NASSCOM®

RESPONSIBLE AI
ARCHITECT'S GUIDE

PRESCRIBING RESPONSIBLE AI BEST PRACTICES IMPLEMENTATION METHODS AND TOOLS FOR ENTERPRISE ADOPTION

Industry Partners

IBM Microsoft

CONTENTS

UNDERSTANDING HUMAN-CENTERED DESIGN

guidelines for the adoption of a human-centered approach to building AI systems

AI holds tremendous potential to augment human intelligence and productivity. But AI system designs must factor in human conditions whilst attempting to augment human intelligence with machine intelligence. A human-centered system design can bring in relevant checks and balances and ensure that the systems are being developed with adherence to the core principles of responsible AI that reduce the risk of potential harm to users of AI systems.

Guidelines for Implementation of Human-Centered Design

Microsoft's 18 Guidelines for Human-AI Interaction Design can be applied by systems designers, architects, and UX designers in systems design workflows at the following stages: initially, during interaction, when wrong, and over time. The guidelines aim to ensure that the system sets the right user expectations and responds to these expectations by giving due consideration to user context. Inaccurate user expectations might hurt user acceptance of the technology, compromise user safety, and paralyse the ability of users to leverage technology for good.

Illustration - 1

Consider a scenario with a semi-autonomous vehicle that is optimised for detecting moving objects. The driver of the vehicle puts the vehicle in self-driving mode and it hits a firetruck parked on the highway, causing the driver grievous injury. If the driver would have been made aware of the vehicle's limitation to detect moving objects only, and not the ones that were stationary, the driver could have been expected to be more cautious and the accident could have been averted.

Illustration - 2

Consider the case of a virtual agent developed to send people reminders about important and due tasks. The agent sends a buzzing reminder to a driver pacing on the highway, causing fatal distraction from road traffic. To avoid such scenarios, the designer of the virtual agent should ensure that it suspends all notifications in sensitive contexts such as driving to prevent harm.

RESPONSIBLE AI LIFECYCLE

compliance with responsible AI principles at each stage of the AI project value chain

The responsible AI lifecycle culminates from conscious adherence to the imperative of operationalising responsible AI principles at each stage of the AI project value chain — namely, envisioning and impact assessment, data collection and processing, prototyping, testing, deployment, building for production, deployment, and monitoring. This enables enterprises to deploy AI solutions that prioritise user trust and safety above all. Enterprises achieve this imperative using a range of technology and management tools and guidance, as part of their commitment to building and deploying responsible AI that prioritises user trust and safety.

Illustration: Operationalising the Principle of Explainability at Multiple Stages of an AI Model Lifecycle

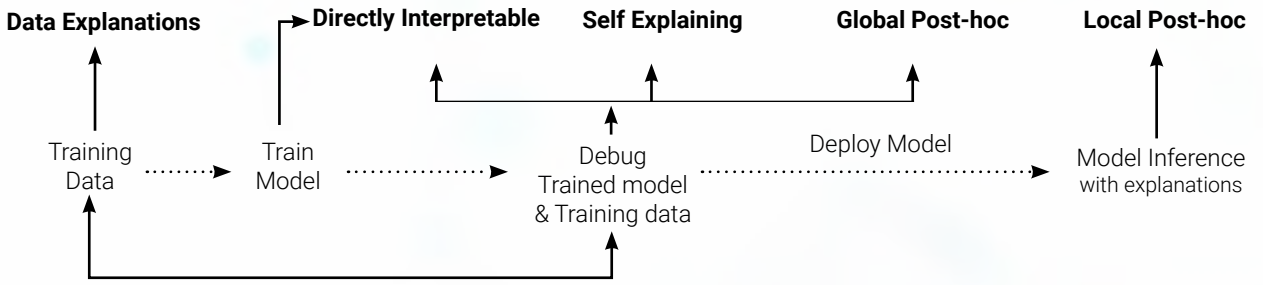
Being intrinsically black-box in nature, a common challenge with AI models is that they cannot meaningfully communicate their inner workings and predictions to their end users. The consequence of this opacity is a cutback in both user trust and the overall usefulness of AI systems. AI explanations, therefore, become critical for fostering trust amongst users by enabling them to understand the behaviour of the AI model. Furthermore, each stage of the AI lifecycle can involve one or more stakeholders who seek different types of explanations.

A multitude of AI explanation techniques are available open-source for use by developers today.

Figure - 1

The mapping of commonly used AI explanation techniques to their use in different stages of the life cycle of an AI model such as data collection, training, debugging and deployment can be seen in Figure 2. Some of these techniques and their use to various stakeholders are briefly explained below:

Figure - 2



Data Explainers

These explainers typically help explain a certain aspect of the data — for instance, commonly occurring samples in a class or certain corner cases or outliers. These are useful when the data is being collected and evaluated during training and debugging of the AI model in order to understand how the data impacts model behaviour and performance. Such explanations are generally used by AI system builders or data scientists.

Directly Interpretable Models

These explainers are used to explain intrinsically transparent models that are understandable by most stakeholders — for instance, a small decision tree or a linear model. Such explainers can be used to build transparency in models for regulated industries. They can also be used as global post-hoc explainers to explain a source blackbox AI model. These are most commonly used by data scientists and regulatory bodies who may look to audit models for bias or discrimination.

Global Post-hoc Explainers

These explainers are used to obtain insights about the overall behaviour of an AI model by training a transparent surrogate model. They are most commonly utilized by data scientists to understand and debug model behaviour, by domain experts to compare the model's behaviour with well-known rules or domain knowledge, and by regulatory bodies.

Local Post-hoc Explainers

These explainers are used to derive justifications about an individual instance or sample. They are generally useful for affected end users who may face the decision of an AI model, and by domain users such as doctors, judges, loan officers, and so on.

Example: AI Explanations for a Finance Model

Here, we discuss a credit approval AI use case and the expectations that various stakeholders may have with respect to AI explanations, and presents how these expectations can be achieved using the existing AI explainability techniques and tools.

Consider a scenario where a financial institution such as a bank uses an AI model to determine if a customer's loan application should be approved or not. This AI use-case would typically involve three stakeholders, namely, a data scientist who may have trained and deployed the model, a loan officer who uses the AI model to review loan applications, and a loan applicant.

The type of AI explanations that are relevant to each of these stakeholders is discussed here:

AI Explanations for a Data Scientist

A data scientist would be primarily interested in understanding the overall reasoning of the AI model and may seek assurance that the recommendations made by the model are reasonable in most cases to avert spurious correlations. In this setting, the data scientist may use a global post-hoc explainer, a directly interpretable model, or a data explainer.

AI Explanations for a Loan Officer

A loan officer would be interested in the justifications for the recommendations given by the AI model for different loan applications. One way for the loan officer to do this is by using a local post-hoc explainer to understand the model's reasoning behind the approval or rejection of a given application, enabling him to also communicate it to the loan applicant.

AI explanations for a Loan Applicant

A loan applicant may want to know the status of the loan application, and more importantly, how they could update their application to receive a favourable decision in the future. These expectations can be met by local post-hoc explainers that provide contrastive or counterfactual explanations.

ENVISIONING AND IMPACT ASSESSMENT

high-level guidelines and recommendations for impact assessment of a system in its early stages

The process of envisioning a responsible AI system should be underlined by a peoplefirst approach. When writing a product vision document describing product requirements and its intended use cases, it is critical to incorporate plans to systematically identify and document the probability, nature, and magnitude of potential
impact of the product and its use-cases on target users and the society at large.

This section provides high-level guidelines and recommendations for impact assessment of a system in its early stages. Early identification of potential harms from the system to both direct and indirect stakeholders could help assess and address future challenges around user adoption, and related ethical, regulatory, financial, and reputational risks. In some cases, this assessment may eventually lead to corporate decisions to defer
deployment of the system or hold back certain features in the system.

Early Identification of Potential Harms

Privacy

Privacy is paramount to building responsible AI systems. Personal data or personally identifiable information also carries legal protections that must be respected. Privacy violations could lead to lack of user trust, negatively affect user adoption, and carry financial penalties for enterprises under applicable domestic, regional, and international rules and regulations (e.g., EU General Data Protection Regulation).

Creating a workflow for an early stage privacy review must include the following considerations:

- Data flow in the system
- How is data collected
- Where the data will reside and for how long (see Right To Be Forgotten giving users the right to remove their data from the system so that it remains untraceable for third parties and the right to social integration without being perpetually stigmatised by specific incidents in their past)
- Drawing the compliance boundaries for secure hardware or encrypted data in the system

Fairness

There are different types of fairness-related harms (e.g., allocation harms, qualityof- service harms, and representation harms). It is crucial to note that a system could be held legally liable for both disparate treatment (intended or explicitdiscrimination in the process) and disparate impact (unintended or implicit discrimination in the outcome).

- **Allocation harms** occur when AI systems are used to allocate opportunities or resources in ways that could otherwise have significant negative impact on the lives and livelihood of people – for example, an AI system used in recruitment or loan adjudication that shows bias against a gender.
- **Quality-of-service harms** occur when a system that works well for one person does not work the same for another, even if it does not allocate or explicitly withhold any opportunities or resources – for example, a facial recognition system that has a higher error rate for darker skinned people or a voice assistant that has a higher error rate for a certain accent.
- **Representation harms** occur when AI systems propagate discriminatory stereotypes that present members of certain demographic groups in a derogatory manner or under- or over-represent their participation in socially significant activities – for example, auto-tagging that applies racist or demeaning labels to photos of people.

It is important to note that these harms are not mutually exclusive and may cooccur. To assess these harms, it is imperative to identify sensitive demographic groups that may be harmed in the intended use cases. Moreover, any quantitative or algorithmic assessment of these harms must also protect all personal data used in this assessment. If complete mitigation of the harms is not pragmatic, a redressal mechanism must be incorporated in the product vision. Interpretability

of the system is useful for user redressal, confronting legal challenges, system debugging, and timely updates to the system.

Regulatory Compliance, Responsible AI Dashboards and Audits

Responsible AI envisioning must factor in both benefits and harms from intended as well as unintended use cases, and explicitly discourage or defer use-cases where harms outweigh the benefits.

Example: Facial Recognition Technology

In the past, some concerns about unregulated use of facial recognition technology were reported. Firstly, the outcomes of facial recognition algorithms showed racial and gender bias and violated anti-discrimination laws. Secondly, their widespread use may have intruded into people's privacy. Thirdly, their use for mass surveillance could have encroached on democratic freedom. It was, therefore, important for the industry to defer the deployment of facial recognition technology and work in tandem with governments towards regulating the inherent risks in this technology.

Responsible AI Licenses (RAIL) provide a way for developers to safeguard their AI source code against irresponsible and harmful applications. These licenses include clauses to restrict applications of a code for potentially harmful applications of AI.

The risks associated with an AI system must be communicated in a transparent manner, both internally and externally. It is vital to build responsible AI dashboards internally to quantify and communicate different types of risks to various stakeholders. Senior leadership, legal teams, social scientists, and data scientists need to work together to develop dashboards that accurately reflect responsible AI metrics and regulatory compliance. Internal audits as well as third-party audits using these dashboards at regular intervals are important because AI systems are in a continual state of evolvement when they learn from newer data. Externally, a responsible AI system must clearly communicate the privacy risks to its users regarding the collection of their personal data, and the ethical and societal risks in the output of the system.

Inclusive Development

For inclusive development, a diverse team could offer mixed viewpoints and potential concerns for consideration. To further mitigate the risk of bias-related harms, it is important to identify and include diverse datasets and sensitive demographic groups in the training, testing, and auditing of AI systems. If prior identification of sensitive groups is not possible due to privacy constraints, a workaround would be to collect diverse and inclusive training and testing data along with redressal mechanisms for any complaints about model performance, bias,

and transparency. Incorporating the takeaways as inputs into the product vision to retrain the model with phased deployment could help assess its robustness and mitigate large-scale amplification of harms.

DATA COLLECTION AND PROCESSING

primer on data bias and its common types, and best practices for responsible data collection and processing

AI systems are generally expected to make fair and faster decisions. The performance of an AI system significantly depends on the quality of training and evaluation data. Data, in its raw form, could be loaded with social biases. Poor data collection practices could allow these biases to flow into the AI engineering pipeline.This section provides quick insights into the problem of data bias and its most common types precluding responsible AI design and development, with recommendations and best practices for countermeasures for adoption across different phases of data collection and processing.

Understanding Data Bias

A dataset is considered biased when certain elements are heavily weighted compared to others ultimately skewing the dataset. This dataset foregoes the capability ofrepresenting the entire population accurately and can result in a variety of unexpected outcomes and analytical errors that lower the expected accuracy levels.Furthermore, it has a denigrating impact owing to its impartial nature eventually veering towards a few specific elements whilst exhibiting inequity towards the others. Therefore, the baseline dataset must try to minimize all bias to ensure a closer-to-true-world representation of the facts and features. Data bias can occur across disparate areas ranging from human reporting and selection criteria to algorithmic and interpretation capabilities. It is important to understand and identify this bias before it can be remedied. The most common types of biases are explained below as a starting point:

Sample Bias

Consider an AI model trained to identify the gender of people. Imagine a scenario where this model is trained on a dataset consisting of images of only a specific ethnicity and a single gender. This model will be prone to considerably lower accuracy with the people of other genders and ethnicities. This is referred to as sample bias or selection bias where your dataset is not inclusive of all the potential users or stakeholders.

Exclusion Bias

Exclusion bias occurs when removal of features from a dataset to reduce noise results in elimination of features that are otherwise critical to providing internal linkage that may go unnoticed by a human. For example, while predicting consumer spending patterns from a specific state, we may choose to exclude the city location given that we are surveying the state. However, eliminating this information might introduce the exclusion bias preventing successful differentiation between rural and urban spending.

Measurement Bias

Measurement bias occurs when the data collected for training differs from that collected in the real world, or when faulty measurements result in data distortion. A good example of this bias occurs in image recognition datasets where the training data is collected with a specific type of camera, but the production data is procured with a different camera. Measurement bias can also occur due to inconsistent annotation during the data labelling stage of a project.

Observer Bias

Observer bias or confirmation bias is the effect of seeing what you expect to see or want to see in data. This occurs when researchers enter a project with subjective thoughts about their study, either consciously or unconsciously. We can also see this occur when researchers let their subjective thoughts control their labelling habits that result in inaccurate data.

Recall Bias

Recall bias is a kind of measurement bias and is common at the data labelling stage of a project. Recall bias occurs when you label similar types of data inconsistently that results in lower accuracy. For example, consider a system that is being designed to label image of phones as damaged,

partially damaged, or undamaged. If one labels an image as damaged, but a similar image is tagged as partially damaged, the data is prone to inconsistency.

Racial Bias

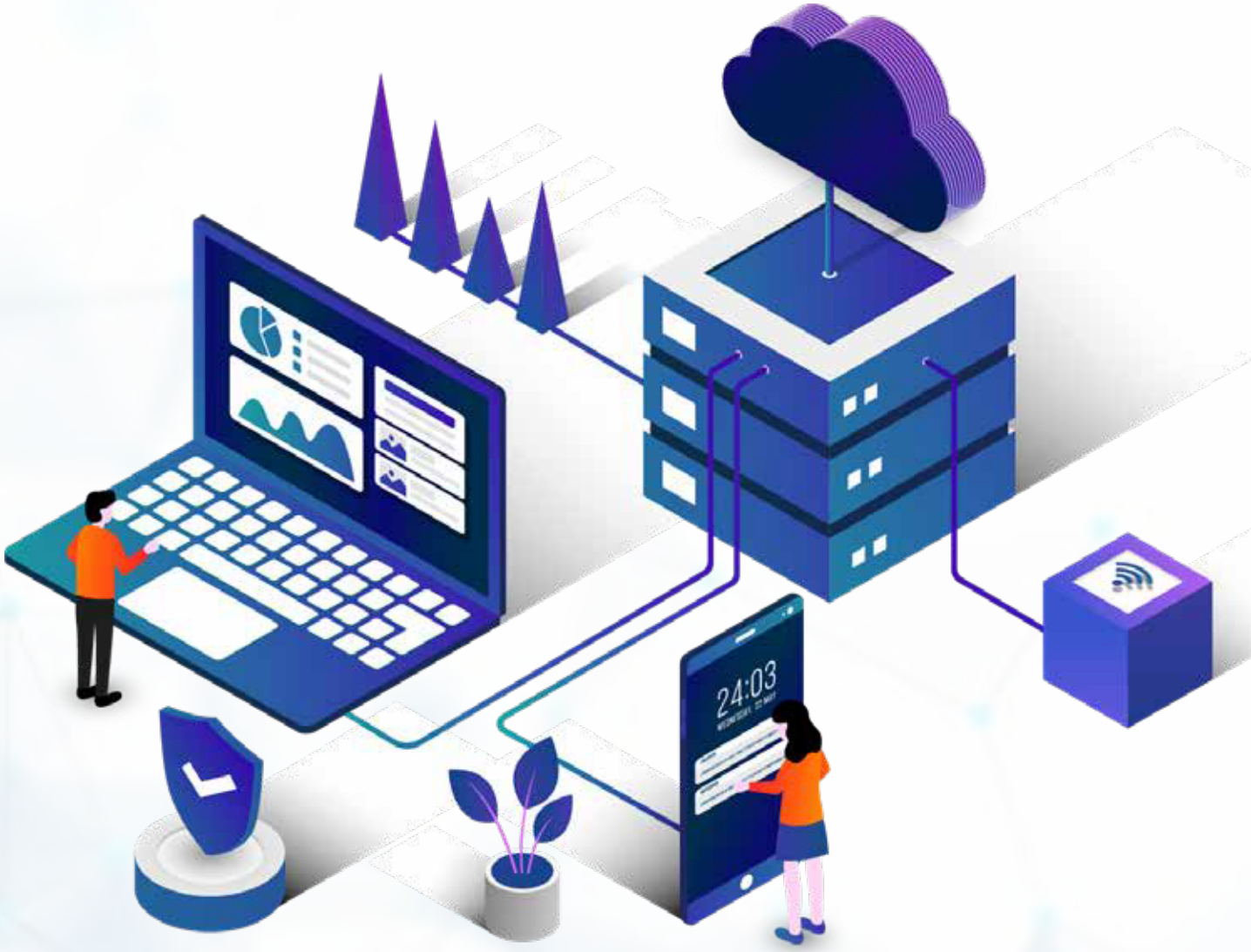
Racial bias occurs when data skews in favour of a specific demography. This can be seen in facial recognition and automatic speech recognition technology which fails to recognize people of colour as accurately as it recognizes Caucasians. This can be an unsupervised model displaying different behaviours for disparate sects of the society. For example, an online promotion model shows different prices to different people based on their ethnicity.

Association Bias

Association bias occurs when the data for an ML model reinforces and/or multiplies a cultural bias. Your dataset may have a collection of jobs in which all men are doctors, and all women are nurses which does not mean that women cannot be doctors and men nurses. However, as far as your ML model is concerned, female doctors and male nurses do not exist. Association bias is best known for creating gender bias.

Harms from Biased Datasets

AI systems built on biased datasets can result



in allocation harms, quality-ofservice harms, and representation harms, as explained in the Envisioning and Impact Assessment section of this guide.

Data Sourcing

Sourcing is paramount to any dataset creation that is subsequently used to train an AI/ ML model. It dictates system performance and response to the inputs received to a very high degree. As a general practice, a well-performing model trained on a representative dataset is assumed to be a good generalisation of the target population. However, this is also the first trigger that can introduce bias in the system from the start.

A series of proposed questions allow a detailed analysis of the source and ensure that the biases are identified and eradicated, wherever possible:

- **How will my dataset look? –** Identifying the key elements required in the model along with the learning parameters and their corresponding importance
- **Who is my target audience? –** Knowing the composition of the target audience is an important factor — the chosen sampled data should be representative of the entire population to a high degree.
- **How is my target variable effected? –** Understanding the target variable and the possible linkages of the data features
- **Have I collected all the data? –** In a given sample of data, the focus should be on identifying all the potential attributes that contribute positively or negatively towards the target variable. The sample data must be able to explain the traits in target variable to a high degree.
- **Is my data not weighted to a particular section of the population? –** Datasets skewed in favour of a particular section of the representative population will a cause selection bias.
- **Am I legally compliant to collect the data? –** There are several compliances mandated by EU General Data Protection Regulation or local laws pertaining to the usage of personally identifiable information that may need to be reviewed, especiallywhen dealing with sensitive information.
- **What is the agreement for processing of data? –** An agreement detailing the possible uses of collected data is a legal obligation that every enterprise should have while collecting data from individuals
- **Is the data collected the bare minimum required? –** The focus should always be collecting the bare minimum data features that can successfully train the model to work as intended. This will also help avoid unauthorized usage of data by third party, thereby minimizing data theft risks.
- **Is unique data being collected? –** Under the several regulations like EU General Data Protection Regulation, there are certain data sets that are termed as 'unique or special'. Some examples are political beliefs, sexual orientation, religious beliefs, memberships, and so on. Collection of this type of data needs to come with an assurance that it will neither be reproduced nor result in unintended use without consent.
- **Is the life cycle and subsequent purging of collected data being tracked? –** Reviewing the planned expiration of the dataset ensures that the AI system remains up to date and the collected information is safeguarded against unintended or unauthorized usage.

Best Practices

- **Using user-friendly research instruments** (such as questionnaires, interviews, experiments, or observations) that detail unambiguous and easy to use directions for respondents to ensure better respondent responses and consistency in collection and recording.

- **Training of research personnel** who are involved in the data collection process and supplying information of the end goals, target groups, and the standard process to be followed. Trained research personnel are usually able to provide uniform sampling that is a good general approximation of the target demography.
- **Evaluation of data for recording errors** that are likely to occur frequently owing to human errors, systemic issues, and so on. Using a robust backup system to evaluate any recording errors and minimizes any unwarranted skewing of the system.
- **Gathering of data from multiple sources** by collecting data samples from various sources to ensure a balanced sample creation and diversifying the risks.
- **Verifying and validating the data** using an exploratory analysis is recommended to review outliers and missing information along with crossreferencing from other available data sources to confirm the veracity of the collected inputs.
- **Checks for alternative explanations** that help in identifying and accounting for alternative reasons for the way in which a particular data sample may have been collected. This also averts confirmation bias.
- **Peer-reviewing the datasets** and collected samples to discover issues that could have been missed and identify gaps in the original data source that may need to be addressed.

Dataset Creation

Once the relevant sources of data are identified along with key attributes and the target variable/ audience, the next step is the dataset creation. The data from heterogenous sources will eventually be extracted into an algorithm-friendly tabular format. There are multiple steps involved in curating the datasets and achieving homogeneity to enable consumption and training.

The methodology can be broadly classified in 3 stages:

Pre-processing

- Aims at reducing bias in the original data before the model is trained and preventing the different harms that could emanate from a resulting system.
- While gathering data from disparate sources, it is important to carefully choose the attributes that can explain your target variable and ensure accounting for exclusion bias that can occur to loss of certain features during data munging.
- In certain scenarios, there may be a need to label or annotate data points. This increases the occurrence of recall bias within the dataset. It happens due to asymmetrical labelling in similar records of data and broad ranges.
- The labelling should be kept as close and sensitive as possible to the real-world descriptions along with a mechanism to update labels as and when an anomaly in the underlying source is observed.
- There are instances where there is a need to group certain data points into buckets for classification. Bias occurs when one group is used (often one's own group) as the standard against which others are evaluated. For example, usage of 'normal' may prompt readers to make the comparison with 'abnormal', thus stigmatizing some model outputs with differences.

Best Practices

- Avoiding labels that single out a particular section of the data and instead generalise the labels into subtle types
- Anticipating exclusion, association, and observer bias during preprocessing
- Including all the data points in the sample set to represent the entire population at this stage

In-processing

- In-processing methods tackle bias during model training. The algorithm will analyze the sample and learn from the labeled facts. The algorithm will then mirror its learning on the test dataset.
- If the judgment applied during sourcing and pre-processing comprises any biases, they will get mirrored and sometimes amplified. The skew in the algorithms stems from either inaccurate representation of the population or imbalanced labelling.
- Multiple methods can be used for information re-balancing such as performing preferential sampling and reweighing (assigning a class specific or dynamic weight) so that the algorithm balances the model fit and avoids both over-fitting as well as under-fitting.

Best Practices

- Accounting for group-dependent label imbalances
- Selecting algorithms that minimise the bias during processing of the data examples - Discrimination Aware Ensembles
- Balancing the model to avoid under or over fitting

Post-processing

- Post-processing reduces bias by calibrating model predictions. Owing to mathematical or analytical errors, the model may present only a specific type of biased output.
- The output is governed by several dynamic data attributes such as timesensitive features, regulatory changes, demography changes and so on, representing the changing realities over time. Hence, recurring updates to the dataset is a quintessential process reviewed at this stage to avoid any influx of bias as a result of outdated data features.

Best Practices

- Performing dataset updates periodically
- Employing pre-processing best practices when the updates are deemed significant

Persisting Data

The collected data needs to be persisted for its usage in the next stage of the AI pipeline. But since the data is usually collected from disparate sources, it may contain sensitive user information and additionally come with several compliance restrictions like GDPR. Therefore, it is critical to select the appropriate data destinations within all the associated constraints. It is also recommended to employ cutting-edge 'Encryption at Rest' mechanisms to prevent inadvertent leakages.

Usage and Distribution implications

Data is considered the new oil with the ability to power multiple avocations apart from its original intended purpose. This can open up new possibilities to make improvements in the quality of life, but with the significant risk of being misused with real-world consequences and serious harm. Especially, when it comes to PII, lack of adequate privacy safeguards may enable AI systems to wholly record and analyze an individual's personal data without their consent resulting in serious repercussions.

For example, consider an online marketing company that collects user demographics and ethnic data for marketing several daily use products. The primary purpose of this data is to generate recommendations to boost up-sell and cross-sell opportunities. This same information can also be utilized to analyze spend patterns and buying capacities of specific groups based on age, location, gender, and so on. Once realized, this additional information has application in various other industries such as hospitality and F&B and can be potentially misused to target a specific group (higher interest loans, preferential pricing, and so on).

Security risks in AI systems arise from its heavy reliance on data and design and deployment environments. Some of these attacks are unique to Machine Learning systems and affects different parts of the Machine Learning development cycle. Adversarial machine learning attacks are designed to leverage vulnerabilities in the Machine Learning model with potentially harmful real-world consequences.

Data Documentation

The documentation journey of data from its collection to serving as training data for an AI system mostly takes place in an ad-hoc manner and is barely documented. As a result, developers have little insight into the quality of collected data and its readiness for building an AI model. They lose track of transformations that have been applied to the data since the acquisition stage and end up spending numerous iterations to explore such properties. Focusing on standardised documentation to accompany AI assets is a fairly recent phenomenon and is critical for explaining the long-term impacts of the AI system. Recent research proposals like Datasheets, Dataset Nutrition Label, and Data Readiness Report highlight these standardisation efforts for various AI assets in a systematic manner. It is recommended that each dataset should accompany a separate shareable document that could certify the baseline quality of ingested data and provide a record of operations and remediations performed on the data. This artefact could become a onestop lookup for understanding data quality and readiness analysis. These efforts are also aimed at benchmarking the key characteristics of data to empower auditing and enable informed reuse.

It is recommended that the following aspects be captured in the document:

Data Origin

This section should include details about who created the dataset and any funding information such as associated grants. Capturing the origin is very critical as it can be useful to the governance officer at a later stage.

Motivation

This section should include details of the specific task in mind when the dataset was collected. This could help identify any potential gaps in the dataset for future goals.

Composition

This section should capture the number of instances in the dataset and the constituents of each data instance. It should also highlight if this dataset includes any sensitive information or if it can be used to construct individual identity. It will help data scientists to make informed decisions about the applicability of the dataset.

Collection Process

The timeframe and the entities involved in the collection process should be captured. These details are crucial as they could help the data scientist understand the scope of the dataset and how well it is expected to generalise the target population.

Raw Data Location

The collected raw dataset should be persisted before any further processing. The original data could help the data scientist repeat the preprocessing of the dataset from a fresh perspective and may also help trace back any issues that originate at later stages. The data may be stored in a central location or distributed across various cloud locations within the compliance boundaries. This section should contain the storage location details and recommend any security measures to be used to protect the persisted dataset from accidental leakages.

Baseline Data Profile

This section should capture the basic characteristics and statistical properties of the collected dataset. For instance, it could include details about the amount of missing information. This would help data stewards and scientists to understand the raw dataset.

Baseline Quality Information

The collected dataset should be tested against several quality measures like class imbalance, inconsistency, redundancy, and so on to reflect the quality of the raw dataset.

Pre-processing

This section should include the details of any pre-processing, cleaning, or labelling activities executed on the raw dataset. It should also have pointers to the algorithms and the implementations used to process the dataset.

Updated Data Profile

The raw data profile may undergo changes once remediations or transformations are carried out to improve the data quality or remove anomalies. All such changes must be documented, as this section serves to provide the relevant updates to the data profile.

Updated Quality Informatio

After checking the data quality on raw data, many transformations or remediations are applied to improve the data readiness. This is typically validated by re-running the quality analysis on the updated dataset and ensuring that data modifications have been effective.

Lineage of Operations

Various personas interact with data in different ways for a variety of reasons. This section will provide a detailed documentation of all the data transformations and operations performed by them in chronological order, along with timestamps and other relevant details. This digital trail will provide a comprehensive record of how the data has evolved due to interactions with humans in the loop.

Data Governance

This section should explicitly mention any policy or legal constraints that restrict the accessibility of the dataset.

Summary of Quality and Readiness Assessment

The quality and readiness of the dataset should be summarised in this section as an executive summary for a quick assessment of the dataset.

PROTOTYPING

processes and tools to design a responsible system prototype

The following questions need to be considered for designing a responsible system prototype:

- How do your system design and implementation decisions affect your users?
- Are you designing your system to be robust to errors?
- Are your system decisions inclusive and fair?
- How does your data affect your system?
- How do your modelling decisions affect your system?
- Have you carefully considered and tested the design choices for your system?
- Is your inclusive and transparent UI or application co-designed with machine learning?
- Have you tested your system with diverse user groups?

This section provides details about building responsible AI dashboards, documentation of data and AI systems, and pre/in/post-processing algorithms for mitigation of fairnessrelated harms, with relevant examples, case studies, and references to responsible AI toolkits.

Responsible AI Dashboards and Metrics

The first step in the prototyping of a responsible AI system is to align all the stakeholders with the regulatory, legal, ethical, and societal concerns, and reach a consensus on what responsible AI metrics to track. The data science and legal teams must collaborate with social scientists and ethnographers to define privacy, robustness, fairness, explainability constraints, quantifiable metrics for the system, and its intended use cases.

The metrics are available in open-source toolboxes (such as fairlearn, interpret.ml, AIF360, AIX360). Also, a combination of these metrics or customized responsible AI metrics for specific use cases can be used. A good responsible AI dashboard must clearly capture the regulatory compliance of a model as closely as possible and raise a flag when the model violates any regulatory compliance requirement. Tracking responsible AI dashboards and their internal audit at regular intervals becomes critical as models continuously evolve with newer data. This step can be incorporated and executed along with the regular efficiency and performance trackers. To build responsible AI dashboards, it is important to first understand the data collection process, check the feasibility of building such a dashboard, and whether it requires any additional or better-quality data and ground truth labels.

Documentation of Data and AI Systems

Datasheets help understand how individual data and sensitive or protected attributes are collected, screened, labelled, stored, and used in the model pipeline. Datasheets also help document dataset characteristics and limitations of the pre-deployment data (training, validation data) and post-deployment data (test data) and any anomalies between them.

On the other hand, a **Data Readiness Report** provides a more comprehensive and holistic view of data quality issues, the remediations applied, and related explanations. It also maintains the lineage of data assessment operations and the role of various personas in a collaborative data preparation environment. The Data Readiness Report is a one-stop lookup for understanding the basic profile, data quality, and readiness analysis including the lineage of transformations applied.

Documentation of datasets can also help understand the features that are being used and determine if they are causal and important for the model's performance. They also verify if there is a high co-relation with sensitive or protected attributes such as the pin code or location as a proxy for membership of an underrepresented community. It is important for a responsible AI system to ensure that sensitive or protected attributes are used in a privacy- and fairness-compliant manner.

Model cards record the capabilities and limitations of models. They also disclose the context in which models are intended to be used along with the details of the performance evaluation procedures.

Factsheets focus on the final AI service as a distinct concept from a single pre-trained machine learning model or dataset. Factsheets are intended to include sections on all relevant attributes of an AI service such as the intended use, performance, safety, and security. They also list how the AI service was created, trained, and deployed. In addition, the scenarios it was tested on, response to untested scenarios, and guidelines for the at tasks they should and should not be used for, and any ethical concerns of their use. Factsheets, therefore, help prevents over-generalisation and unintended use of AI services by solidly grounding them with metrics and usage scenarios. Documenting the capabilities, limitations, and evaluation benchmarks of an AI service can help target the violations, if any, of responsible AI metrics and makes it easier to debug the AI pipeline.

Datasheets, Data Readiness Report, Model cards, and FactSheets together complete the AI documentation pipeline and bolster the trust and reusability of the data and AI service. They are helpful in marking the compliance boundaries and compliance framework clearly for model training, inference, and impact assessment.

Co-design UI and ML for Transparency, Redressal, and Fine-tuning

Careful consideration of design and testing decisions can be helpful in making a system transparent and inclusive. For this, an inclusive and transparent UI or application should be co-designed along with machine learning. In addition, building redressal mechanisms for any user complaints about model performance, bias, transparency, and methods to incorporate them as inputs to retrain or fine-tune the model periodically is documented.

Mitigation of Harms

A model must be tested with diverse user groups for robustness, fairness, and inclusivity. Understanding the AI security risk assessment as well as fairness and transparency is crucial to verify if the responsible AI dashboard for a model is in compliance with the regulations and internal responsible AI standards developed by an enterprise. To mitigate any fairness related harms, preprocessing, in-processing, and post-processing algorithms should be used. Local, global, attribute-based and causal explanations can also be used to assess fairness and transparency. Since various responsible AI metrics have an inherent trade-off, they must be paid careful attention – in terms of whether they are acceptable under the prevailing regulations or responsible AI standards and understand why some of these violations cannot be mitigated.

Dogfooding, Ring-testing and Post-deployment Strategy

Dogfooding, ring-testing, and post-deployment strategies need to be developed to avoid large-scale amplification of harms. In a phased deployment, periodic data collection and assessing performance and reliability deviations occurring due to data shifts are crucial. This is because a model that is trained, validated, and assessed to be compliant with a responsible AI dashboard must remain compliant in its entire life cycle. Models should incorporate mechanisms to flag data shifts and deviations in the responsible AI dashboard in a timely manner so that mitigation strategies can be applied against any large-scale amplification of harms.


TESTING
methods and techniques for model testing to ensure compliance with responsible AI principles
MacBook Air

This section provides detailed information on three types of tests that should be performed on the output of AI models from the responsible AI standpoint. Tests are executed for the following attributes:

- Behaviour - the model should perform the expected task
- Fairness - the predictions should not be biased against minority communities in the real world
- Adversarial – the model should be robust and withstand known adversarial attacks.

Data engineers can ensure that the data made available for consumption from standalone applications, data lakes, data mesh, or data fabrics pass certain basic tests for data quality and flag potentially unrepresentative data. They must maintain the provenance and lineage of the data and verify if it is collected with the informed consent of data principals. Provenance helps downstream data users to determine any ethical issues while collecting the data. Lineage helps with governance during the lifetime of the data. Data scientists and end user representatives must run the relevant tests mentioned in this section and verify the results before further use of the data and models. Project or product managers must ensure that the stakeholders have a test plan that includes tests for responsible AI and ensure compliance with user governance in place.

Blackbox Testing

Test driven development and other paradigms in software development have emphasised the need to have a testing strategy in place before development. From the responsible AI point of view, finalising the test scenarios in 'black box' mode has several advantages. Potential bugs could arise from several parts of the AI development life cycle. The input data for training the model could introduce bias due to a coding error in annotation or transformation. The model may learn personally identifiable information (PII) and violate user privacy, even if it not biased. For these reasons, writing tests in the 'blackbox' mode helps to holistically view the model from an ethics and fairness standpoint holistically before determining where to implement the responsible AI solutions.

Checklists for Behavioural Testing

Many natural language processing (NLP) tasks such as sentiment classification, machine translation, reading comprehension have direct applications in products used by people in their daily lives. However, there are other tasks such as named entity extraction, entity classification, and relation extraction that could be used to extract data. Applicationspecific ML models are trained and inferred on the extracted data. For example, a graph neural network model could be trained on data extracted from unstructured documents. The pre-processing tasks are typically performed in an NLP pipeline. Since these NLP pipelines can be used to extract personal data from unstructured documents, they need to be tested for any potential bias against minority communities. Recently, checklists have been proposed for behavioural testing of NLP models. The idea of testing the model for expected behaviour for a given input can be extended to test the entire set of NLP pipelines. This type of testing could be considered more in line with integration testing instead of unit testing of the individual models. For unit testing of models, specific input and output combinations that are created can be verified. For testing the entire pipeline, the end NLP application for which data is being extracted can be considered. However, the propagation of errors from NLP applications to data preprocessing tasks tend to be quite noisy.

Fairness Testing using Post-hoc Models

Fairness can be measured in terms of group and individual fairness. To evaluate group fairness in the generated data, metrics such as disparate impact could be used. For individual fairness, counterfactuals are generated by perturbing the input and checking if the predictions vary or not. More specific metrics such as the unfairness score can also be used to evaluate individual fairness.

Adversarial Testing for Robustness

For checking models against adversarial attacks and to evaluate their robustness, we can use libraries such as TextAttack. TextAttack and other similar tools perform transformations on the dataset with respect to certain constraints to produce new samples. Development teams can also use adversarial samples specific to their data and applications and the unit tests written during development can be used to create adversarial input.

Getting Started

After having decided that the tests for responsible AI metrics are a mandate, a typical question that enterprises and software development teams may have is how and where to get started. These teams may already be following test driven development, with checklists for behavioural testing and A/B testing frameworks in place.

An easy way to incorporate tests for fairness or other responsible AI metrics is to run the existing test suites but with modified input data. If a model for predicting loans exists, the same tests and metrics that were written from an accuracy point of view can be executed on a subset of the input with only a minority of applicants.

If there is a substantial difference in performance, it could indicate that the underlying model is trained on less diverse of biased data.

BUILDING FOR PRODUCTION

responsible ML toolkits as a superior alternative to DevOps for model development

System development and deployment processes have evolved into advanced developer operations or DevOps. The introduction of tools and platforms into workflows have been designed for Continuous Integration and Continuous Deployment (CI/CD). Continuous Integration implies that each time a code is checked in, it will trigger a set of automated unit, smoke, and integrated tests depending on the environment. Dev/test/preproduction/ production will ensure that the code being checked in is of acceptable quality, solves the functionality or provides bug fix and does not break or impact other parts of the system. Only a successful build is checked into the repository, otherwise is prone to failure.

A successful build triggers manual, semi-automated, or automated process for deploying the changes in the code. Today's systems have techniques to roll-back to previous versions, deploy across multiple environments, regions, and so on and provide for greater control over the quality of code, ability to upgrade, as well as correct any issues to mitigate complications. Most enterprise now have a fairly well-established process of DevOps that align with their deliverables and meet the required security and compliance requirements.

Integrating Model Development with DevOps

In most enterprises developing and deploying AI systems, the process of model development is still isolated and has a traditional wall between the data scientists who build the model and developers who integrate the model into the system's workflows. Some of the common challenges in this approach are as follows:

Model Development in Isolation

Models are typically developed by data scientists on historical data exported by a system or made available offline from other sources. Once the model is ready, the model is made available as a packaged file.

Figure 3

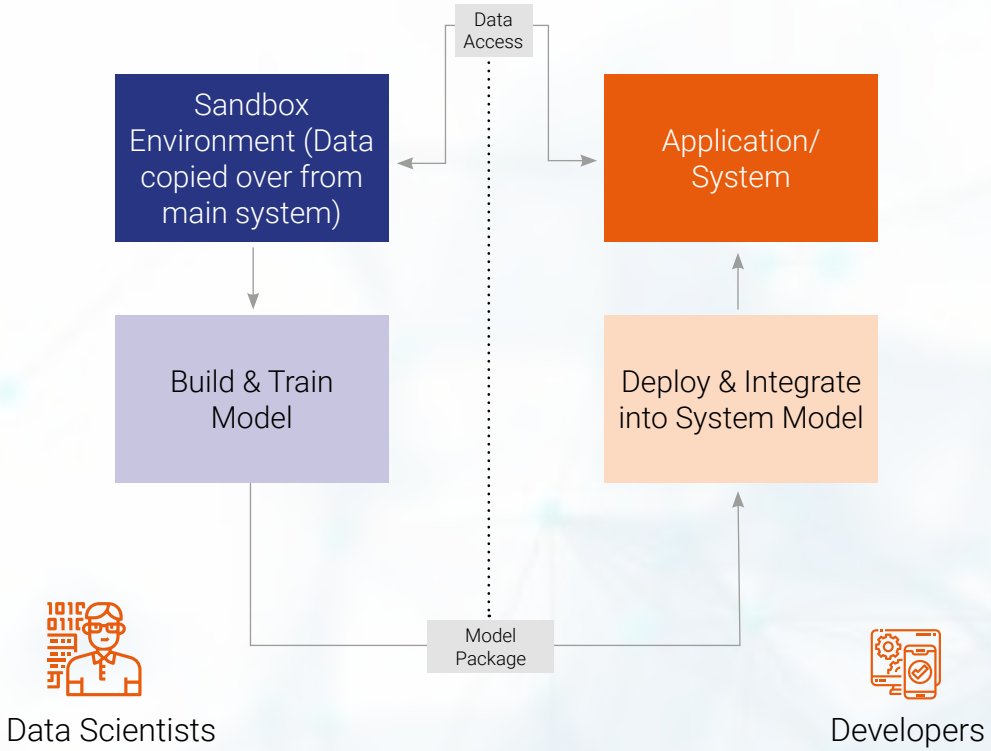


Figure 3

The processing steps and learnings from various tweaked parameters are not available to the system and more often than not remain with the specific data scientist working on it. This may also result in a lot of versioning issues.

Data Security Risks

Data is often provided in a sandbox environment which is typically a copy of the data from the main systems that may have the right level of access control, monitoring and triggering mechanisms, and audit logs. The sandbox environment exposes the data to risks of unauthorized access, privacy breaches, and lack of control.

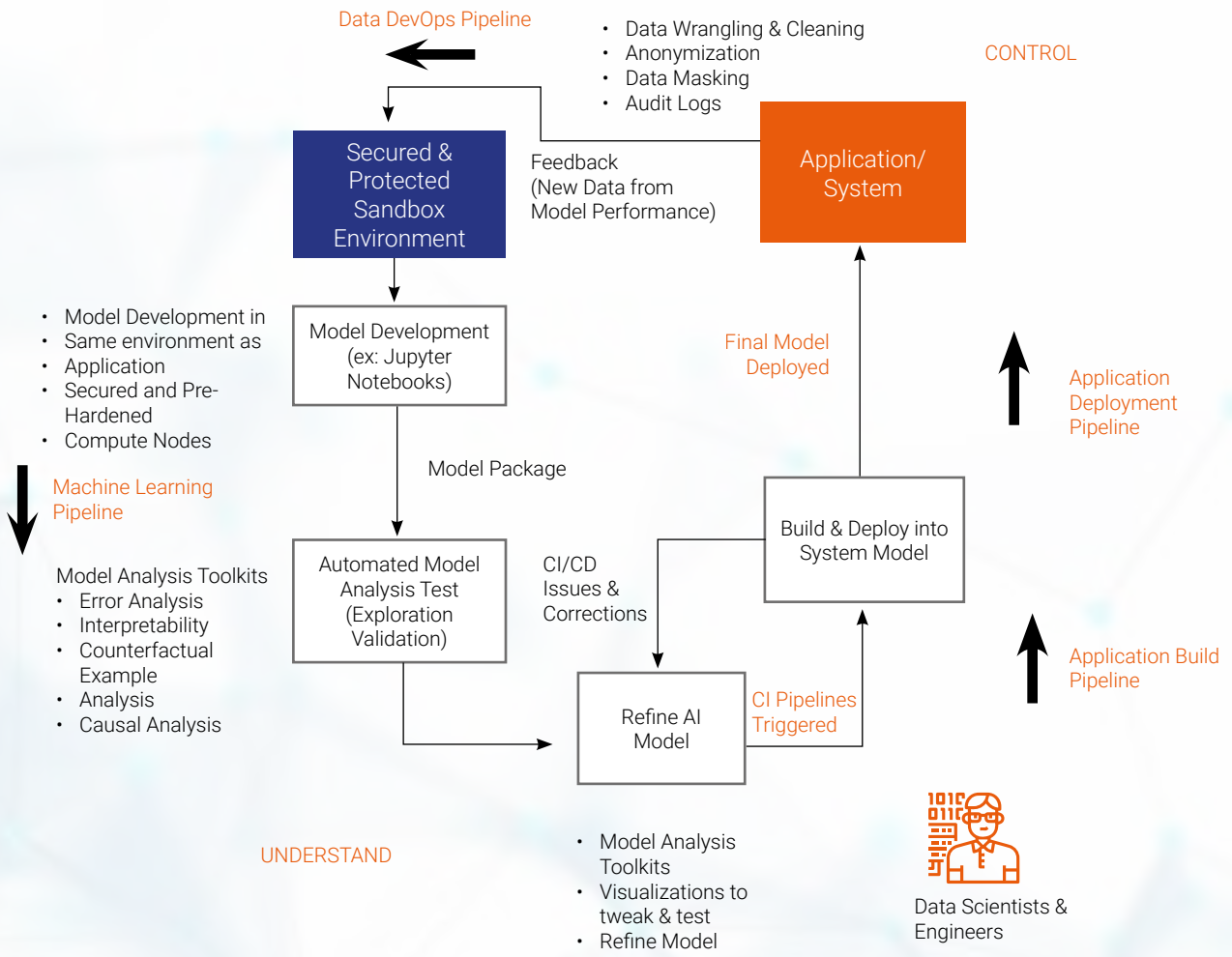
Delays in Feedback for Model Deployments

Since the process of development, refinement, and publishing of the model is different from where the model is integrated and used by the systems, there is a significantly delay in the time the feedback on the performance of the model is captured to when it is analyzed and used for model refinement.

Figure 3 depicts the disconnected process between model building and model deployment.

An effective Responsible ML Development warrants mitigation of these challenges as a prerequisite and require that the model development be integrated and made mainstream in development of the application. **Figure 4** below explains integrated development processes that can be used in deployment.

Figure 4



The processes depicted are not specific to a DevOps platform. The intent is to represent a generic process that can be implemented by any popular DevOps platform preferred by enterprises.

The Integrated DevOps processes can be better understood by exploring the various engineering processes that are involved in building an AI-driven system. They can be categorised as:

Data Engineering

Data pipelines that manage data acquisitions, processing, managing storage, and logging and monitoring

AI/ML Model Engineering

Model building, training, refining and publishing

Application Engineering

APIs, client and front-end development, integration with AI/ML models, and data systems.

Each of these engineering processes typically have their own DevOps processes and the CD pipelines should be designed to integrate these processes.

Figure 4 illustrates a typical integrated machine learning DevOps process. The sandbox environment is created as part of the data engineering processes,

where data can be copied or made available for specific access. The environment can be updated automatically with data being extracted from the live system, especially on performance of the model. For example, if the deployed model identifies a user's action based on voice input, the monitoring data at each time the model ran but failed or required a manual input could be further used for refining the model towards enhanced accuracy and eliminating potential biases against untested scenarios.

The model developers typically use tools such Jupyter notebooks which can be deployed on their personal machines or servers where they can be made available to data scientists for model building and development. The machine learning pipeline could potentially include these tools to create, train, and test the models. The process of publishing the model and integrating with the application through an API can also be used to push upgrades to the model and follow the traditional CD process.

Additionally, the machine learning pipeline can further include automated tools that enable Responsible ML development as explained in the next section.

Building AI with Responsible ML

The three pillars of using Responsible ML are understand, control and protect.

01 Understanding model behaviour is also inclusive of the ability to assess and mitigate unfairness

02 Control requires that the process should be repeatable, reliable, and hold stakeholders accountable

03 Protection is about the access to sensitive data, masking and encrypting sensitive fields, yet enabling the model to be trained while supporting all data sensitivity controls

Figure 5

There are several explainability/interpretability and fairness toolkits. A machine learning DevOps pipeline typically enables the incorporation of these toolkits into development of the models so that the process of re-training, validating, and improving models are seamlessly integrated with the rest of the application development processes.

Controls can be built into the pipelines to ensure that every model that is being developed follows the process of evaluating the model against the toolkit.

The Responsible AI Toolbox is an open-source framework for helping data scientists and ML developers build ML-powered products that are responsible and reliable. The toolbox integrates together ideas and technologies from several open-source toolkits in the areas of error analysis which identify cohorts of data with a higher error rate than the overall benchmark.

These discrepancies may occur when the system or model underperforms for specific demographic groups or infrequently observed input conditions in the training data. Model interpretability powered by InterpretML explains blackbox models that help users understand the global behaviour of their model or the reasons behind individual predictions.

Counterfactual Example Analysis powered by InterpretML DiCE shows feature perturbed versions of the same datapoint which would have received a different prediction outcome. For example, Taylor's loan has been rejected by the model. But they would have received the loan if their income was higher by 10,000 USD.

Causal Analysis powered by EconML focuses on answering 'what-If-style' questions to apply to data-driven decision-making – how would the revenue be affected if a corporation pursues a new pricing strategy? Would a new medication improve a patient's condition with all else equal?

For example, a DevOps pipeline can be built to run these toolkits each time a code/ model version is checked in.

In conclusion, it is important to consider streamlining model development with application development practices while using AI in production grade systems. Enterprises must revisiting their existing DevOps processes to incorporate ML pipelines that enable developing, analysing, training, and deployments. Additionally, they should consider templatising pipeline components that incorporate various toolkits to ensure that model developers, data scientists and data engineers use it to incorporate Responsible ML best practices.

DEPLOYMENT

tools and best practices for responsible deployment of the system

Deployment helps materialise the industrious efforts made in the previous stages by bringing the AI system to life. Being a crucial stage in the AI life cycle, deployment often incurs an enormous technical debt if not done responsibly. In the long-term, the deployment must be carried out responsibly with feedback channels and escalation processes in place. To ensure successful and responsible deployment of the system, reviewing of the following traits is recommended:

Robust Deployment

There have been instances when a production system has failed to handle any sudden peaks in the workload, leading to substantial financial loss and damaged reputation. Hence, it is critical to understand the production requirements and have a detailed plan in place to handle any anomalies in the expected workload. Therefore, deployment must be robust to ensure stable performance of the model.

Secure Deployment

The security of an AI system is another essential dimension. Research has shown that a deployed AI system is susceptible to several adversarial threats including evasion, poisoning, extraction, and serviceability. An adversary forces a deployed model to be misclassified in an evasion attack by making imperceptible disturbances to the input samples. In a poisoning attack, adversaries attempt to deliberately influence the training data with the aim of manipulating the outcomes when models are periodically updated with new data. In an extraction attack, adversaries abuse a model's query API and launch a series of intelligent queries to steal the hosted model, thus averting future query payments. In the serviceability attack, the adversary tries to bring the system down by using a Denial-of-Service (DoS) attack. Hence, security is a significant factor to be considered for reliable and successful deployment. Both robustness and security attributes help sustain an AI system to prevail in extreme and adversarial environments.

Continuous Deployment

Due to clenching competition, the size of the application, and time-critical demands, most of the industries adopt Continuous Integration and Continuous Deployment (CI/ CD) pipelines. It is recommended to design these channels strategically. This section provides a detailed insight into the related pitfalls and best practices.

As the AI systems are used to make sensitive decisions such as lending or VISA approvals, it is crucial to ensure that the decisions made by the models remain fair. In long-term deployment, the deployed system may quickly become obsolete due to the evolving input data distribution. More importantly, the model may become biased towards a particular group. Consequently, to make the deployed AI system reliable, it needs to be monitored and fixed. In order to ensure enduring profits, a deployed system must be continuously improved and have feedback channels and escalation processes in place. In order to have a trustworthy deployment in place and extract the maximum out of the AI service, every enterprise should prepare a deployment checklist. At the bare minimum, the following aspects must be validated before deploying to a production environment:

- **Requirements Check**
  The user and business requirements must be re-accessed along with the use cases identified in the envisioning stage to ensure that the AI systems achieve the planned objectives.
- **Evaluation in Production Environment**
  The AI system must be tested in the pre-production environment to handle the expected workload and for any biases before it can be put in production systems.
- **Frequency of Model Access**
  The production environment must be configured appropriately to handle the varying frequency of model access using optimal resources.
- **Batch or Single Instance**
  Specific optimisations must be done to handle batch and individual accesses.
- **Load or Number of Users or Scale**
  The production system must be able to predict the load and scale up the deployment appropriately to keep the AI service active.

- **Latency Requirements of the User**
  The deployed system is only helpful when it can serve the user request within an acceptable latency.
- **Handle Unforeseen Issues**
  Catastrophic and unseen scenarios must be pre-meditated to ensure minimal downtime.
- **Monitoring, Maintenance and Upgrade Plan**
  The deployed system must be continuously monitored, and any assumptions made in the requirements must be validated to ensure reliable system behaviour.
- **Feedback Channel and Escalation Plan**
  The deployment must also offer a feedback channel and an escalation plan. These channels can help to quickly identify invalid assumptions, harmful biases, unacceptable system behaviour and assist in quick fixes without causing a significant business loss.
- **Auditable and automated process**
  Continuous integration/continuous delivery (CI/CD) pipelines must be configured to ease future updates.

Documentation

The journey of any AI service usually does not end with successful deployment. In addition to satisfying the current objective, a deployed AI service may attract more prospective clients or be employable in additional larger contexts. Therefore, the expedition of the AI service in its current life span must be documented. This document can also help different stakeholders understand and reason the AI service predictions and performance. Focus on standardised documentation to accompany AI assets is a fairly recent phenomenon and has been explored by efforts like Model Cards and Factsheets.

Therefore, to do the job right and with responsibility, the deployed model must also accompany a separate document that provides information about the AI system including general details, intended use cases, evaluation data and metrics, model performance measures, and so on. The motivation is essentially to increase transparency and bridge the expertise gap between the producer and consumer of an AI service by communicating the various attributes of the AI services in a standardised way. It may also help advertise the AI service better and assist in earning the confidence of prospective consumers.

Typically, the document must address the following areas:

- The intended usage and scope of the AI service
- The output of the service and any available additional features such as explainable outcomes
- Training, testing, and validating datasets
- Scope of testing and evaluation results
- Quality of the AI service concerning fairness and other domain-specific criteria
- Any associated data readiness documents
- Security features to counter any adversarial attacks
- Scalability measures
- Expected performance on unknown data distributions in terms of whether to expect a complete shutdown or graceful degradation


MONITORING
best practices
for responsible
monitoring of the
deployed system

AI systems learn continuously from newer data, so they must be monitored at regular intervals for their compliance with responsible AI standards and regulations.

The monitoring for compliance assessment can be a self-assessment via internal dashboards and processes or it can also include an audit by a regulator or third-party auditor. A dynamic dashboard for continuous monitoring as self-assessment is important because of data drift that can result in a negative feedback into the system and affect its responsible AI metrics (e.g., for fairness, robustness) over time.

Apart from dynamic dashboards, a cadence for compliance assessment could depend on how frequently an AI system receives major updates or when we notice significant change or drift in its data distribution or when there is any change or update in the standards or regulations for its end-application. One reason why the standards or regulations may change is because they are updated by the regulator or the body that sets them. Another reason why standards or regulations may change is because the same AI system may be used for different end-applications, e.g., face recognition for photo-labelling in photo album vs. face recognition for access to an essential service. To monitor any AI system, it is important to build and monitor dashboards about the properties of its training-vs-test data and data drifts, its responsible AI metrics on a representative evaluation data, and most importantly, keeping track of the right standards and regulations relevant for its end-application. The representative evaluation data for compliance assessment can either be requested from or provided by the regulator or the auditor (e.g., biometrics benchmark data sets released by NIST, https://www.nist.gov/itl/iad/image-group/resources/biometrics-evaluations) or if that is not available, AI technology providers can work with the regulator to volunteer some of their data to create and synthesise common benchmark databanks.

Regulatory sandboxes, where a system is stress-tested by a regulator in a controlled environment in a limited release for a limited amount of time, is another approach to monitoring a responsible AI system and limit the amplification of its harms. As an example, one can look at the regulatory sandbox set up by the AI regulator in Norway, https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificialintelligence/. To stress-test AI models and finding bugs, building check-lists and templates to test how the behaviour of a system changes under particular synthesised changes in its data or inputs is a good approach that is inspired by the best practices in software engineering, https://aclanthology.org/2020.acl-main.442/.

TOOLS FOR RESPONSIBLE AI

strategies and tools to mitigate privacy and security risks

The usage of AI in the industry has become widespread, with many of the AI systems working by learning non-trivial representations of data. With the decisions from these systems becoming increasingly harder to understand, there is a scope for multiple issues to be introduced into these systems that cause privacy and security concerns. This chapter discusses strategies a practitioner can adopt to mitigate these issues and a few tools that can help them such as AIX360 and ART.

Adversarial Robustness

Many of the Deep Neural Networks (DNNs) that form the backbone of modern AI systems have been prone to adversarial attacks and can often mislead the systems by altering their inputs. An adversary can also construct such inputs due to the lack of knowledge about the internal works of an AI systems leading to serious privacy and security issues in practices. Other concerning trends are adversarial attacks which do not manipulate the objects in the physical world and not the input. For example, an adversary can easily mislead autonomous vehicles' recognition systems by sticking tailored patches on traffic signs.

There are four types of adversarial threats to systems that manifest in several ways -evasion, poisoning, extraction, and inference.

Evasion Attacks

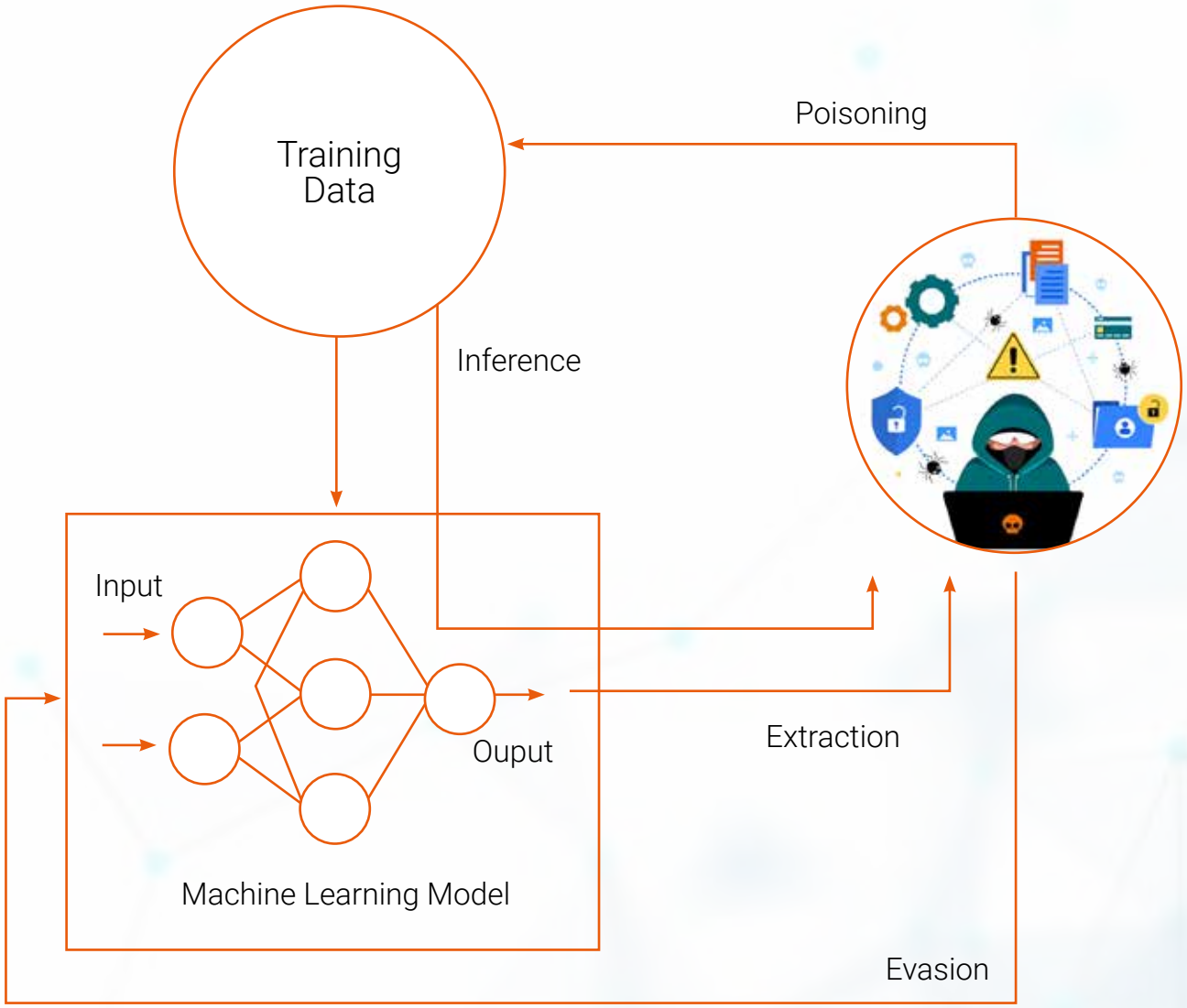
These are primary attacks that occur when an adversary tries to manipulate the input such that the system malfunctions. For instance, in the case of object recognition systems, an adversary aims to manipulate minimal pixels of the image such that the system gives spurious classification. Evasion attacks broadly fall into two categories - targeted and untargeted. In the case of untargeted attacks, the attacker or adversary aims to alter the examples and spurious outputs are produced by the system. As opposed to untargeted attacks, targeted attacks occur when an adversary manipulates the input to achieve a particular goal. For instance, in the case of spam filtering systems, the adversary can construct spam mails in such a way that they can bypass the spam filtering system.

These attacks can also be categorised based on the way the adversary has access to the AI systems - white-box or black-box attacks in these cases. In the case of white-box attacks, the adversary has intricate knowledge of the inner workings of the systems. In contrast, black-box attacks occur when the attacker has minimal knowledge about the deployed AI system. In the cases of black-box attacks, an adversary typically tries to construct a surrogate model based on the responses of the AI system for several of their queries and then uses the constructed surrogate to generate malicious examples.

Poisoning Attacks

This class of attacks occur when the adversary exposes AI systems during data collection. In practice, developers of AI systems often assume that they can completely trust the training data sets. However, they have minimal knowledge of how the data was collected as they are either crowdsourced or gathered by a third party. In turn, this could lead to an adversary manipulating the data collection process to achieve specific end goals, which can vary from reducing the system's performance to inserting back doors.

Figure 6



Extraction Attacks

This class of attacks involve an adversary repeatedly probing a black-box AI system to either create its clone
or extract the training data that forms its backbone. Such attacks can be disastrous especially when the backbone model or the training data are confidential. For example, several banks are now developing an automated system that assists their employees in trading and investments. If the internal working of such system is exposed, it could lead to several issues for the banks as an adversary can easily manipulate the system to make profits.

Inference Attacks

These attacks occur when an adversary can learn sensitive information about the internals of the deployed system by repeatedly probing the system. For instance, an attacker can repeatedly query the system to access sensitive information present in the training data.

Due to these threats, it is paramount that the practitioners build and deploy defenses and test them against adversarial attacks. A developer can implement the following strategies to protect AI systems against adversaries:

- Certifying and verifying the robustness of the underlying Machine Learning (ML) models and

improving its robustness with approaches such as pre-processing inputs

- Augmenting training data with adversarial samples
- Leveraging run-time detection methods to flag any inputs that an adversary might have modified

Testing AI Systems

Once the system is ready for deployment, it must be measured for robustness. A developer can assess the robustness by measuring the system's performance on manipulated examples. Another approach would be to calculate the variations in the internal working and the system's output based on minor input changes. The complete system must be evaluated by using penetration testing to identify vulnerabilities. Adversaries can exploit these loopholes to make your model susceptible to any adversarial attack. Identifying these vulnerabilities in advance can help plan ahead for any breaches waiting to happen.

Adversarial Training of Models

The process of adversarial training is an alternate strategy for addressing a malicious threat by strengthening the systems for robustness against manipulations. The simplest way to improve the robustness of a system would be to add these adversarial examples with correct labels to the training set, and then train the system on such a more extensive set.

Run-time Detection

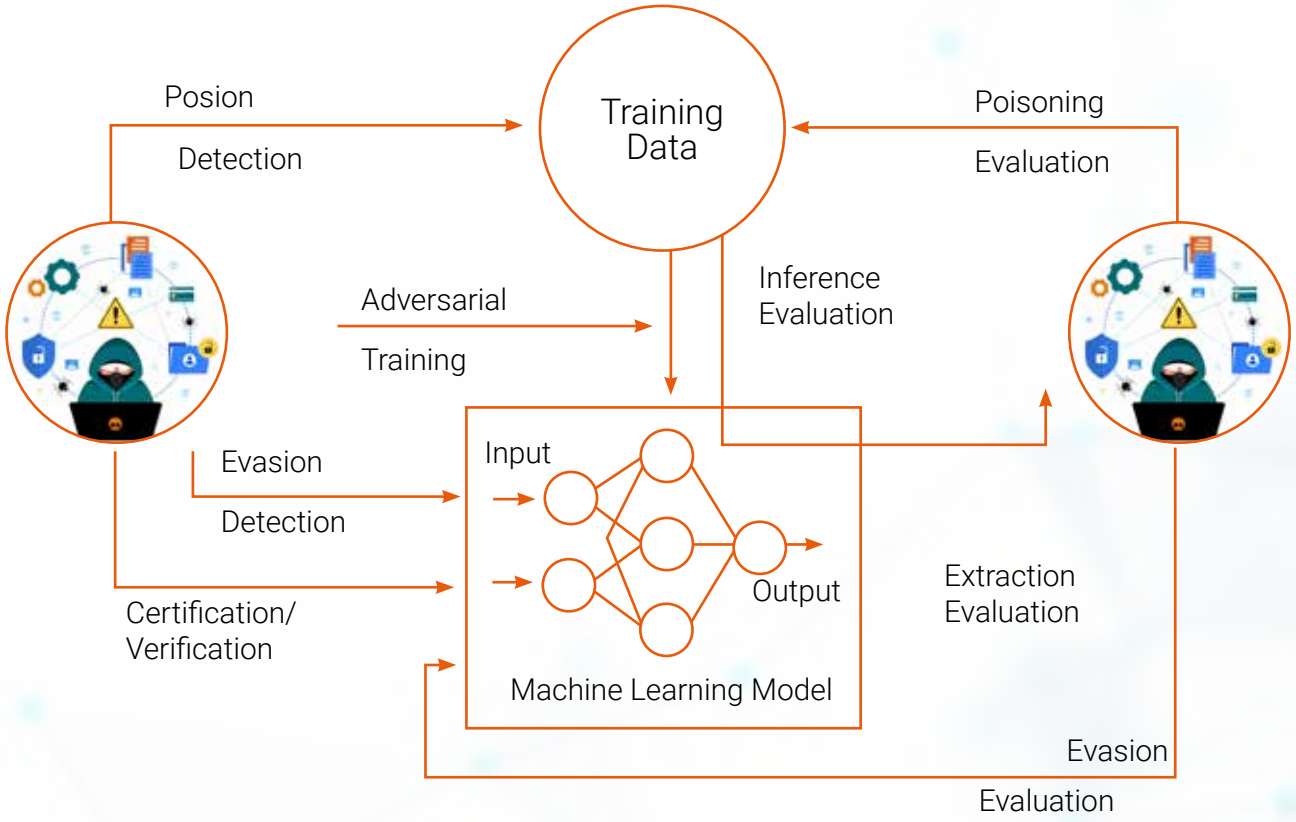
A developer can also employ run-time detection methods to flag any malicious inputs to the system. An alternate strategy would be to build a model that can easily filter out malicious inputs from the inputs that the system receives. These methods typically try to exploit abnormalities in the internal working of the AI systems caused by the adversarial inputs.

AI system developers must be mindful of the potential hazards connected with these systems. It is recommended to perform stress tests on the deployed systems frequently to detect as many potential flaws as feasible. Adversarial Robustness Toolbox (ART) is a Python library for AI Security developed by IBM Research.

ART is designed to support researchers and developers in creating novel defense techniques and deploying practical defenses for real-world AI systems. ART provides tools that enable developers to defend and evaluate the underlying DNNs and applications against the malicious threats of Evasion, Poisoning, Extraction, and Inference. The library also provides interfaces that support comprehensive defense systems' composition using individual methods as building blocks.

The techniques implemented in ART allow a practitioner to stress test underlying ML models against several state-of-the-art threat models. ART supports all popular Machine Learning frameworks (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, and so on), all data types (images, tables, audio, video, and so on), and downstream tasks (classification, object detection, speech recognition, generation, certification, and so on).

Figure 7



In practice, the developers can also leverage strategies used for the security testing of software systems. The tools ART provides (red and blue) indicates that the testers can be grouped into red and blue teams, where the red team can simulate real-world adversarial threats to bring down the deployed AI system. The blue team can help guard against these threats.

AI Explainability 360 toolkit

The AI Explain ability 360 toolkit (AIX360) is an open-source library that supports interpretability and explainability of datasets and machine learning models. The AI Explainability 360 Python package includes a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics.

The toolkit contains 10 different explainability techniques that cover Data Explainers, local post-hoc explainers, directly interpretable models, global post-hoc explainers, and self-explaining models. Multiple educational materials both introduce explainability algorithms included in AIX360 and demonstrate how different explainability methods can be applied in real-world scenarios.

The AIX360 toolkit currently includes five industry tutorials in the form of Jupyter notebooks that show data scientists and other developers how to use different explanation methods across several

application domains. The tutorials serve as an educational tool and potential gateway to AI explainability for practitioners
in these domains. Beyond illustrating the application of different methods, the tutorials also provide considerable insight into the datasets that are used and, to the extent that these insights generalise, into the respective problem domains.

AI Fairness 360 Toolkit

Machine learning models are increasingly used to inform high-stakes decisions about people. Although machine learning is inherently a form of statistical discrimination, the discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage.

Biases in training data due to either prejudice in labels or under and over-sampling yields models with unwanted bias.

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.

AI Fairness 360 (AIF360) includes various metrics and algorithms which may result in a daunting problem of making the right selection for a given application. Fairness is a multifaceted, context-dependent, social construct that defies simple definition. The metrics and algorithms in AIF360 may be viewed from the lens of distributive justice, and clearly do not capture the entire scope of fairness in all situations.

The toolkit must only be used in a very limited setting - allocation or risk assessment problems with well-defined protected attributes in which one would like to have some sort of statistical or mathematical notion of sameness. However, the code and collateral contained in AIF360 is only a starting point to a broader discussion among multiple stakeholders on overall decision-making workflows.

Bias mitigation algorithms attempt to improve the fairness metrics by modifying the training data, learning algorithm, or predictions. These algorithm categories are known as pre-processing, in-processing, and post-processing respectively. The choice amongst the algorithm categories can partially be made based on the user persona's ability to intervene at different parts of a Machine Learning pipeline.

- If the user is allowed to modify the training data, then pre-processing can be used.
- If the user is allowed to change the learning algorithm, then in-processing can be used.
- If the user can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used.

CASE STUDIES

Demonstrating industry adoption of responsible AI

This section contains case studies of responsible AI adoption from Microsoft, IBM, and Fractal Analytics.

Fractal's COVID Social Distancing Compliance

Problem Statement

Monitoring CCTV footage is time-consuming and resource-consuming and becomes even slower when driving compliance via interventions. To help drive public health awareness and compliance, Fractal collaborated with a drone manufacturing company, under public authority oversight and use, to track non-compliance with social distancing guidelines via streaming video feeds. The essential purpose of this project aimed to reduce human effort in monitoring CCTVs and provide instant alerts to authorities to enable quick decision-making.

Fractal independently evaluated the problem statement and the purpose of the effort to ensure it is the right fit and aligns with the principles we laid out. After rigorous assessment, Fractal concluded that the project not only passed the newspaper test but also helped drive a beneficial outcome for society.

Solution

The designed solution flagged instances of social distance violations in real-time. On top of the real-time detection, A dashboard helped visualize the instances and created KPIs like the number of infringements and severity based on determined thresholds by location. The goal was to empower the users to make quick decisions and give a complete view of social distance compliance.

Figure 8

Putting Responsible AI into Practice

To make sure that we designed the solution in compliance with the responsible AI framework, we asked the right questions in each phase of the AI development life cycle.

*The questions shown in the figure above are not exhaustive.

During this exercise, we recognized that we would have to design the solution with 3 primary principles in mind, namely, - privacy, transparency, and accountability.

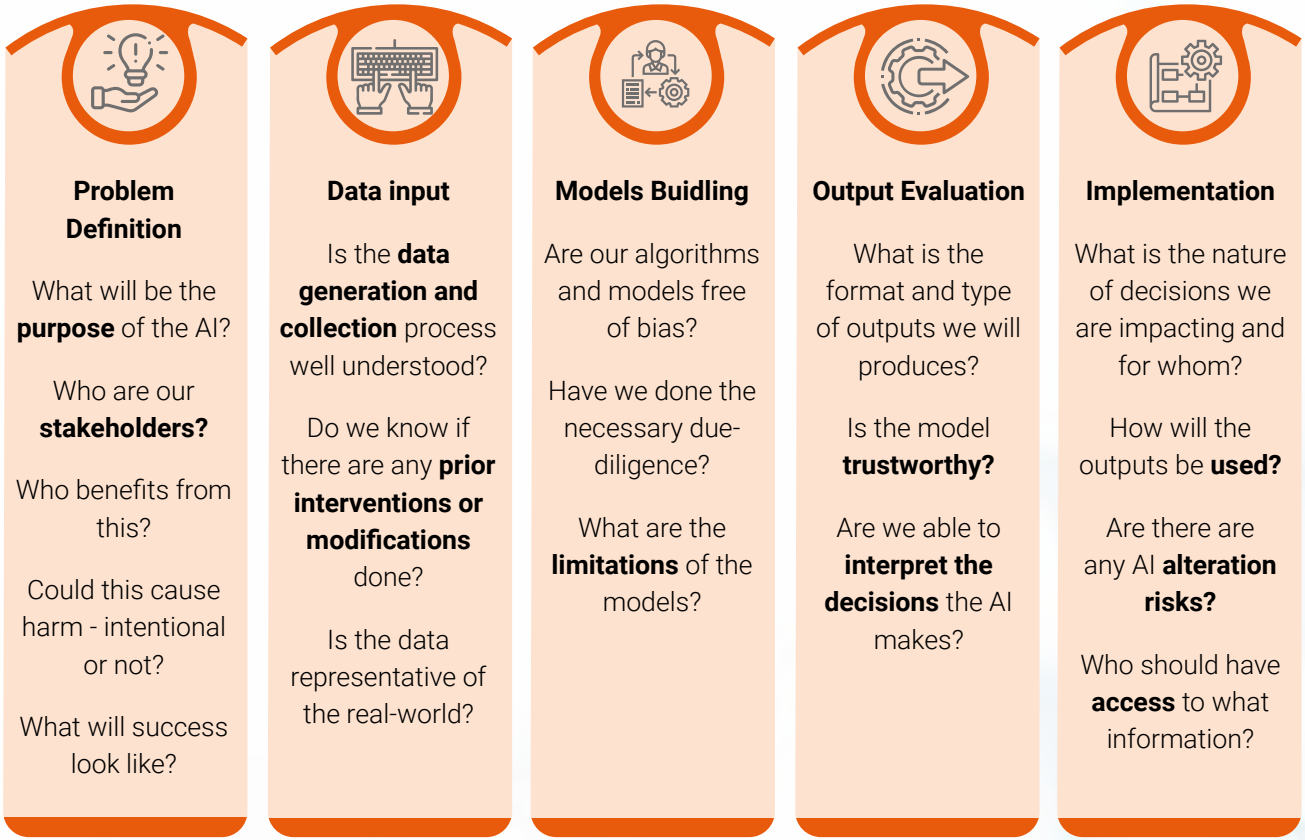
Privacy

The primary objective was to allow for categorization of objects while maintaining their privacy. To do this, we blurred all the detected objects so that any of the views on the video streams would also hide the privacy of the object. The identity of the object was masked and encrypted to restrict access to the same.

Transparency

Since we are creating a solution that might have direct impact on society and its collective behavioural traits, we ensured that complete transparency is enabled. We restricted the transparency view and its access control under the supervision of the public authority. Only the stakeholders who have the right authorization can use the encrypted keys to get full transparency to the operations on object identity retrieval.

Figure 9*



Accountability

Since restricted access divulges information on a need-to-know basis, accountability became of utmost importance. To establish accountability, we ensure that each such access is recorded. A standard audit trail for access includes but not limited to, the accessor details, accessing objecsts details, timestamp, reason for process - access, retention period-record, etc.

IBM Helping US Employer to Put Anti-Bias First for Fair Hiring

Problem Statement

AI-based hiring tools or recruiting systems built on biased datasets could reinforce historical discrimination while screening candidates for a particular role.

One major US corporation was eager to tackle this problem on a large scale and turned to IBM for help. The corporation's mandate included driving efforts to ensure workforce diversity and inclusion. When it came to its hiring practices, it was critical to ensure that its AI/ML models were fair and trustworthy.

The corporation's data science leaders wanted to be able to translate the models' decisions and results easily — in a way any hiring manager could understand. It wanted to establish fairness by accelerating the identification of any bias in hiring and explain decisions made by AI models. The corporation also knew it needed to operationalise AI governance to get more of its business users on board — so it set out to find a solution that could achieve all of these goals.

Solution

The answer was IBM Watson, an AI monitoring and management tool that filled a much-needed gap. It provided explainable AI as a set of processes and methods that allowed users to comprehend and

trust the results and output created by AI algorithms, including its expected impact and potential biases. It helped characterize model accuracy, fairness, transparency, and outcomes in AI-powered decision-making. Explainable AI is crucial for an organization in building trust and confidence when putting AI models into production.

Now, the company is proactively monitoring for and mitigating bias in its hiring processes. Because automation has reduced the workload within DevOps, the company's data scientists can focus more on the new model development and refinements.

Putting Responsible AI into Practice

According to IBM, if a business is involved in making decisions on automation that's driven by AI, it needs to be transparent. The business must know it's making decisions that align with company policy — and that people who are making the decisions based on AI can trust it.

We start by asking the more precise questions like

- How do we understand what AI models are doing?
- How do we ensure AI accuracy and fairness?
- How do we speed up the production and adoption of AI models?
- Can we trust the output?

Sensitive features such as gender, ethnicity, and age, even if not included in AI, could influence the training of the data, the source of the data, and even how the data got to the system from a training dataset. In other words, even if there's no intent or access to those features, in the beginning, those perceptions can lead to incorrect decisions.

Microsoft's Response to Novel Attacks

Problem Statement

In the year 2016, Microsoft released a chatbot on Twitter called Tay. Tay was taught to learn unsupervised from interactions with Twitter users, so she could replicate human communication and personality traits better. However, in a span of 24 hours, Twitter users realized that she could learn and began to feed her bigoted rhetoric, turning her from a polite bot into a medium for hate speeches. This case study taught us that while technology may not be unethical on its own, users do not always have the right intentions and the human element must be considered when designing AI systems.

Solution

Microsoft developed systems that were resilient to new types of attacks that influence learning datasets, especially for AI systems with automatic learning capabilities. To ensure that a similar incident does not recur, we developed technologically advanced content filters and introduced supervisors for AI systems with automatic learning capabilities.

OUR CONTRIBUTORS

Akbar Mohammed
Architect, Fractal Analytics

Akshat Chaudhari
Customer Service Support Manager, Big Data, Microsoft

Amit Deshpande
Researcher, Microsoft Research India

Amit Kumar
Associate Director, Data Science and Machine Learning, Deloitte India

Anantha Sekar
Product Management Group, Experience and Intelligence, Tata Consultancy Services

Anjali Pathak
Product & Social Media Lead (INDIAai), NASSCOM

Ankit Bose
Head of AI, NASSCOM

Aparna Gupta
Executive Director, Customer Success, Microsoft

Balaji Ganesan
Senior Research Engineer, IBM Research

David George
Director, Risk Advisory, Deloitte India

Deepak Vijaykeerthy
Research Engineer, Data and AI, IBM Research

Johar Batterywala
Partner, Deloitte Haskins & Sells

Keerthana Kennedy
Assistant Consultant, Tata Consultancy Services

Krutika Choudhary
Consultant, Fractal Analytics

Madhav Bissa
Program Director, CoE for DSAI, NASSCOM

Manish Kesarwani
Advisory Research Scientist, IBM Research

Mitesh Kapadia
Associate Director, Analytics and Data Science, Deloitte

Payal Agarwal
Partner, Deloitte

Pranay Lohia
Senior ML Researcher, Microsoft

Rohini Srivathsa
National Technology Officer, Microsoft India

Sagar Shah
Client Partner, Fractal Analytics

Sai Kavitha Krishnaiyengar
Director, Support Engineering, Microsoft

Sameep Mehta
Distinguished Engineer, AI and Hybrid Data, IBM Research

Sangeeta Gupta
Senior Vice President, NASSCOM

Shweta Gupta
Engineering Lead for Data and Analytics, Asia, Microsoft

Supriya Samuel
Branding & Marketing Manager, NASSCOM

Swapna Choudhury
Associate Director Deloitte India

Tarun Kumar
Data Strategist, Centre of Excellence - Data Science & Artificial Intelligence, NASSCOM

Vijay Arya
Senior Researcher, IBM Research

Raj Shekhar
Responsible AI Lead, NASSCOM

Contact Person

For more information, Contact
Raj Shekhar
Responsible AI Lead, NASSCOM
raj@nasscom.in

NASSCOM®