# Explainable Artificial Intelligence (XAI)

# User Manual

## Document Guidelines

This document serves as the basis regarding explainable artificial intelligence (XAI), its principles, risk management, methodology of implementation in the department and tests to comply with the guidelines. It aims to provide information about the significance of XAI concepts within the Data Science (DS) teams of Aboitiz.

The document shall serve as a guideline in assessing if a particular model, system, or data science practice adheres to XAI principles. It shall provide a practical methodology for assessing the alignment of Data Science and Artificial Intelligence **(**DSAI) systems with the XAI principles.

The scope of this document includes all models and artificial intelligence (AI) systems in development and currently deployed across DSAI. The scope of work applying the methodology shall be specific to the system under study and is calibrated to the level of risk associated with the system's operation. The XAI methodology and XAI-specific methods should be applied in consideration of and complexity of the models in study, their impact on the customers, and the on the financial performance or reputation of the organization.

Limitations of this document include correcting for change on systems not owned by the organization and that AI systems must adhere to laws and regulations by regulatory bodies of the organization or of the government. We would like to emphasize that this document (and accompanying toolkit) does not define ethical standards; it only aims to provide a way for DSAI System Developers and Owners to demonstrate their claims about the performance of their DSAI systems according to the XAI principles. It also does not guarantee that any DSAI system tested using this Framework will be free from risks or biases.

# Introduction

## DSAI Systems in Context

### What is DSAI?

DSAI systems are defined as technologies that aim to assist or replace traditional human decision-making. DSAI systems can give organizations the ability to learn complex relationships from large volumes of data and perform quantitative risk-aware decision-making at a vast scale.

These technologies can improve business processes, allowing businesses to serve customers better and more efficiently, and potentially avoid the variability from individual humans making decisions. However, DSAI systems operate according only to the data and objectives explicitly provided by their designers and do not understand the broader context or moral constraints that humans take for granted and adding a *"human in the loop"* does not automatically eliminate all fairness risks. As a result, DSAI systems can cause unintended harms and perpetuate or reinforce existing disadvantages that exist in society, which in turn can create reputational, operational, and legal risks for companies. The heightened risks for businesses and customers arise whenever consequential decisions are being made at high speed or volume, whether the underlying decision-making algorithms are relatively simple or more complex (such as neural networks trained with machine learning).

DSAI systems encompass more than mathematical models and data. An DSAI system often includes downstream or upstream business rules (such as eligibility criteria), employees who may review or change automated decisions, and processes and mechanisms to monitor and review decisions and outcomes.

In general, DSAI systems produce decisions which in turn creates outcomes. An example would be an DSAI system that can produce a decision on approving a customer's loan application which would in turn have the outcome that a customer may either have access to credit or in some cases default. With this, the FEAT principles are developed which considers both whether the decision-making process and the outcomes from DSAI-driven decisions are justified.

### What are the FEAT principles?

The FEAT Principles were developed by the Monetary Authority of Singapore (MAS) in 2018 to ensure fairness, ethics, accountability and transparency in the use of DSAI in the financial sector. The section below will give a brief summary of the FEAT Principles.

**Fairness**

*Justifiability*

1. DSAI-driven decisions are not systemically causing a disadvantage to individuals or a group of individuals, unless those decisions can be justified.
2. If any personal attributes are used as input factors in model building, they are required to be justified.

*Accuracy and bias*

1. There must be a periodic review and validation of the data and models used for DSAI-driven decisions, to check for their accuracy and relevance and to reduce biasness.
2. There must be a periodic review of DSAI-driven decisions so as to determine that the models'' design and intent is correct.

**Ethics**

1. The firm's code of conduct and ethical standards must align with the use of DSAI.
2. Before relying on DSAI-driven decisions, they must be assessed to have at least upheld the ethical standards of human driven decisions.

**Accountability**

*Internal accountability*

1. The use of DSAI by a corporation must be approved by an appropriate internal authority.
2. Firms must be held accountable for any internal or externally sourced models that uses DSAI.
3. Management and the Board of a corporation should always be made aware of the use of DSAI in their projects.

*External accountability*

1. Data subjects should be given access to channels to enquire about and request for reviews of DSAI-driven decisions that involve or affect them.
2. When performing a review of DSAI-driven decisions, any relevant supplementary data provided by the data subjects must be considered.

**Transparency**

1. When using DSAI, corporations are recommended to proactively disclose this to their data subjects, so as to boost public confidence.
2. Clear explanations are to be provided to data subjects on what data is required and how this data influences their DSAI-driven decisions
3. Clear explanations are to be provided to data subjects on any potential consequences that the DSAI-driven decisions may have, with regards to them.

The FEAT Principles are an effective, industry-backed set of guidelines that will help DSAI systems in the financial service contexts to produce well-rounded DSAI systems. However, since DSAI is exposed to models outside the financial services industry, we cannot rely on these principles alone to ensure total effectivity with regards to XAI principles.

## What is XAI?

Explainable and Responsible AI (XAI) is a blanket term of multiple concepts related to how DSAI systems should be created while achieving technical excellence. Inspired by the work of the Monetary Authority of Singapore and the Infocomm Media Development Authority (IMDA), we hope to define a set of principles that data scientists from Aboitiz Data Innovation (ADI) can use while developing and deploying their DSAI models. These principles, which will be described in further detail, are the following:

- Transparency

- Explainability
- Repeatability/ Reproducibility
- Safety & Security
- Robustness
- Fairness
- Data Governance
- Accountability
- Human Agency and Oversight
- Inclusive Growth, Societal & Environmental Well-being (Ethics)

If there is an existing set of established principles such as FEAT that are already being used by DSAI systems, what then is XAI and why do we need to define a separate term? We posit that XAI concepts can be applied to DSAI systems not limited to the financial services domain. In order to elevate the effectiveness of DSAI models in the fields that Aboitiz has expertise in – such as power, real estate, construction, assets, food – similar guidelines and tools should be made that are inclusive to the needs of these domains. A broader Framework such as this should be developed so that all models that ADI create will not be prone to excessive bias or other XAI-related inconsistencies.

With this, we define three (3) XAI-based intervention stages. These are arbitrary phases encompassing the DS pipeline by which XAI principles can be applied differently:

- **Input** – Data preparation and understanding, inputs to model, extraction, preprocessing, what data to include, data quality. This includes synthetic data, and any intermediate features (in an intermediate layer in a neural network)
- **Model and Output (M&O)** -- development, initial results/output of model and interpretation
- **Deployment and Monitoring (D&M)** – sharing to stakeholders, dashboard and input to business process

# XAI Principles

As proposed by the IMDA, the Explainable and Responsible AI (XAI) principles can be grouped into five pillars:

- Transparency on the use of AI and AI Systems
- Understanding how DSAI System reaches a Decision
- Safety and Resilience of AI Systems
- Fairness/ No Unintended Discrimination
- Management and Oversight of AI System

These pillars hold previously mentioned XAI principles. We will also be looking at each principle on its viability through each XAI-based intervention stage: Input, Model and Output (M&O), and Deployment and Monitoring (D&M). We describe technical tests that can be employed throughout each intervention stage. Do note that not all principles have technical tests applicable on each stage; we will be mentioning this at its specific section.

## Transparency on use of AI and AI Systems

Appropriate info is provided to individuals impacted by AI system. A better word is 'apparent', as in, readily understood and not hidden (within a black box).

### Transparency

The implementation of AI in the system will be disclosed to end-users. Empowered individuals can make an informed decision if they want to use the AI-enabled system.

#### *Input*

No technical tests are seen to be applicable for this XAI principle at this stage for System Developers, although transparency can be practiced in data collection stage especially when collecting data from surveys or mobile applications.

#### *M&O*

No technical tests are seen to be applicable for this XAI principle at this stage, but it should be clear to DSAI System Owners and Assessors what features found in the data are and are not being used and transformed.

#### *D&M*

Data collection policies should be transparent to alleviate concerns related to AI. Process checks of documentary evidence (such as company policy and communication collaterals) of providing appropriate information to users who may be impacted by the AI system. The information can include, under the condition of not compromising IP, safety, and system integrity, use of AI in the system, intended use, limitations, and risk assessment.

## Understanding how DSAI System reaches a decision

Ensuring AI operation/results are explainable, accurate and consistent. This enables users to know the factors contributing to the AI model's output (decision/recommendation).

### Explainability

Explainability refers to the concept that the end user can understand why a prediction is made by an AI system. Global explanation methods attempt to explain the model behavior as a whole while local explanation methods focus on explaining an individual prediction.

***Input***
This is the constituent features used in the model that are being explained, and presented in a data dictionary. This also includes explanations with respect to an intermediate layer of the network. Internal explanations are necessary for models with an inherent structure that is sequential, such as intermediate layers of a neural network (Leino et al., 2018) or branching structures within decision trees (using Information gain to understand the variable where it was split).

***M&O***
Interpretable models are machine learning (ML) models with a simple structure (such as sparse linear models or shallow decision trees) that can 'explain themselves' i.e., are easy for humans to interpret. Post-hoc explanation methods can analyze and explain a relatively more complex ML model after it has been trained.

***D&M***
This assesses the impact of features on model outcomes to stakeholders. Example is Partial Dependence Plots (DPD) which provides visual interpretation of marginal chances in model outputs when a feature is changed. Process checks include verifying documentary evidence of considerations given to the choice of models, such as rationale, risk assessments, and trade-offs of the model.

### Repeatability / Reproducibility
AI results are consistent, and can be replicated by owner, developer or another third party.

***Input***
Ensure that any attempts to randomize or stratify data during splitting are done with set seeds.

***M&O***
Ensure that model files are saved with specified parameters and can be linked to clear datasets. Create a requirements.txt file or similar file that documents the necessary packages needed to run files on programming languages. Another good practice is to ensure that preprocessing steps or other methodologies applied to the dataset and model creation are documented well.

***D&M***
Assess through process checks of documentary evidence including evidence of AI model provenance, data provenance and use of versioning tools.

## Safety and Resilience of AI Systems
This principle states that the AI system should be reliable, performs as intended, and will not cause harm.

### Safety and Security
AI system is safe. Conduct impact and risk assessment. Known risks have been identified and/or mitigated.

***Input***
No technical tests are seen to be applicable for this XAI principle at this stage. However, should the data contain any private and personally identifiable data, it will be explored through company-specified environments and platforms.

***M&O***

No technical tests are seen to be applicable for this XAI principle at this stage.

***D&M***

Assess through process checks of documentary evidence of materiality assessment and risk assessment, including how known risks of the AI system have been identified and mitigated.

### Robustness

This assesses whether the AI system can still function despite unexpected inputs. Model robustness refers to the degree that a model's performance changes when using new data versus training data. Ideally, performance should not deviate significantly.

***Input***

This refers to data quality, which is the accuracy, completeness and clarity of the data being used to train the model.

***M&O***

Technical tests attempt to assess if the model performs as expected even with unexpected inputs.

***D&M***

Process checks include verifying documentary evidence, review of factors that may affect the performance of the AI model, including adversarial attacks.

## Fairness / No Unintended Discrimination

### Fairness

End-users need to know that the training data is reflective of the characteristics of the population being analysed. An AI system will be considered biased if it discriminates against certain features or groups of features. This provides an opportunity to address biases by detecting them and measuring them at each stage of the ML lifecycle.

***Input***

The training data may not have sufficient representation of various feature groups or may contain biased labels. This imbalance can conflate the bias measure, and our models may be more accurate in classifying one class than in the other. We need to choose bias measures that are appropriate for the application and the situation. So, we want to utilize metrics that can be computed on the raw dataset before training.

***M&O***

Any bias that arises post-training may emanate from biases in the data and/or biases in the model's classification and prediction. Models built on biases would reproduce or exacerbate those biases during predictions in each stage of the ML cycle. After training the ML model, we gain additional information from the model itself, in particular, the predicted probabilities from the model and the predicted labels. These allow an additional set of bias metrics to be calculated and analyzed.

***D&M***

It is quite possible that after the model has been deployed, the distribution of the data that the deployed model sees, that is, the live data, is different from that of the training dataset. This change, also known as *concept drift*, may cause the model to exhibit more bias than it did on the training data. The change in the live data distribution might be temporary (e.g., due to some short-lived behavior like the holiday season) or permanent. In either case, it might be important to detect

these changes. Same bias metrics during modelling can be monitored at continuous intervals. This frequency can be two days, a week, a month, etc. depending upon the DS team and the use case.

### Data Governance

Source and data quality should be sound. Good data governance practices should be in place when training AI models.

***Input***

No technical tests are seen to be applicable for this XAI principle at this stage.

***M&O***

No technical tests are seen to be applicable for this XAI principle at this stage.

***D&M***

We need to hold ourselves accountable when deploying AI applications, especially when users are concerned about how we access and use confidential information. The two takeaways are segmentation and visibility. We have to ensure that we can monitor and restrict how our models use data at all stages. Segmentation prepares and mitigates the impact of a breach to keep user information and data as safe as possible. Also, data collection policies should be transparent to alleviate concerns related to AI. Taking a strong approach towards maintaining high privacy and governance standards will further ensure we are legally compliant.

## Management and Oversight of AI

Ensure human accountability and control, and that the AI-enabled system is developed for the good of humans and society.

### Accountability

Proper management oversight of AI system development

***Input***

No technical tests are seen to be applicable for this XAI principle at this stage.

***M&O***

No technical tests are seen to be applicable for this XAI principle at this stage.

***D&M***

Assess through process checks of documentary evidence, including evidence of clear internal governance mechanisms for proper management oversight of the AI system's development and deployment

### Human Agency and Oversight

AI system designed in a way that will not decrease human ability to make decisions.

***Input***

No technical tests are seen to be applicable for this XAI principle at this stage

***M&O***

No technical tests are seen to be applicable for this XAI principle at this stage.

***D&M***

Assess through process checks of documentary evidence that AI system is designed in a way that will not reduce human's ability to make decision or to take control of the system. This includes defining role of human in its oversight and control of the AI system such as human-in-the-loop, human-over-the-loop, or human-out-of-the-loop.

## Inclusive Growth, Societal, and Environmental Wellbeing (Ethics)

Beneficial outcomes for people and planet. As with the concept's nature, there are no standard technical tests nor process checks that can be made readily available. However, DSAI System Developers and Owners should take steps to critically consider such factors despite not being of a technical nature.

***Input***

No technical tests are seen to be applicable for this XAI principle at this stage.

***M&O***

No technical tests are seen to be applicable for this XAI principle at this stage.

***D&M***

No technical tests are seen to be applicable for this XAI principle at this stage.

# XAI Methodology

As we have described the XAI Principles, we will present methodologies that exemplifies these principles throughout the three intervention stages of 1) Input, 2) Model and Output (M&O), and 3) Deployment and Monitoring (D&M). Note that theoretical assumptions that adjustments made to input data, model training, and other aspects of the ML process will decrease the accuracy or performance of the models, remain unfounded (Rodolfa et al., 2021). The aim of this section is to provide a technical reference to integrate XAI methods into DS projects, so that each DS team can make design choices, and take practical steps to enable explainability to other developers and stakeholders.

## Input

Usually when dealing with datasets, there may be groups or segments that are underrepresented and that can be disadvantaged when a model is trained for prediction. For example, certain age or ethnic groups may be disadvantaged in application models due to their lack of prevalence in datasets. In general, we should always consider any potentially disadvantaged groups that should be protected.

### Dealing with Data

#### Test for (binary) class in data

Examples of tests for class imbalance: proportions, Difference in Positive Proportions in Labels (DPL), KL divergence (KL), Jenson-Shannon divergence (JS), Lp-Norm (LP), Total variation distance (TVD), Kolmogorov-Smirnov (KS), Conditional Demographic Disparity in Labels (CDDL).

#### Test for multicategory labels

In this case, there are two approaches (a) collapse categories to binary and compute label imbalance measures. Or (b) compute label imbalances across all multiple categories. (a) is a special case, but it requires a human to examine and consider labels to be grouped and discover which ones are more significant.

#### Errors, Biases, and other Properties in Data

What errors, biases, properties in data used by system may impact system's adherence to XAI principles? Tools like EvoML by TurinTech (https://github.com/EvoML/EvoML) perform data quality inspection, and apply relevant techniques to ensure data is AI-ready.

- **Sampling procedures** (i.e. unfair representation from previous data that is targeted, or presence of attrition that under-represents another long-term outcomes)
- Systematic **errors in measurement**
- **Predictions from other models** used as data for a system
- Data which **requires defining measurable proxies** (i.e. low socio-economic status by 'earns less than X per annum')
- **Relevancy of the dataset**, since only attributes that provide meaningful information to the domain should be considered
- **Accuracy of any labels/annotations**. This can be examined by recall or Intersection over Union (IoU). IoU has a range of within [0,1] that provides the mean average precision, specifying the amount of overlap between predicted and ground truth.

- **Size of the raw data corpus** (training, validation, test sets), since dataset size will depend on domain of the task, and complexity of the model. A point to note is also there is a threshold dataset size after which model performance plateaus.
- **Variance of each class in data**, as the number of classes proportionally increase data size
- **Type of classification** (e.g., are we predicting class for a data point, an entire image or each pixel in an image?)
- **Comprehensiveness of the dataset,** since features used in the model should have enough samples of 'edge' cases for better generality during training
- **Data representation due to change over time** (e.g., more customers are performing transactions in an app over time, than via a website)

## M&O

### Showcase diagrams of the entire data science pipeline

This enables us to showcase to users/ stakeholders the customized tools and processes we use to achieve maximum value from the data, and how we reach our final decision-making.

**Data flow diagrams (DFDs)**

DFDs are graphical representation of data flows through the AI system. This will include data stores and any data movement through subprocesses.

Scikit-learn offers a display of the multiple processes/ functions used during model development. The visual display of each step taken to prepare the data, tune the model and transform the predictions, will enable developers and users to evaluate specific data transformations and model configurations. This will further ensure consistency and reproducibility when data or model is being updated.

### Qualitative inspection

Process checks include verifying documentary evidence of having a strategy for the selection of fairness metrics that are aligned with desired outcomes, and the definition of sensitive attributes are consistent with legislation and corporate values. Establish qualitative guidance on how models should be classified. For example, adopt framework to classify models 'can result in systematic disadvantage to individuals or groups' to have 'high impact', while models with outputs that are 'benign and customers can opt in/out' to be of 'low impact'.

Other relevant details during the modeling process should, if necessary, be disclosed to ensure transparency to users and other stakeholders. Examples of relevant details can be:

- ***Coverage period***, which refers to the period of time which the model is trained and tested. This helps to ensure that the coverage period is representative of the data that the model was trained on, and that the model is reliable and accurate.
- The ***portfolios or products*** that the model is intended for use should be reported. This includes specifying the type of data that the model was trained on (such as numerical data, text, or images, etc). The context or domain of the data should be considered to ensure that the model is appropriate for the specific application.
- ***Exceptions*** are cases where the model would perform differently from the norm. This can include edge cases or unexpected events. This is important to inform and explain to users how the model will handle these exceptions.

## Model should be robust

Unexpected inputs should be converted or raise an exception (e.g., the use of numerical data points as inputs where string is intended). Randomly generated noise (not necessarily normally distributed) can be added to the test data to determine model performance. Noise deviating over several standard deviations (or scales) can be used to determine decision boundaries, i.e. the acceptable accuracy or acceptable range of change in model performance due to various amounts of noise in dataset.

## Explainability tools

Quantitative measures provided from explainability methods can supplement/ complement qualitative criteria by providing additional perspectives that can contribute to a more holistic assessment of fairness in the AI system. Explanation libraries can either be model-specific (e.g., designed for neural networks or other differentiable models) or model-agnostic (i.e., applicable for any ML model, after training). Explanations are presented in the form of Shapley values, baselines and counterfactuals, attribution methods, aggregate attributions, comparison to alternative global explanation methods/ another structure of a similar model.

- Model-specific tools/libraries:
    - Explain Like I'm Five (ELI5) (sklearn regressors and classifiers, XGBoost, CatBoost, LightGBM, Keras)
    - Activation Atlases (neural networks)
    - What-if Tool (WIT)(TensorFlow models, XGBoost and Scikit-Learn models).
- Model-agnostic tools/libraries:
    - skater (deep neural networks, tree algorithms, and scalable Bayes)
    - InterpretML (LIME, SHAP, linear models, and decision tree)
    - Alibi Explain, azureml-interpret
    - Rulex Explainable AI (Logic learning machine)
    - Model Agnostic Language for Exploration and Explanation (DALEX)(xgboost, TensorFlow, h2o).
- Other explainability techniques:
    - *Feature importance analysis/techniques* – This aims at generating a feature score that is directly proportional to the feature's effect on the overall predictive quality of the model. Examples: mean decrease impurity (MDI), mean decrease accuracy (MDA), single feature importance (SFI).
    - *Force dependence plots* – This presents a scatter plot that shows the effect of a single feature on model predictions.
    - *Baselines and counterfactuals* – This selects a baseline that introduces the concept of a baseline score (to compare against).
    - *Causal inferences* – This technique tests the causal relationships based on model outcomes.

## Examine model output for unintended bias and uncertainties

After training the ML model, we gain the predicted probabilities from the model and the predicted labels. Additional set of bias metrics can be calculated and analyzed for bias. This is to inspect any disparities in outcomes despite checks for class imbalance before training. Some examples include:

- Difference in positive proportions in predicted labels (DPPL)
- Disparate (Adverse) Impact (DI)

- Difference in conditional outcome (DCO)
- Recall Difference (RD)
- Difference in label rates (DLR) - note difference in acceptance rates (DAR) or difference in rejection rates (DRR) and Precision Difference (PD), Accuracy Difference (AD)
- Treatment Equality (TE)
- Conditional Demographic Disparity in Predicted Labels (CDDPL)
- Counterfactual Fliptest (FT)

**AI predictions always contain a level of uncertainty**, due to imperfect and noisy data, and the presence of incomplete distribution of possible data points. For example, predicting traffic congestion at 5am, would carry a high level of uncertainty, since the training/test data is collected from 6am to midnight. ML models need to present uncertainties to end-users for better explainability, provide indicators during decision making, and identify limitations of the AI-enabled system. Providing an estimate of uncertainties when using AI models is already mandatory in some domains (like medicine) as proposed by the European Commission(European Commission, 2021). In any case, quantifying and reporting uncertainties will build trust with the use of AI models. Quantification methods to determine uncertainties can be based on resampling methods, such as the distribution of the leave-one-out residuals estimated by perturbed models. Other methods include the Jackknife+ method, which presents a confidence interval that a new observation would lie within a prediction interval (Barber et al., 2019).

### Performance vs perceived goodness of explanation

There might be trade-offs between quantitative estimates of a system's performance and XAI objectives. For example, ML models with high performance are often based on complex algorithms with low explainability, and vice versa. While XAI objectives can fill the gap between complexity and interpretability, a lot is still dependent on how the end-user will integrate the given information with the decision-making. The choice of models being utilized and deployed will depend on the specific context and goals of the project in each DS team. Nevertheless, we encourage DS teams to intentionally design ML systems that maximize both model performance and explainability (Rodolfa et al., 2021).

To evaluate how end-users evaluate and perceive explainability, satisfaction of explainability can be performed by receiving feedback from developers, users, stakeholders, etc, with the use of Explanation Goodness Checklist, Explanation Satisfaction Scale, Trust Scales (Hoffman et al., 2018). To determine possible trade-off between performance and interpretability, we can provide estimates such as percentage likelihood in probability of harm, which will help to align understanding across different stakeholders who perform and validate those assessments.

## D&M

### Check for model drift

Model drift is the reduction in a model's predictive capability as a result of changes in the external environments. Model drift could be caused by many reasons, which include changes in the technological environment and the consequential changes in relationship between variables. Model drift can be broadly classified into two categories: data drift and concept drift.

Some methods can be used to inspect for model drift. The Kolmogorov-Smirnov test (KS test) which is a nonparametric test, compares the training and post-training data. If the KS test states that the data distributions from both datasets are different, this will confirm the presence of model

drift. The population stability Index (PSI) compares the distribution of the target feature in the test dataset with the training dataset used to develop the model. Another method, the adaptive Windowing (ADWIN) algorithm adopts a sliding window approach to detect model drift. ADWIN slides a fixed-size window to detect any changes in the new data, and an alert is set, if the changes exceed a pre-set threshold.

### Check for data drift

Data drift can happen when there is a significant gap between the time the data is collected and when the model is used to predict outcomes using real-time data. If this problem is not addressed in a timely manner, this will negatively impact business decisions that were reliant on the model's predictive capabilities.

One of the common practices in catching data drifts is by using out-of-time (OOT) testing. OOT testing is the process of testing the model using unforeseen data and inspecting the model's performance (ie., any decrease in the predictive model performance). A suggested threshold for data drift and retraining is if model performance decreases by more than 15%. However, this threshold value can be selected based on each DS team and use case. Furthermore, the Population Stability Index (PSI) or the Characteristics Stability Index (CSI) can be used to quantify the magnitude of the data drift, and these indices can be communicated to the business team that retraining of the model may be required.

### Check for concept drift

Same bias metrics during modelling can be used to inspect concept drift. Inspections can be set at frequency selected based on DS team and use case (for example, DDPL bias can be computed every two days, and alert is set if bias metric exceeds confidence interval).

# References

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2019). Predictive inference with the jackknife+. *ArXiv: Methodology*.

European Commission. (2021). *Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. *CoRR*, *abs/1812.04608*. http://arxiv.org/abs/1812.04608

Leino, K., Li, L., Sen, S., Datta, A., & Fredrikson, M. (2018). Influence-Directed Explanations for Deep Convolutional Networks. *CoRR*, *abs/1802.03788*. http://arxiv.org/abs/1802.03788

Rodolfa, K. T., Lamba, H., & Ghani, R. (2021). Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, *3*(10), 896–904. https://doi.org/10.1038/s42256-021-00396-x

# Document Version Control

| Date / Version | Edited By | Changes |
|---|---|---|
| | | |
| | | |