AI Ethics

Fairness, Accountability, and Transparency

Fairness, Accountability, Transparency

Fairness

What is Fairness?

While there is no universally agreed upon definition for fairness, we can broadly define fairness as the absence of prejudice or preference for an individual or group based on their characteristics.

Fairness Components

- **Data Bias**
  - Historical Bias
  - Representation Bias
  - Measurement Bias
- **Model Bias**
  - Evaluation Bias
  - Aggregation Bias

Data Bias

Data Bias is a bias that arises during the data collection period of developing an AI system. The bias resides within the AI system.

Data Bias

Historical bias is the already existing bias in the world that has seeped into our data.

For example, a machine learning model trained on Wikipedia produced gender-biased analogies like: **man : doctor** :: **woman : nurse**, or **man : commander** :: **woman : school teacher**. The model inherited the historical biases of society by learning from the huge corpora of text, and produced work further reinforcing those biases.

Data Bias

Representation bias this happens from the way we define and sample a population to create a dataset.

For example, the data used to train Amazon’s facial recognition was mostly based on white faces, leading to issues detecting darker-skinned faces.

Data Bias

Measurement bias occurs when choosing or collecting features or labels to use in predictive models. Data that’s easily available is often a noisy proxy for the actual features or labels of interest.

For example, survey interviewers asking about deaths were poorly trained and included deaths which occurred before the time period of interest.

This would lead to an overestimate of the mortality rate because deaths which should not be included are included.

Model Bias

Model Bias is a bias that arises from the methods in which the AI system evaluates the situation and makes decisions.

Model Bias

Evaluation bias occurs during model iteration and evaluation. A model is optimized using training data, but its quality is often measured against certain benchmarks. Bias can arise when these benchmarks do not represent the general population, or are not appropriate for the way the model will be used.

An example of evaluation bias would be focusing on optimizing a model solely on loss and ignoring other hyperparameters such as R2 or accuracy. Evaluation bias can also be impacted by representation bias in the evaluation (test) data set.

Model Bias

Aggregation bias arises during model construction where distinct populations are inappropriately combined. There are many AI applications where the population of interest is heterogeneous, and a single model is unlikely to suit all groups.

For example an NLP model trained to detect aggressive/violent language in a general context, may end up falsely flagging text that is quoting popular sayings, song lyrics, slang in local contexts.

Accountability

What is accountability?

Accountability refers to the method and resources used to ensure failures within a system can be corrected to mitigate or reduce harm.

Accountability Components

- Good Governance
- Understanding Data
- Define Performance Goals and Metrics
- Monitoring Performance

Good Governance

Governance can can be broken down into two parts

- **Organizational level**: Governance at the organizational level helps entities ensure oversight and accountability and manage risks of AI systems.
- **System level**: Governance at the system level helps entities ensure AI systems meet performance requirements.

Good Governance - **Organizational** Level

- **Clear goals:** Define clear goals and objectives for the AI system to ensure intended outcomes are achieved.
- **Roles and responsibilities:** Define clear roles, responsibilities, and delegation of authority for the AI system to ensure effective operations, timely corrections, and sustained oversight.
- **Values:** Demonstrate a commitment to values and principles established by the entity to foster public trust in responsible use of the AI system
- **Workforce:** Recruit, develop, and retain personnel with multidisciplinary skills and experiences in design, development, deployment, assessment, and monitoring of AI systems.
- **Stakeholder involvement:** Include diverse perspectives from a community of stakeholders throughout the AI life cycle to mitigate risks.
- **Risk management:** Implement an AI-specific risk management plan to systematically identify, analyze, and mitigate risks.

Good Governance - **Systems** Level

- **Specifications**: Establish and document technical specifications to ensure the AI system meets its intended purpose.
- **Compliance**: Ensure the AI system complies with relevant laws, regulations, standards, and guidance.
- **Transparency**: Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.

Understanding Data - Components

When determining how accountability plays into our use of data, we must consider two primary types of data

- Data used for Model Development
- Data used for System Operation

Data Used for **Model Development**

- **Sources**: Document sources and origins of data used to develop the models underpinning the AI system.
- **Reliability**: Assess reliability of data used to develop the models.
- **Categorization**: Assess attributes used to categorize data.
- **Variable Selection**: Assess data variables used in the AI component models.
- **Enhancement**: Assess the use of synthetic, imputed, and/or augmented data.

Data Used for **System Operation**

- **Dependency**: Assess interconnectivities and dependencies of data streams that operationalize the AI system.
- **Bias**: Assess reliability, quality, and representativeness of all the data used in the system’s operation, including any potential biases, inequities, and other societal concerns associated with the AI system’s data.
- **Security and Privacy**: Assess data security and privacy for the AI system.

Define Performance Goals and Metrics - Components

GAO developed nine key practices for this principle, grouped into two categories:

- Component Level
- System Level

Define Performance Goals and Metrics - **Component Level**

- **Documentation**: Catalog model and non-model components along with operating specifications and parameters.
- **Metrics**: Define performance metrics that are precise, consistent, and reproducible.
- **Assessment**: Assess the performance of each component against defined metrics to ensure it functions as intended and is consistent with program goals and objectives.
- **Outputs**: Assess whether outputs of each component are appropriate for the operational context of the AI system.

Define Performance Goals and Metrics - **System Level**

- **Documentation**: Document the methods for assessment, performance metrics, and outcomes of the AI system to provide transparency over its performance
- **Metrics**: Define performance metrics that are precise, consistent, and reproducible.
- **Assessment**: Assess performance against defined metrics to ensure the AI system functions as intended and is sufficiently robust.
- **Bias**: Identify potential biases, inequities, and other societal concerns resulting from the AI system
- **Human supervision**: Define and develop procedures for human supervision of the AI system to ensure accountability.

Monitor Performance regularly

GAO developed five key practices for this principle, grouped into two categories:

- **Continuous monitoring of performance**: This category involves tracking inputs of data, outputs generated from predictive models, and performance parameters to determine whether the results are as expected.
- **Assessing sustainment and expanded use**: This category involves examining the utility of the AI system, especially when applicable laws, programmatic objectives, and the operational environment may change over time. In some cases, entities may consider scaling the use of the AI system (across geographic locations, for example) or expanding its use in different operational settings.

Transparency

What is transparency?

Transparency refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to its environment, and to the governance of the data used.

Components of Transparency

- Traceability
- Communication
- Intelligibility/Explainability