIME – ensures accuracy & consistency of explanation*
- *Average marginal contribution of a feature value over all possible coalition (**Local+Global**)*
- *Applicable for Tabular, Text & Image data*
- *Slower than LIME & doesn't return a model as output (as LIME does)*



AI Fairness - Overview



A formal definition:

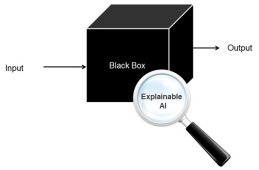
" Explainable AI is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models. With it, you can debug and improve model performance, and help others understand your models' behaviour "

brillio

Responsible AI - Principles

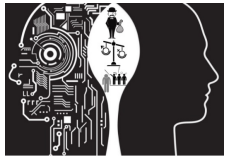
AI FAIRNESS

Justness



Transparency

ETHICAL AI



Governance

Privacy

brillio

Governance

Data Governance

- Data Security & Data Loss Prevention
- Data Integrity
- Data Lineage
- Data Completeness

Process Governance

- Formalizing steps in ML Lifecycle
- Formalize bringing Human in the Loop
- Ensure validation checks before Deployment
- Example of actions involved:
  → Reviews
  → Sin-Offs
  → Capturing supporting Materials (Documentation)

Model Governance

- Versioning to ensure traceabilit
- Experiment Tracking to select appropriate model
- Continuous Integration
- Continuous Deployment

Responsible AI sees strong Governance as the key to achieving fairness and trustworthiness.

brillio

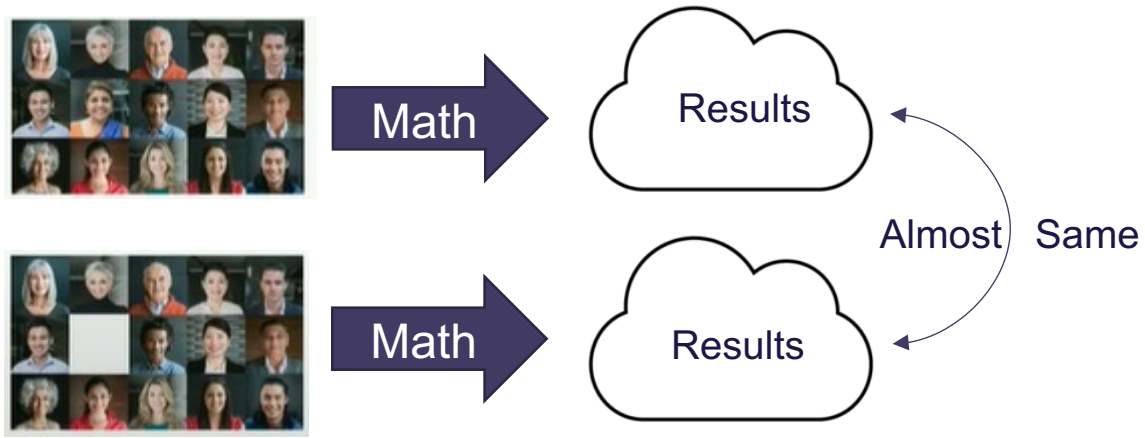
Ethical AI - Privacy

Methods to achieve Privacy in AI

"Protect the data before it enters the model"

Differential Privacy

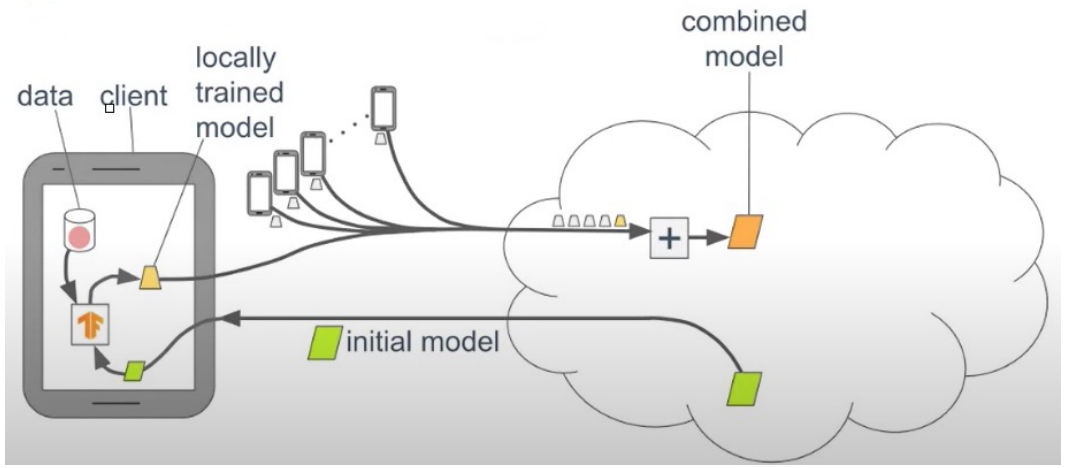
- Differential Privacy is a property & not a Technology
- AI system that is *differentially private* allows analysis while protecting sensitive data behind a veil of uncertainty



"Building protection into the Model"

Federated Learning

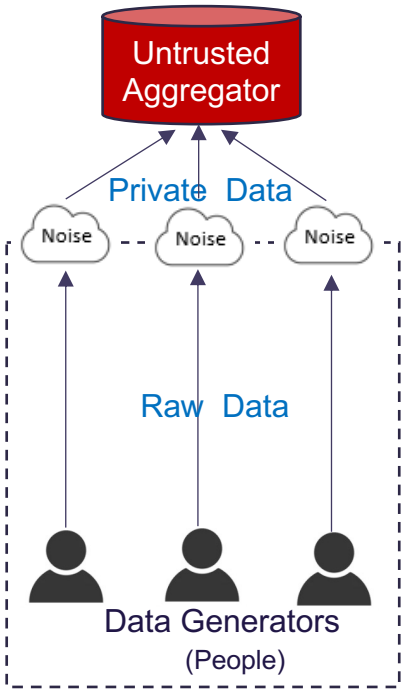
- Federated Learning is Decentralized Machine Learning
- Equivalent of pooling your data without sharing it



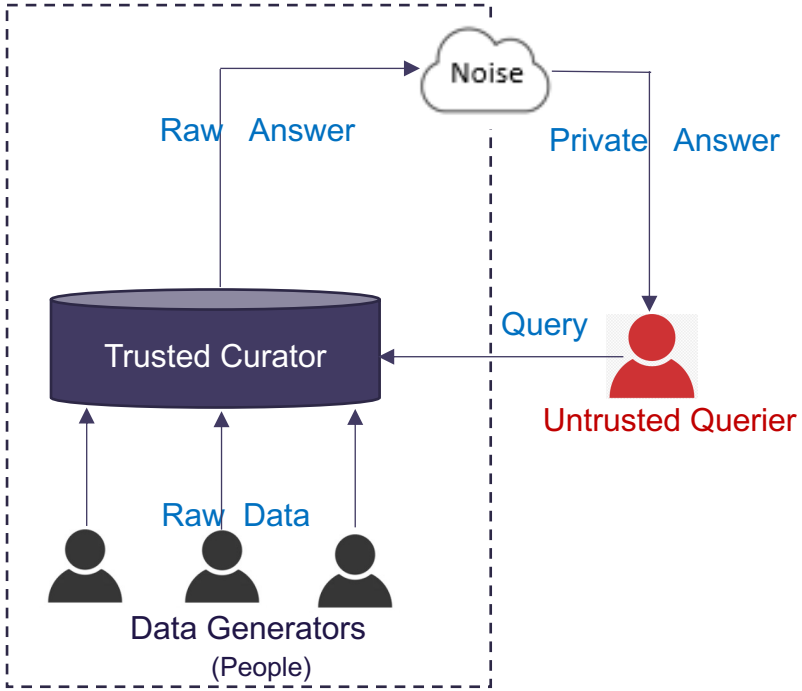
brillio

Ethical AI – Differential Privacy

- Differential Privacy , unlike most privacy-preserving tech , doesn't rely on Encryption
- There are 2 ways of achieving Differential Privacy : **Local** & **Global**



Local Differential Privacy

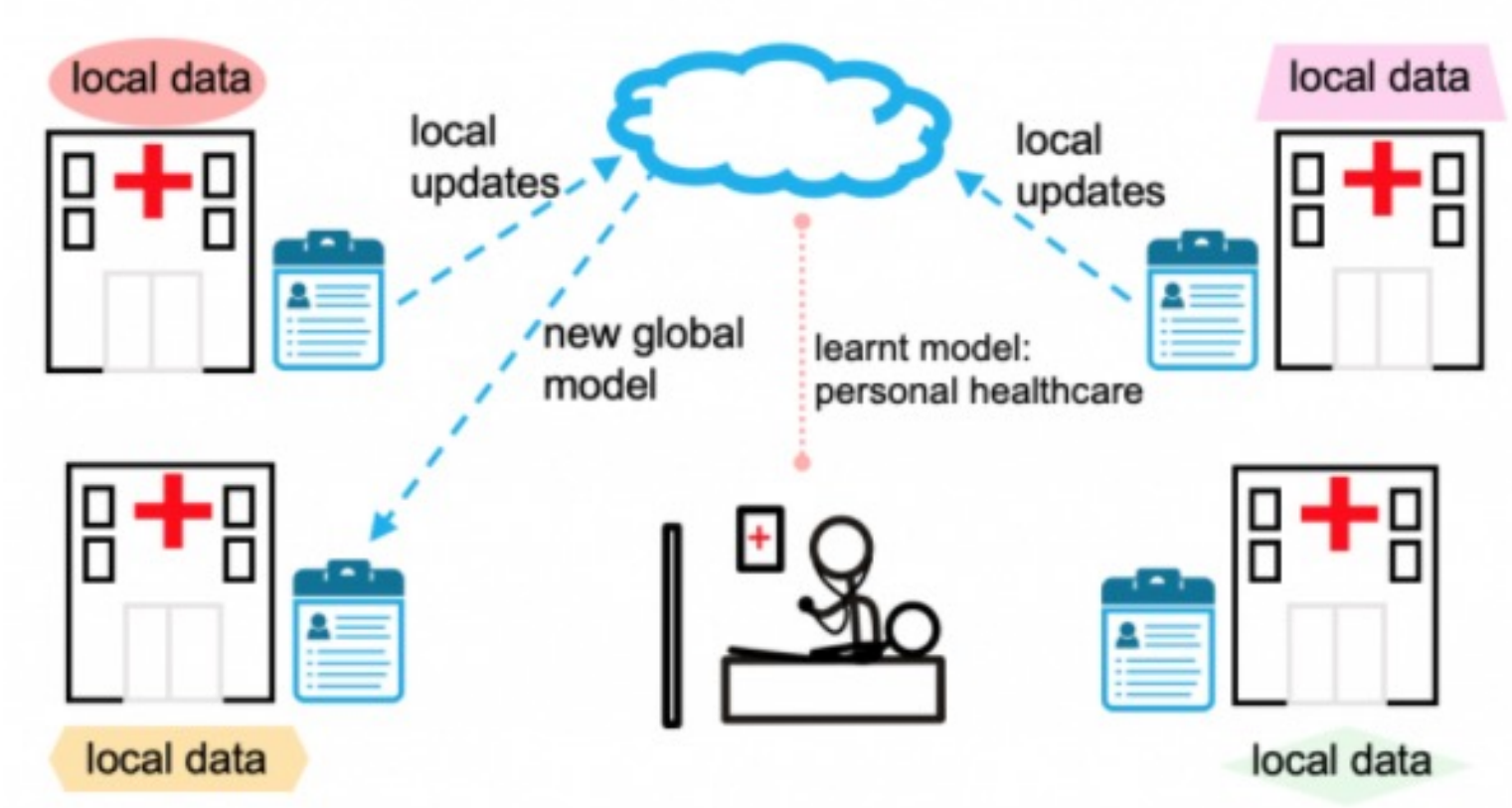


Global Differential Privacy

brillio

Ethical AI – Federated Learning

- Hospitals can be viewed as remote devices that contain multitude of patient data for predictive Healthcare.
- Federated learning reduces strain on the network and enable private learning between various devices/organizations



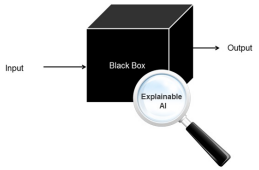
Federated learning for personal healthcare via learning over heterogeneous electronic medical records distributed across multiple hospitals.

brillio

Responsible AI - Principles

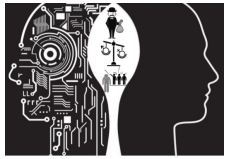
AI FAIRNESS

Justness



Transparency

ETHICAL AI



Governance

Privacy

brillio

Tools Overview

brillio

AI Fairness Tool Comparison

		Detect & Remove Bias		Enable Transparency			
	AI FAIRNESS TOOLS	*Removing Data Bias*	*Removing Algorithm Bias*	*Data Explanation*	*ML Model Evaluation*	*DL Model Evaluation*	*Production Monitoring & Explainability @ Scale*
OPEN SOURCE	IBM AI Explainability 360	✓	✗	✓	✓ Lime/SHAP & Others	✓ Lime/SHAP & Others	✗
	IBM AI Fairness 360	✓	✓	✗	✗	✗	✗
	XAI	✗	✗	✓	✓ (Alibi)	✗	✓ Monitors & Provides Explainability
	SHAP/LIME	✗	✗	✓	✓	✓	✗
CLOUD	AWS Clarify	✓ Only detects	✓ Only detects	✓	✓ Lime/SHAP	✓ Lime/SHAP	✓ Only monitors
	Google's Explainable AI	✓ Only detects	✓ Only Detects	✓	✓ Lime/SHAP & Others	✓ Lime/SHAP & Others	✓ Only monitors
	Microsoft's Responsible ML	✓ Only detects	✓	✓	✓ Lime/SHAP & Others	✓ Lime/SHAP & Others	✓ Only monitors

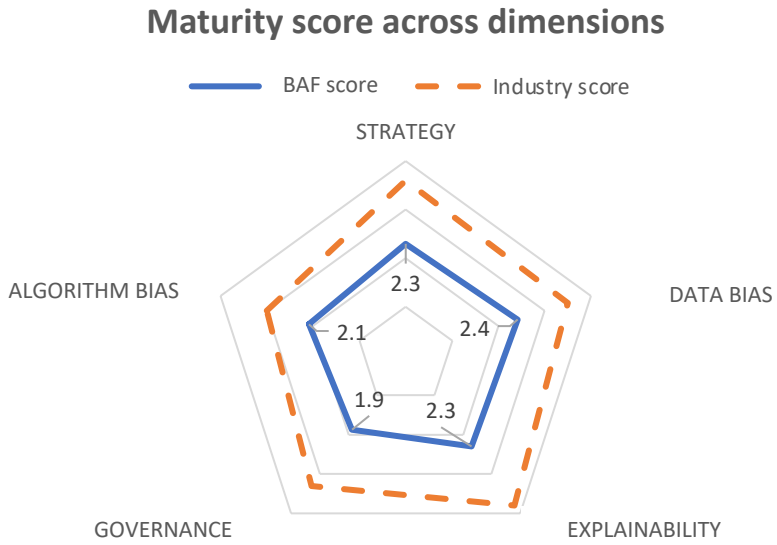
Engagement Models

brillio

Brillio's BAF – Bias Assessment Framework Helps Assess Current State and Map Gaps Quickly

Through our discussions with AI IT & business stakeholders, the importance and performance scores would be captured against each of the following dimensions

DIMENSIONS	SUB- DIMENSIONS
1 STRATEGY	Vision, Principles, Trainings, Maturity level, Problem formulation process
2 GOVERNANCE	Decision-making process with human-in-the-loop, Roles, Change management
3 DATA BIAS	Single version of truth, Data Specific Techniques to remove Bias, Fairness metrics to detect bias,
4 ALGORITHM BIAS	Model lifecycle management, Bias-Variance Trade-off, Penalizing Cost function during Model Training (tackling class imbalance), In-Processing Algorithms, Bias Mitigation Algorithm to predict Labels(Post Processing)
5 EXPLAINABILITY	Transparency through the ML process, ML/DL model evaluation, Model validator checks, Model monitoring & deployment process



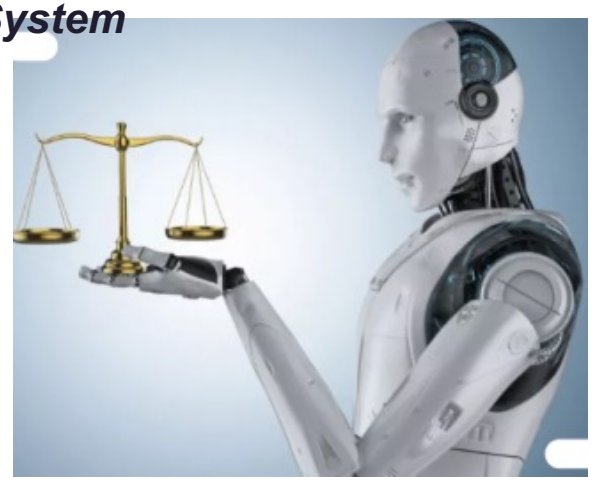
Overall Maturity: XX

How We Can Start Our Journey Together For “Responsible AI”

Existing AI/ML Systems

Brillio Accelerators

Unbiased Explainable AI/ML System



Discovery Phase (1-2 weeks)

- Assessment of existing data, Models and AI Fairness Maturity Level through **BAF**
- Target state - AI Fairness Maturity and Gap Analysis

Foundation Phase (2 weeks)

- Building Roadmap for enabling “Responsible AI” through Brillio Accelerators – Strategy, Technology, Process and Governance Framework

Execution Phase (1-3 weeks)

- Implementation: Explainable AI techniques, Data/Algorithm Bias Detection & its Removal
  - *Tools : Open Source & Cloud Platform*
- Workshops with data science team for smoother transition process
- Strategy, Role definitions, Process and Governance workshops with key stakeholders for evangelisation

brillio

Thank You

brillio