Ethical Application of Artificial Intelligence Framework

ACT-IAC WHITE PAPER

ETHICAL APPLICATION OF ARTIFICIAL INTELLIGENCE FRAMEWORK

EMERGING TECHNOLOGY COMMUNITY OF INTEREST
Artificial Intelligence Working Group

Date Released: October 8, 2020

Synopsis
Artificial Intelligence (AI) is becoming an influential cornerstone of the digital future as we increasingly rely on it to support and inform our world. As organizations become more dependent upon the many technologies which comprise AI, there is a need to determine how much confidence and trust to place in them. To accomplish this ethically, a means to determine its performance and continually monitor for mission veracity or any adverse impacts is needed. This paper addresses the overall ethical framework through an index that evaluates five core parameters underpinning the impact of AI: Bias, Fairness, Transparency, Responsibility, and Interpretability. By addressing and incorporating these five components, the framework can be used to understand the ethical automation of solutions to meet desired mission outcomes.

American Council for Technology-Industry Advisory Council (ACT-IAC)

The American Council for Technology-Industry Advisory Council (ACT-IAC) is a non-profit educational organization established to accelerate government mission outcomes through collaboration, leadership and education. ACT-IAC provides a unique, objective, and trusted forum where government and industry executives are working together to improve public services and agency operations through the use of technology. ACT-IAC contributes to better communication between government and industry, collaborative and innovative problem solving, and a more professional and qualified workforce.

The information, conclusions, and recommendations contained in this publication were produced by volunteers from government and industry who share the ACT-IAC vision of a more effective and innovative government. ACT-IAC volunteers represent a wide diversity of organizations (public and private) and functions. These volunteers use the ACT-IAC collaborative process, refined over forty years of experience, to produce outcomes that are consensus-based.

To maintain the objectivity and integrity of its collaborative process, ACT-IAC welcomes the participation of all public and private organizations committed to improving the delivery of public services through the effective and efficient use of technology. For additional information, visit the ACT-IAC website at www.actiac.org.

Emerging Technology Community of Interest

ACT-IAC, through the Emerging Technology Community of Interest, formed an Artificial Intelligence Working Group to give voice to and provide an authoritative resource for government agencies looking to understand and incorporate AI/ML technology and functionality into their organizations. This working group includes government and industry thought leaders incubating government use cases. The ACT-IAC Emerging Technology Community of Interest (ET COI) mission is to provide an energetic, collaborative consortium comprised of leading practitioners in data science, technology, and research, engaged with industry, academia, and public officials and executives focused on emerging and leading technologies which transform public sector capabilities.

Disclaimer

This document has been prepared to contribute to a more effective, efficient, and innovative government. The information contained in this report is the result of a collaborative process in which several individuals participated. This document does not – nor is it intended to – endorse or recommend any specific technology, product, or vendor. Moreover, the views expressed in this document do not necessarily represent the official views of the individuals and organizations that participated in its development. Every effort has been made to present accurate and reliable information in this report. However, neither ACT-IAC nor its contributors assume any responsibility for consequences resulting from the use of the information herein.

Copyright
©American Council for Technology, 2020. This document may not be quoted, reproduced and/or distributed unless credit is given to the American Council for Technology-Industry Advisory Council.

For further information, contact the American Council for Technology-Industry Advisory Council at (703) 208-4800 or www.actiac.org.

Table of Contents

Executive Summary

This paper and its index are intended to be an advisory framework to highlight that humans are ultimately responsible for the ethical application of Artificial Intelligence (AI) solutions. By monitoring and measuring critical elements of AI throughout the lifecycle of development, implementation, and operations, one can assess an AI application's level of credibility, and thus, the level of confidence to place in that instance of this rapidly evolving technology. This confidence can be demonstrated through an index that incorporates five core parameters underpinning the impact of AI on those systems: Bias, Fairness, Transparency, Responsibility, and Interpretability.

1. **Bias**: AI algorithms learn from large quantities of data. The machine learning models that the AI builds can amplify some of the biases inherently present in the data. Accountable owners of AI systems should identify and address bias in AI to prevent negatively impacting desired mission outcomes or individuals in protected classes or statuses.
2. **Fair:** AI systems should be designed to avoid potential risk of unfair impact within the context of use, whether intentional or unintentional.
3. **Transparent**: AI systems should be developed so that models, data, and results are auditable and explainable to decision-makers and the general population to the extent and manner appropriate or possible.
4. **Responsible**: The implementation of an AI solution must be relevant to the purpose of the task. It must ensure that both data and model sources are uncompromised. It must produce repeatable, legal, authentic, auditable, and effective results.
5. **Interpretable**: Stakeholders should thoroughly understand what AI has been asked to provide. They should be able to ensure that both data and model sources are credible, and will produce repeatable, trustworthy, and effective results.

The Ethical Application of AI Index (EAAI) framework allows for an establishment of a consistent measure upon which one may qualify and quantify components used to create, operate, and improve AI capabilities. It provides the means to monitor and measure the influence AI has on systems throughout the entire lifecycle. Each component should be reviewed and scored based on its applicable indicators and resulting implications. The indicators quantify the quality of the characteristics and the implications quantify the impact. These components will deliver an overall score that should be used to evaluate the ethics of an AI system, which should continually be monitored and checked over time.

For additional background, it is recommended that readers review the following documents: ACT-IAC Artificial Intelligence Primer and the ACT-IAC Artificial Intelligence Playbook[1]

[1] https://www.actiac.org/act-iac-white-paper-artificial-intelligence-machine-learning-primer
https://www.actiac.org/act-iac-white-paper-artificial-intelligence-playbook

Ethical Application of Artificial Intelligence Framework (EAAI)

Artificial Intelligence (AI) is becoming an influential cornerstone of the digital future as we increasingly rely on it to support and inform our world. As organizations become more dependent upon the many technologies which comprise AI, there is a need to determine how much confidence and trust to place in them. To accomplish this ethically, a means to determine its performance and continually monitor for mission veracity or any adverse impacts is needed. This paper presents five subsections addressing the overall ethical framework. Each subsection is able to stand alone and can be integrated into the whole of the ethical application of Artificial Intelligence. The subsections logically overlap (see Figure 1) by design. When all five subsections are addressed, the framework can be used to understand the ethical application of AI solutions to support the organization's mission.

The origin of data and model bias is mostly human-made, human-thought, and human-omitted. In recognition of our ethical and moral obligations as public stewards, one must recognize the challenges involved when combining technical and human factors. Therefore, there is a scorecard example in the last section that demonstrates how the components can be combined to culminate in an overarching score (Index). These metrics are both qualitative and quantitative in nature, and aim to assess the level of potential risk/uncertainty as they relate to the ethical implications of AI across its lifecycle. The Artificial Intelligence Working Group developed the Ethical Application of AI Index (EAAI) that allows for identification of the risks associated with AI systems. The framework could be evolved into tools to measure the ethics of an AI system along the dimensions of the indicators and implications. It can be used in individual scenarios or use cases to determine the structure and philosophy that combines the mission's goals and manages the AI system strategically and ethically.

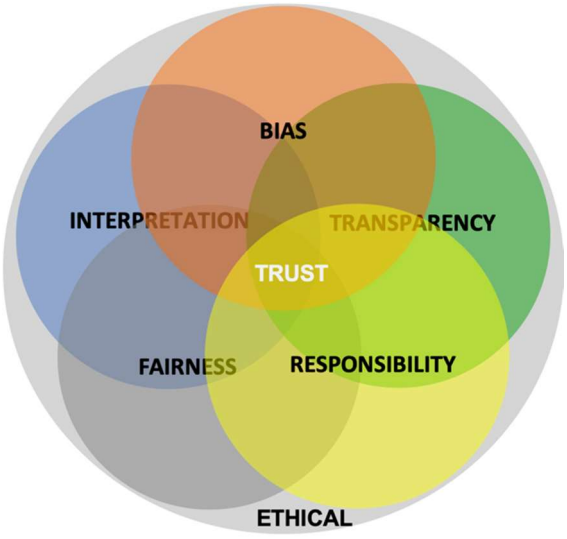


Figure 1: AI Ethics Components

Figure 2: Efficacy of Trust

Each component is described along the lines of the following efficacy of trust characteristics:

Define Distinction: WHAT the component is.
Differentiate Differences: HOW the constituent parts comprise the totality (Indicators).
Distinguish Differentials: WHERE the application of the indicators influences the outcome (Implications).
Designate Determine: WHEN each indicator impacts outcomes (Monitor).
Discern: Present/Future – WHY the results impact our view of the environment (Measure).

BIAS COMPONENT

Bias in an Artificial Intelligence algorithm is a reflection of the organization(s) and person(s) who implement and integrate the AI. The General Services Administration (GSA) has made Bias training mandatory. According to the GSA Online University, "*Whether we know it or not, we all possess unconscious biases affecting us inside and outside the workplace*.[2]" Unconscious bias may result from ingrained stereotypes, omission resulting from lack of awareness as to the variable's relevance (under-fitting the model), or even confirmation bias of data that results from prior association of the data with similar models. Implicit and explicit bias in the adjudication of benefits or in the adjudication of risk is neither novel nor unique. Even before the introduction of Artificial Intelligence ecosystems at scale in the federal, state, and local governments, there were many instances where bias was introduced into systems by humans through written and unwritten policies and procedures related to interactions with government agencies.

Given the increased use and reliance on AI capabilities, it is critical to understand how bias influences and affects the inputs to AI, the algorithms that operate it, and the interpretations that provide insights to decision-making. Since AI algorithms are informed by large quantities of data from which they analyze and provide answers, the machine learning models that the AI builds can amplify some of the biases inherently present in the data. It is imperative that the adoption, implications, and impact of data are monitored and measured. Through this process, one can identify and potentially eliminate explicit bias in AI. This requires collaboration across disciplines to develop and implement technical improvements, and operational practices to address the inherent bias.

It's important for all of us to remember that there is no universal and unchanging set of ethics, and that regional and cultural diversity are key to any conversation about AI ethics.

Definition

Bias *Influences* decisions and is pivotal to all ethical subgroups.

[2] https://corporateapps.gsa.gov/hr-apps/gsa-olu/

Indicators

Indicators identify those qualitative areas of practice which best ensure that the technology will be appropriately employed. If monitored closely, these practice areas can increase confidence and potentially reduce the chance of abuse given any use case. Based on the objective of the AI use case, the following are important indicators for consideration:

Context of Use: The AI use case should be developed with a clear understanding of the objective and usage of the system.

When developing an AI solution, one should understand the following:

- Goal of the use case;
- Data sources used to achieve the objective;
- Decision owners/governors;
- Decision triggers (e.g., algorithm or business process);
- Clarity about the objective and usage of the system.

Diversity: The AI solution should include diversity among various aspects including the team, protected classes, and stakeholders.

Although most government entities have gotten much better at identifying and rectifying explicit bias, there have been many challenges with identifying and rectifying implicit bias resulting in disparate impact to individuals in protected classes or protected status. There are numerous areas where one must include diverse and inclusive perspectives to ensure these biases are addressed, including:

- Diversity of the team involved in data provenance, collection, and labeling;
- Diversity across the nine protected classes (sex, race, age, disability, color, creed, national origin, religion, or genetic information) as per the US federal law;
- Diversity of the engineering team involved in creating the algorithms;
- Inclusion of Stakeholder committee members including representation from various departments – Legal, HR, Sales, Marketing etc.;
- Training for diversity, equity and inclusion (DEI).

Data Bias: AI systems should identify and address bias in the data to prevent negatively impacting individuals in protected classes or status.

Cleansing the data to remove links in the relationship between outcomes and protected characteristics is essential to editing metadata to produce representations of the data that do not contain information about sensitive classes and status. Conducting an independent validation and verification process is significant to determine if outputs from a model are similar to the original inputs when reverse engineered. The bias inherent in the data must be evaluated from the following areas:

- Statistical distribution of data and methods dealing with skewed data;
- Appropriateness of data sets (confirm the data sets include complete information and no key information is missing);

- Engineering and impact of synthetic data;
- Over-fitting and under-fitting of the data.

Data Modeling: Since AI algorithms learn from large quantities of data, the data models must consider solutions to address the biases inherently present in the data.

AI solutions and their models need to consider the following**:**

- Size and variability of the test and training data;
- Reproducibility of the results;
- Factors used for predictions (e.g. psychological, behavioral, geographical or any other societal inferences);
- Training data considerations for protected classes;
- Explainable methodology for modeling.

Implications

Implications refer to the level of impact on outcomes. Compromised indicators result in compromised outcomes. In this discussion, every indicator has a corresponding implication. The probability of a compromised indicator multiplied by the level of impact will produce the index. This index represents the technology scored. For this exercise, the implications are expressed with "if…, then…" logic.

Objective: *If the Objective of the Use Case is not understood, then the resulting solution may benefit some individuals more than other individuals or groups, missing intended outcome.*

Diversity: *If diverse perspectives and solutions are not included, then psychological, behavioral, geographical or any other societal inferences may be used for predictions that may adversely impact certain individuals or groups' social and economic interest, health, access or mobility.*

Data Bias: *If the data bias is not identified and addressed, then the solution may have unintended consequences or detrimental conclusions toward an individual or class.*

Data Modeling: *If data models do not address the biases, the solution may restrict an individual from access to a specific business product or service.*

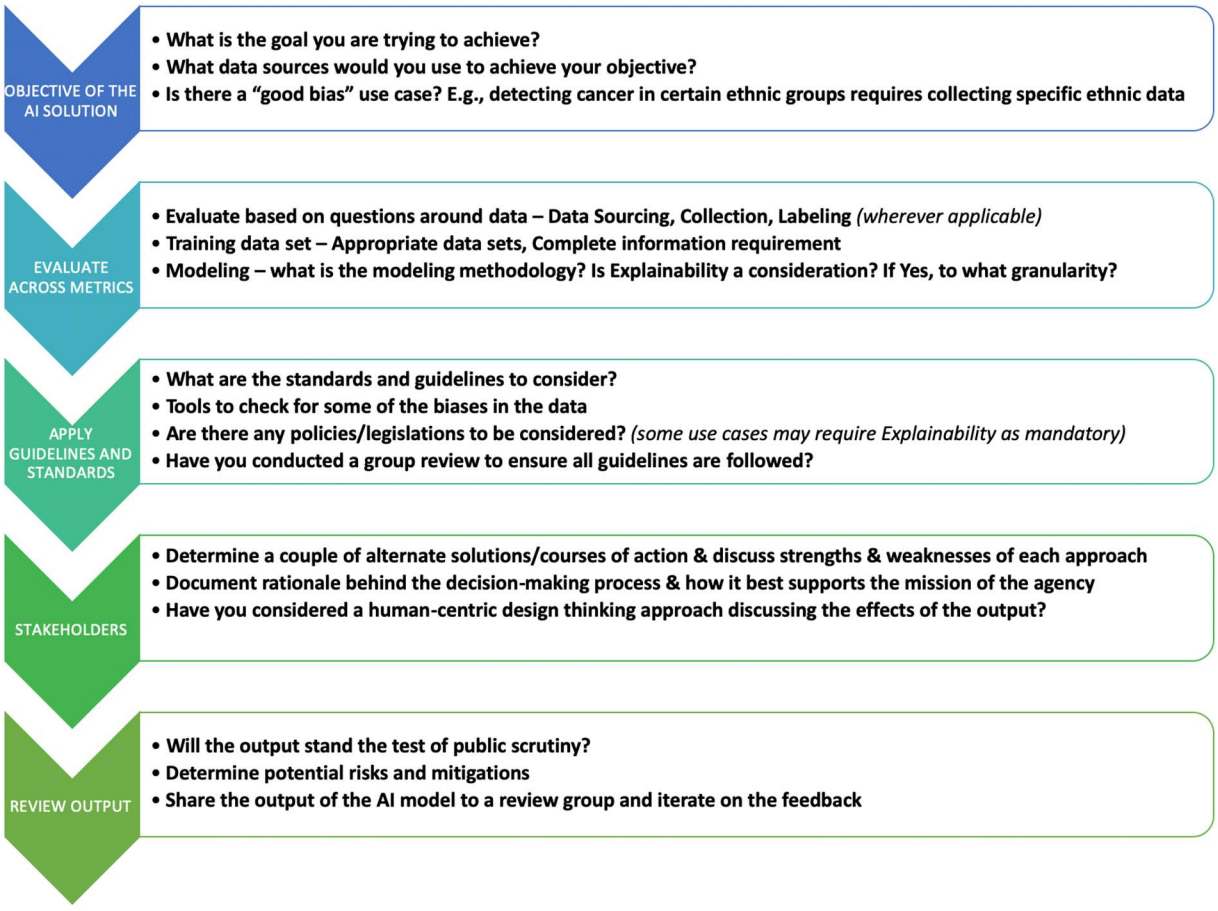
Monitor and Measure

It is important to monitor and measure the indicators and implications of the components throughout the project lifecycle. Monitoring includes auditing for performance of algorithms against key value-driven metrics such as accountability, bias, and transparency. Measurement is tracking smart metrics and key performance indicators (KPIs) throughout the lifecycle.

- Elaborate end-to-end testing
  - *Data sources and conditions testing*
  - *Algorithmic testing*
  - *System and regression testing of outputs*
- Monitoring of the AI systems includes auditing for performance of algorithms against key value-driven metrics such as accountability, bias, and transparency;
- Using alerts and notifications ensuring that the stakeholders are notified whenever there is an abnormal change or anomaly to the outputs;
- Tracking defined business parameters;
- Using feedback mechanisms to find key business incidents as soon as they occur;
- Robust scrutiny of outputs.

Five Steps to Monitor Bias

	Monitor Implications	Measure Impact
Context of Use	Detect anomalies	Robust scrutiny of outputs
Diversity	Use cases for various customer personas	Measure outcomes for various personas



	Monitor Implications	Measure Impact
Data Bias	One class in the training set dominating the others	Data distribution and balance
ML Modeling	Algorithmic, system and regression testing	Audit expected results against key value-driven metrics

FAIRNESS COMPONENT

AI system owners should be concerned about fairness because of the potential of discriminatory, unwanted, undesirable, or unacceptable social, economic, health, or legal outcomes. Measuring fairness in AI requires adequate identification of potential risks that might be introduced intentionally or unintentionally. A main objective in a successful and trustworthy implementation of AI is to arrive at a trusted solution as free as possible from enabling the potential for unfair strategic advantage or undesirable outcome.

If AI solutions produce unfair outcomes and behaviors, especially if these are harmful, future implementations of AI may face limited adoption and stall its full potential. By establishing a framework that considers key indicators of fairness throughout the various phases of the AI lifecycle, including design, development, implementation, and monitoring, it would be possible to effectively enhance the users' ability to detect, and understand related implications and unwanted biases. Fairness depends on the context of use and the intent of the AI implementation. Fairness in AI points to enabling awareness of underlying data and processes involved to identify inclusive and impartial representation and treatment of all relevant attributes needed to achieve desired objectives.

The level of diversity, the inclusion of multiple stakeholders with different perspectives, the understanding of the context of use, the level of transparency, and interpretability of algorithms are key indicators for fairness. Fairness should be continuously monitored throughout the AI process life cycle. It is very important to frequently test for quality and compliance. User feedback, quantitative and qualitative metrics need to be continuously captured and evaluated to detect, understand, and address unwanted bias that may lead to unfair outcomes or behaviors.

Definition

Fairness in AI refers to inclusive and impartial representation and treatment to achieve the desired outcomes within the context of use. Given equal initial inputs, outcomes are fair if they minimize variance and have an equal probability of occurrence. Equality is starting in the same place and fairness is ending in the same place.

Fairness in AI is best represented in terms of **Intent**, **Impact**, and **Evaluation**, which are the building blocks towards establishing the Trust of an AI solution. The context of use of the AI solution, the transparency of, and potential bias in the data and algorithms used for its development, implementation, and operation also play significant roles in driving fairness in AI. These building blocks should be continuously monitored and evaluated in an iterative fashion according to pre-established performance criteria and ongoing user-generated feedback.

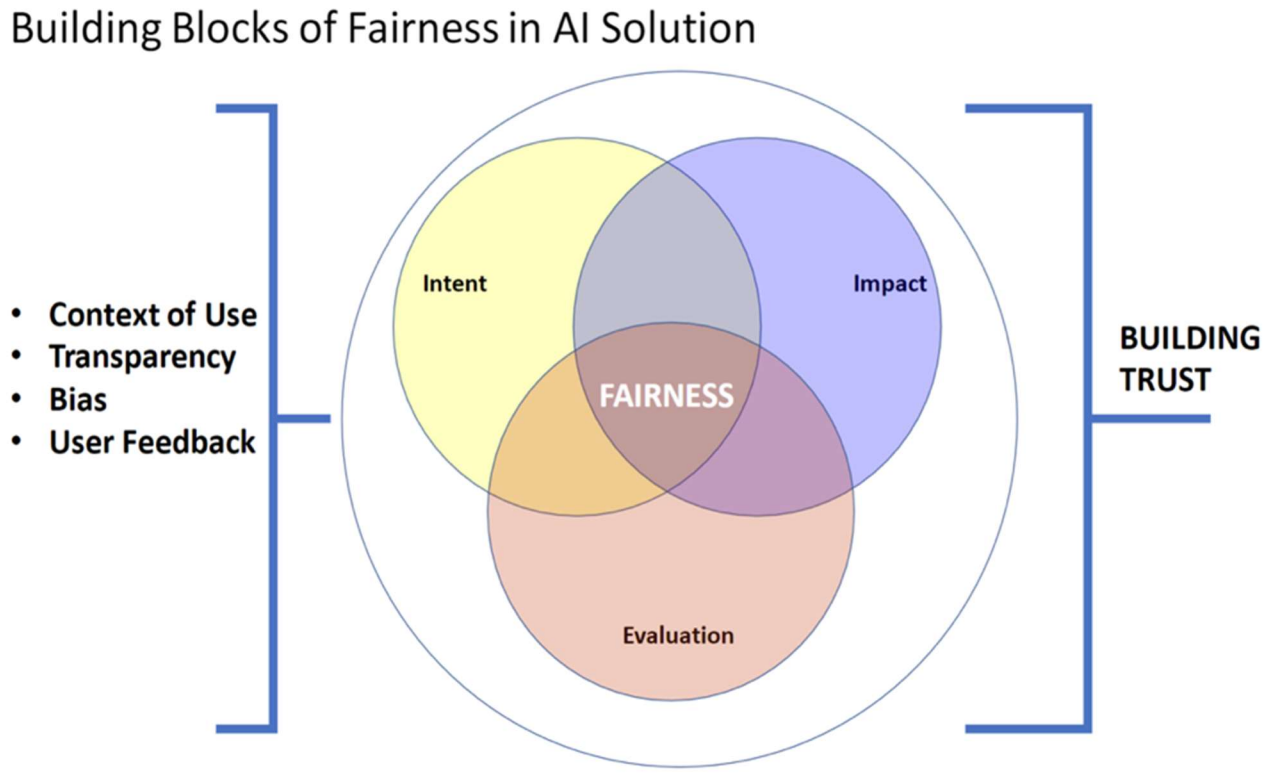


Figure 3: Fairness in AI Solution

Indicators

Indicators identify those areas of practice which best ensure that the technology will be appropriately employed. If monitored closely, these practice areas can significantly reduce the chance of an unfair outcome. From a product development perspective, these indicators also serve as a basis to identify relevant design requirements and critical attributes to identify attributes of fairness in a new AI system across the development, implementation, and operational stages. Limitations of this framework should be considered and evaluated on the basis of application specific constraints, assumptions, and overall impact on intended or unintended outcomes of the AI solution.

Understanding of Context of Use (Scope): The AI solution will adequately answer the problem expressed in the use case.

Review the Fit for Use according to the following criteria:

- What is the goal of the Use Case?
- Does the AI solution seem to adequately answer the problem expressed in the use case?

Level of Diversity **and Inclusion:** The AI solution must demand the greatest levels of diversity and inclusion.

The solution should be reviewed for diversity and inclusion according to the following criteria:

- Criteria that embraces/ensures group, individual, agent, or entity differences, including, but not limited to: age, ethnicity, gender, educational discipline, cultural perspectives, diverse domains of expertise and perspectives, performance capacity, reach, scope, and dimensionality.
- Criteria for fact-based determination of potential biases in decision making, and overall impact on society, behavior, and performance outcomes.
- Criteria that embrace universal ethical values.

Data Indicators: The AI System will provide robust Data Lifecycle management, discriminatory behavior, represent objective and subjective indicators, and provide a data framework that ensures transparency.

The solution should be reviewed per the following criteria:

- **Data Life-Cycle –** A robust approach to fairness is essential at every step of an AI solution:
  - Data Seeding
  - Data Discovery and Acquisition
  - Data Representation and Quality
  - Data Operations, including Text Preprocessing, Understanding, and Language Feature extraction
- **Discriminatory behavior –** Group fairness calls for analysis on statistical parity or equal group error rates for protected groups, while individual fairness says analytics should aim only at accurate predictions.
- **Representative – Objective and subjective indicators** - Includes relevance of historical data, timeliness of data**: Subjective data** are information from the client's point of view (“symptoms”), including feelings, perceptions, and concerns obtained through interviews. **Objective data** are observable and measurable **data** (“signs”) obtained through observation, physical examination, and laboratory and diagnostic testing.
- **Transparency** – Data framework to understand all processes and inherent limitations of data availability: (1) The ability to easily access and work with data no matter where

they are located or what application created them. (2) The assurance that data being reported are accurate and are coming from the official source.

Methodology and Technology Life Cycle Indicators: The methodology and technology by which the AI system is employed will provide the ability to explain outcomes, provide transparency with respect to purpose and function and identify potential.

The solution should be reviewed from the following perspectives:

- **Explainability:** The extent to which outcome is related to its model prediction in such a way that humans understand determining factors.
- **Transparency:** Openness about validation of the purpose, quality of the objective function, structure, and underlying actions.
- **Limitations and Assumptions:** Identify and evaluate in terms of overall risk and impact.
- **Risk and Impact Analysis:** Evaluation of possible outcomes for unfair strategic advantage; and evaluation of intended and unintended outcomes.

Figure 4: AI Lifecycle

Iterative Evaluation and Improvement across Development Life Cycle: Continuous improvement of the AI system will be by an iterative evaluation mechanism designed to evaluate expected and actual outcomes.

A pre-processed training dataset is first introduced into the model. After processing and model building with the given data, the model is tested, and the results are matched with the desired result/expected output. The feedback is then returned to the system for the algorithm to further learn and fine tune its results, repeating until outcomes exhibit minimal variance and have an equal probability of occurrence when similar initial data is input.

Implications

Implications refer to the level of impact on outcomes. Compromised indicators result in compromised outcomes. In this discussion, every indicator has a corresponding implication. For this exercise, the implications are expressed with the "if…, then…" logic.

Context of Use Fit for Use: *If intent and goal are not evaluated, then there is no guarantee of use case is good fit for AI process or not.*

Discrimination: *If level of diversity and inclusion are not addressed during the data collection, then outcome can have adverse social, economic, health, or legally acceptable impact and can result in unfair strategic advantage to specific group.*

Data Indicators: Unintended outcome: *If data does not include objective and subjective data, then result will produce unintended outcome.*

Data Indicator: Unintended consequence: *If auditability and traceability are not considered, then a potential unfair outcome might produce unintended consequence.*

Data Indicator: Unfair Outcome: *If data collection process is not transparent, then outcome might not be trustworthy and/or fair.*

Methodology and Technology Life Cycle: Explainability: *If process and outcome are not explainable, then individual or group might not understand the process, which could impact fair outcomes.*

Continuous Improvement: Design & Development: *If continuous data evaluation is not done and data bias is not identified, then outcome might not be trustworthy and fair.*

Continuous Improvement: Unfair Outcome: *If user feedback is not collected and data bias is not identified, then outcome might not be trustworthy and/or fair.*

If fairness is not implemented correctly, then the AI solution could produce an outcome that provides an unfair strategic advantage or have adverse social, economic, health, or legally acceptable impact.

Monitor and Measure

It is important to monitor and measure the indicators and implications of the components throughout the project lifecycle. Monitoring the AI systems includes auditing for performance of algorithms against key value-driven metrics such as accountability, bias, and transparency. Measurement is tracking smart metrics and KPIs throughout the lifecycle.

Monitor their impact:

- Continuous monitoring with frequent testing for quality and compliance
- User feedback / Implement lesson learned/ Continuous improvement
- Data Bias checking
- Judgment and Interpretability
- Monitor for protected attributes, key considerations
- What is the Risk of AI solution?
    - Impact Analysis –evaluate possible outcomes of AI model
    - How is outcome affecting behavior of stakeholders involved in process?

- What is the societal impact?
  - Is it "Fit for Use"?
  - Does solution result in unfair strategic advantage?

Measures and Index (Qualifiers): How to measure fairness parameters?

- Disparate Index – Preventing unfair strategic advantage/disadvantage - Ensuring equality
  - *Measures of adverse impact on entities/agents/places/events/assets*

- Data Documentation Indices
  - *Data discovery and traceability*
  - *Data prioritization*
  - *Metadata and data lineage*
  - *Data quality*
  - *Policies and processes*
  - *Data privacy and security*
  - *Data retention lifecycle*

- Is outcome satisfactory to stakeholders? Based on iterative process
  - *Goals – Context of Use / Fit for Use*
  - *Version - Training levels*
  - *User feedback*

- Quantitative vs. Qualitative metrics (quantitative evaluation must be verified and validated through a qualitative matrix)
  - *Technical and non-technical evaluation*
  - *Objective vs. Subjective*
  - *Is there enough adequate data to derive a quantitative metric?*
  - *Understanding complex system needs consideration of multiple parameters - (e.g., Spider diagram)*
  - *SMART measures and assessment tools.*

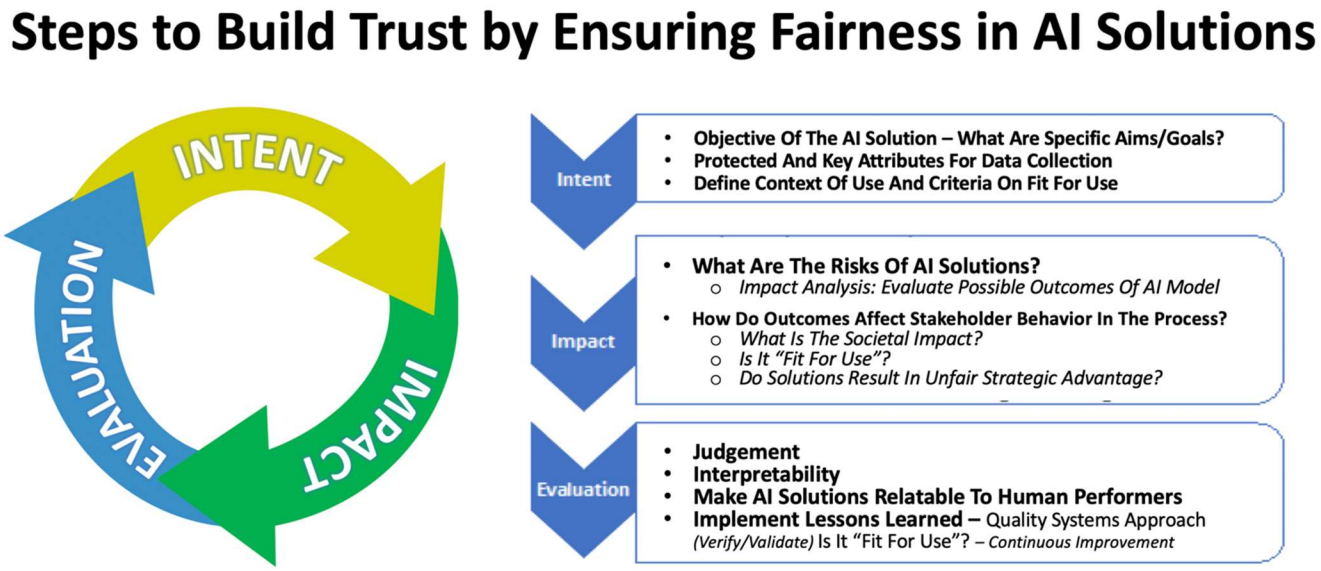


Figure 5: Ensuring Fairness in AI Solutions

	Monitor Implications	Measures Impact
Context of Use	Intent and goal	Intent and goal evaluation; Fit for Use
Diversity	Use cases for various customer personas and from diverse perspectives	Measure outcomes for various personas
Discriminatory behavior	Data collection and distribution	Statistic used in Data distribution and balance; Statistic used in misleading fashion; incomplete or unrepresentative outcome
Representative	Objective and subjective data	Data distribution and balance
Transparency	All processes and data availability	Openness about validation of the purpose, quality of the objective function, structure and underlying actions
Explainability	Process and outcome	Outcome in such a way that humans understand
Iterative Evaluation	Data at every stage of the process	User feedback, outcome with minimum variance and have an equal probability of occurrence
Risk and Impact	Outcome	Unfair strategic advantage, unintended consequence, limitations and assumptions

TRANSPARENCY COMPONENT

Transparency in the context of AI means that AI is explainable to any user, decision maker, or impacted population. As AI becomes increasingly ubiquitous in all aspects of our lives, it is critical to ensure that these AI systems are developed with data that is fair, interpretable, and representative. In the context of AI Ethics, the definition of transparency takes on an additional, sometimes paradoxical, meaning as it is used to infer the trust and reliability upon the AI use case and the decisions that follow.

Transparency in the context of AI is openness about the purpose, structure, and underlying actions of the algorithms used. In an ideal state, AI would be fully transparent to users, decision makers, and those impacted, but this is difficult to achieve in practice given intellectual property concerns. Transparency can be used to identify issues of fairness, bias and trust — all of which have received increased attention. Used effectively, transparency creates a means to instill credibility and establish confidence; explainable AI improves transparency.

Definition

The capability to understand how and why a system arrives at a given outcome.

Indicators

Indicators identify those qualitative areas of practice, which best ensure that the technology will be appropriately employed. If monitored closely, these practice areas can increase confidence and potentially reduce the chance of abuse given any use case. Based on the objective of the AI use case, the following are important indicators for consideration:

Open Data, Architecture and Algorithms: When the results from an algorithm can be tracked back from a data, architecture and algorithmic perspectives, it lends itself to be transparent.

An AI system where the **data** use is open enables transparency. The caveat is in those instances where proprietary data or data with PII needs to be used. Transparency works as a diagnostic to improve your model**.** Models can run away from you and your improvement is limited. Data undergoes complex transformations within the data pipelines that feed AI systems. This results in “data derivatives” and the risk that the initial meaning of transformed data is lost, which could result in it being misinterpreted.

AI systems should use open **architecture** which makes adding, upgrading or swapping components easy and allows users to see inside all or parts of the architecture without any proprietary constraints.

An AI system where the **algorithm** is not proprietary and the logic behind the algorithm can be shared creates transparency in an AI system. Opaque black-box algorithms, such as those used in deep neural networks, incorporate many implicit and highly variable interactions into their

predictions. By contrast, transparent "glass box" algorithms, such as those used for logistic regression, are usually simpler. There is no single approach to understand all algorithms. To aid in the future adoption of deep neural networks (DNN), tools are being developed to address explainability.

Observable: An AI system that is observable has access to relevant information and results in explainability.

Explainability is the extent to which the internal mechanics of an AI system can be explained in human terms and provides the ability to answer the why, how, and what to provide insights into where an algorithm succeeds, fails, and errs, allowing for greater understanding of the process.[3]

Explainability should not be confused with interpretability, which is about the extent to which a cause and effect can be observed within a system. Rather, it is the extent to which you are able to predict what is going to happen, given a change in input or algorithmic parameters.

While Interpretability is about being able to discern the mechanics without necessarily knowing why, Explainability is being able to quite literally explain what is happening in terms of the lifecycle of: Aggregation, Assessments, and Answers.

Implications

Implications refer to the level of impact on outcomes. Compromised indicators result in compromised outcomes. In this discussion, every indicator has a corresponding implication. The probability of a compromised indicator multiplied by the level of impact will produce the index. This index represents the technology's score. For this exercise, the implications are expressed with the "if, then" logic.

[3] https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

Data Veracity: Fit for Purpose –*if there is data that contains implicit racial, gender, or ideological biases, then it can have adverse impacts.*

- Data can contain implicit racial, gender, or ideological biases that distorts reality. It can be poorly researched, with vague and unsourced origins. For some, end results can be catastrophic: Qualified candidates can be disregarded for employment, while others can be subjected to unfair treatment in areas such as education or financial lending. In other words, that age-old saying, "garbage in, garbage out" still applies to data-driven AI systems.

Correlation: Operationally Relevant –*If there is not continuous measuring of the degree of correlation between the indicators to understand how closely aligned AI and decision-making results are with the ethical risks of missions and goals, then the solution is operationally relevant.*

- The way AI and decision-making results demonstrate relationships, correlations, with business process productions is by measurements of correlation. Each individual organization may be characterized by quantitative and qualitative factors as proxy indicators of the degree of goodness of the organization needs, structure, and capabilities. Stakeholders from across the organization can provide assessments or scores of particular AI values contributing to enhanced human values, efficiency, effectiveness and benefits of systems functions, and comprehensiveness or data science lifecycles. By continuously measuring the degree of correlation between the exemplary indicators we can derive a sense for how closely aligned AI and decision-making results will be with the ethical risks of missions and goals.

Process: *If the processes and outcomes are consistent and repeatable, meaning the results for similar queries are the same and can be tracked back, then the solution's process is ethical.*

- Transparency in AI requires processes and outcomes to be consistent and repeatable, meaning the results for similar queries are the same and can be tracked back. Similarly, the processes and outcomes are reliable, traceable, and reproduceable. Reliability is important to any system, regardless of AI or not, but in this context, outcomes must be reliable in the AI for users to have confidence that instills trust.
- An intended audience's confidence in AI increases when it gets a reasonable explanation of how an AI system came to a particular conclusion. The explanation does not have to be deeply scientific, but should address basic risks and the concerns of people who will use the AI system.
- If Transparency is not achieved, then there is increased Risk of AI Gone Wild with potential flaws/issues being realized too late**.** Transparency works as a diagnostic to improve your model**.** Data undergoes complex transformations within the data pipelines that feed AI systems. This results in "data derivatives" and the risk that the initial meaning of transformed data is lost, which could result in it being misinterpreted.

Monitor and Measure

It is important to monitor and measure the indicators and implications of the components throughout the project lifecycle. Monitoring the AI systems includes auditing for performance of algorithms against key value-driven metrics such as accountability, bias, and transparency. Measurement is tracking smart metrics and KPIs throughout the lifecycle.

	Monitor Implications	Measures Impact
Open	Open data, open architecture, open algorithm (to be explainable)	Explainability of the AI
Observable	Access to required information	Credibility. Openness about the structure and underlying actions
Business Rules	Confidence in the process and results	Trust

RESPONSIBILITY COMPONENT

The responsible application of AI demands accountability through appropriate usage. In the AI Ethics dialogue, responsibility has less to do with the technology and more to do with its use. Responsibility deals with the intent of application and protects outcomes from any compromised motive. Responsibility will not address technology and never weighs its feature set. Responsibility weighs heavily on the appropriate use of the technology and how to protect stakeholders and targets from outcomes that do not apply to its mission. The implementation of an AI solution must be relevant to the purpose of the task. It must ensure that both data and model sources are uncompromised. It must produce repeatable, legal, authentic, auditable, and effective results.

Definition

Responsibility in the ethical use of artificial intelligence ensures accountability for the application and use of the technology and its resulting impact and consequence due to outcomes.

Indicators

Indicators identify those qualitative areas of practice which best ensure that the technology will be appropriately employed. If monitored closely, these practice areas can potentially reduce the chance of abuse given any use case. For this exercise, indicators are expressed in user story format.

Purpose – The AI System will only be used in accordance to its designed purpose in support of the mission.

AI is a family of technologies, each with purpose and targeted function. Solutions may use one or a combination of these technologies to a single application designed for a specific purpose. The goal is to focus the portfolio of apps and tools to support its stakeholder(s). Trying to apply an application portfolio to support another stakeholder's requirements, regardless of how similar, will likely produce unknown outcomes, some of which could prove harmful.

Privacy – The user(s) of the AI system must protect the privacy of all stakeholders. Personally identifiable information (PII) cannot be compromised without detriment to the individual or parties impacted by its use.

Protecting individuals' privacy is a matter of Public Law 93-579, known as the Privacy Act of 1974. Even if "security by design" practices are foundational in the engineering of AI tools and utilities, threats outside of the technology must be considered." Again, this section deals with the behavior of the technology's usage; thus, threats perimeter to the technology are exploited through simple mechanisms such as malicious code, social engineering, and improper use. Users of AI must follow sound cyber practices to protect credentials, remote access, removable media, mobile devices, email, social networking, and the like.

Pedigree – The application of the AI system will first ensure data veracity before consuming data for usage.

It is not the function of the AI to guarantee the pedigree of the data it is consuming for analysis. However, it is the responsibility of its users to ensure that input data is both untampered with and all-inclusive. All participating stakeholders and affected groups must be represented within the data sets. Finally, the source of the data must be verified before it can be declared appropriate for use. Purity, completeness and the validation of the source of the data produces repeatable and trustworthy outcomes.

Provenance – The user(s) of AI will audit the evolution and maturation of data and model structures.

Pedigree, above, deals with the veracity of the data. Provenance is the foundation for auditable changes to the data, whether the data is consumable or algorithmic. In short, it is proving its authenticity. Coupling technology, such as blockchain, may provide users the ability to monitor and measure the potential impact from malicious alteration and automate auditability of planned change.

Implications

Implications refer to the level of impact on outcomes. Compromised indicators result in compromised outcomes. In this discussion, every indicator has a corresponding implication. The

probability of a compromised indicator multiplied by the level of impact will produce the index. This index represents the technology's score. For this exercise, the implications are expressed with the "if, then" logic.

Applicability – *If the AI system is used for a purpose other than the mission it was designed for, then the applicability of its output is unknown.*

Legality – *If the AI system cannot maintain the privacy of the individual(s) under surveillance, then legal aspects of the system are challenged.*

Authenticity – *If the data lacks authenticity, then confidence in outcomes is misplaced, negatively impacting the mission.*

Auditability – *If the AI system lacks the ability to be audited, to include changes to data, model design, testing, and acceptance, then its behavior (including results produced) will not be reliable.*

Monitor and Measure

It is important to monitor and measure the indicators and implications of the components throughout the project lifecycle. Monitoring the AI systems includes auditing for performance of algorithms against key value-driven metrics such as accountability, bias, and transparency. Measurement is tracking smart metrics and KPIs throughout the lifecycle.

As one considers the complete Artificial Intelligence System Lifecycle, they can determine the dimensions of responsibility for its proper application. These include accountability, participation, security, and clarity of the data that feeds the AI system.

	Monitor Implications	Measures Impact
Purpose	Only used as **Intended** in support of the mission	Opportunity for abuse
Pedigree	Data purity	credibility
Provenance	**Confidence** in the ability to Audit	accountability
Privacy	Identity protection	legality

INTERPRETATION COMPONENT

The results of an AI solution must be operationally relevant and focused upon the goals and objectives of the organization in their ongoing efforts to serve their specific mission. The results of leveraging and applying AI capabilities must ensure that both data and model sources are uncompromised, produce a consistent result, and are reliable and trustworthy.

By ensuring outcomes are consistent, the AI application will instill credibility and establish credibility with all who utilize their services. To achieve this, the outputs must be clear and free from ambiguity and assure it is consistently understood by a variety of people. This final step is a culmination of all that goes into the EAAI by ensuring the data that feeds the algorithm, the means upon which it analyzes information, and the resulting assessments that produce the knowledge base provide consistent understanding given current circumstances.

Definition

Provide a consistent perspective is produced as a result of the AI technology.

Indicators

Causality/Influences- understand the relational dependencies and how they are interpreted.
Correlation/Effect - ascertain how the culmination of analysis produces answers.
Consequences/Impact - understand how the insights influence and affect the environment.
Consistency/Reliable - assess how the predictability of the outcomes provide repeatable results given the same circumstances.

Implications

Inferences/Associations - *identify how the results are susceptible to one's perspective.*

Implications/Answer - *understand how the outcomes influence current perceptions.*

Impacts/Actions - *identify how results will generally inform and transform paradigms.*

Credible/Confidence - *data assessments to inform analysis that ascertains knowledge.*

Monitor and Measure

It is important to monitor and measure the indicators and implications of the components throughout the project lifecycle. Monitoring the AI systems includes auditing for performance of algorithms against key value-driven metrics such as accountability, bias, and transparency. Measurement is tracking smart metrics and KPIs throughout the lifecycle.

Given the past frames of reference influence individual perceptions and current circumstance impact their perspective that inform future paradigms. Thus, how things are interpreted given their past experiences will impact their view of how the outcomes will inform the outcomes of

the AI tool. Thus, the results should inform perceptions and transform paradigms to collectively and consistently evolve perspectives.

	Monitor Implications	Measures Impact
Perceptions	**Causality/Effects**	Inferences/Associations
Paradigms	**Correlation/Enlighten**	Impact/Answers
Perspective	**Cognition/Empower**	Insights/Awareness
Provenance	**Consequences/Evolve**	Implications/Ascertain

The EAAI Scorecard

This section shows how the different components of the framework can be assembled into a scorecard index. In this example, the EAAI index follows the following formula:

$$EAAI\ Index = Sum\ (Component\ Score)$$

$$Component_i\ Score = avg(Component_i\ Indicators\ Score) * avg(Component_i\ Implications\ Score)$$

For each component, the indicators are measured from 1 - 10 and the implications are measured from 1 - 5. The EAAI score would range from 5 – 250.
The following table shows a possible way the Bias and Fairness components go from components to indicators to indices.

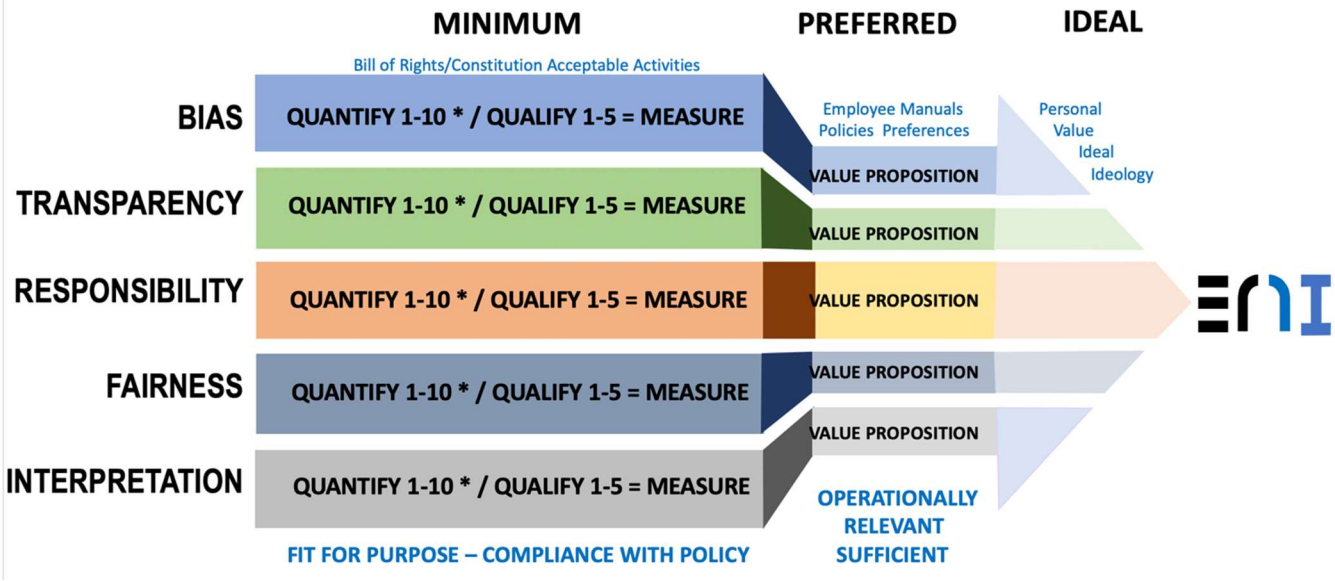


Figure 6: Ethical Index for AI

- Counterfactual fairness is a researched methodology that helps consider protected classes and how to compensate for biases effectively.
- Quantitative vs. Qualitative metrics (quantitative evaluation must be validated and justified by some qualitative one.)

Component of Bias and Fairness	Indicators	Measures / Indices
Intent Are the goals and outcomes of the AI system clear	**Understanding of Context of Use (Scope)** Criteria of Fitness for Use – What is the goal of the Use Case?	**Quantitative vs. Qualitative metrics** (quantitative evaluation must be validated and justified by some qualitative one.) • Technical and non-technical evaluation • Objective vs. Subjective • Is there enough adequate data to derive a quantitative metric? • Understanding complex system needs consideration of multiple parameters - (e.g., Spider diagram) • Surveys and other assessment tools Is outcome satisfactory to stakeholders? Based on iterative process • Goals – Context of Use / Fit for Use • Version - Training levels • User feedback
Intent **Design Methodology**	**Inclusive representation**	**Disparate Index** – Identify unfair strategic advantage/disadvantage. Monitor equality – Measures of adverse impact on entities/agents/ places/events/assets. • Social context example: (Total positive for under privileged group / total instance for under privileged group) / (Total positive for privileged group / total instance for privileged group) (Fairness Metrics Overview[4])
Intent **Outcomes**	**Bias influences decisions and hence impacts any intended outcomes**	**Disparate Inde**x – Preventing unfair strategic advantage/disadvantage - Ensuring equality – Measures of adverse impact on entities/agents/ places/events/assets

[4] https://cloud.ibm.com/docs/services/ai-openscale?topic=ai-openscale-anlz_metrics_fairness

		• Social context example: (Total positive for under privileged group / total instance for under privileged group) / (Total positive for privileged group / total instance for privileged group) (Fairness Metrics Overview[5])
Design Methodology	**Level of diversity and inclusion** • Criteria that embraces/ ensures team member differences, including, but not limited to: Ages, ethnicities, genders, educational disciplines, and cultural perspectives, multi-domain team members with different perspectives. • Engage in fact-based conversation about potential biases in human decisions. • Human-centric design using design thinking principles of Empathize, Define, Ideate, Prototype, Test	**Quantitative vs. Qualitative metrics** (quantitative evaluation must be validated and justified by some qualitative one.) • Technical and non-technical evaluation • Objective vs. Subjective • Is there enough adequate data to derive a quantitative metric? • Understanding complex system needs consideration of multiple parameters - (e.g., Spider diagram)
Design Methodology	**Data Indicators** • Consider *The Open Government Data Life-Cycle*[6] as it relates to AI solution • Unbiased - Statistical fallacies (**need to collaborate with Bias Subgroup) • Representative – Objective and subjective indicators - Includes ladder of inference and relevance of historical data, timeliness of data • Transparency – Framework to understand all processes and inherent limitations of data availability	**Data Indicators** • Data collection • Data accountability • Data prioritization • Metadata and data lineage • Data quality assessments • Policies and processes • Data privacy and security • Data retention
Design Methodology	**Algorithm Indicators** • Explainability • Limitations and assumptions of the technology	**Quantitative vs. Qualitative metrics** (quantitative evaluation must be validated and justified by some qualitative one.) • Technical and non-technical evaluation

[5] https://cloud.ibm.com/docs/services/ai-openscale?topic=ai-openscale-anlz_metrics_fairness
[6] https://www.researchgate.net/figure/Open-Government-Data-Life-Cycle_fig2_281349915

	• Data requirements
	• Consider thru system development life-cycle
	• Accountability
	  ○ Responsibility
	  ○ Auditability
	  ○ Traceability
		• Objective vs. Subjective
		• Is there enough adequate data to derive a quantitative metric?
		• Understanding complex system needs consideration of multiple parameters - (e.g., Spider diagram)
		
		Is outcome satisfactory to stakeholders? Based on iterative process
		• Goals – Context of Use / Fit for Use
		• Version - Training levels
		• User feedback

Design Methodology Outcomes	**What is the Risk of AI solution?** • Impact Analysis –evaluate possible outcomes of AI model • How is outcome affecting behavior of stakeholders involved in process? • What is the societal impact? • Is it "Fit for Use"? • Does solution result in unfair strategic advantage?	**Risk and Impact Analysis** • Evaluation of possible outcomes for unfair strategic advantage • Evaluation of intended and unintended outcomes (include stakeholders in manual evaluation of outcomes)

Conclusion

This framework and related documentation aim not to be prescriptive, but to strengthen awareness of the techniques and methods available to develop and implement AI ethically and optimally. The framework is a joint effort between government and industry, draws from the existing body of work published by the ACT-IAC Artificial Intelligence Working Group, and is designed with the federal government in mind.

As a reminder, the cross-functional ACT-IAC collaborators who contributed to this and related documents recognize that many papers, guides, and efforts on ethics in AI exist. A recent MIT review identifies several initiatives in defining the ethical use of AI. As new information and viewpoints arise, the Working Group welcome and strongly encourage the broader community of informed and impacted voices and organizations to reach out and help keep this framework adaptive and evergreen to maximize its value to government and its stakeholders. For additional information or to provide thoughts, visit the ACT-IAC website at www.actiac.org, email ACT-IAC@actiac.org, or phone (703) 208-4800.

Acknowledgement

The Artificial Intelligence Working Group thanks the authors and contributors who provided a tremendous amount of time, hard work, and good humor to bring this document to completion. The Working Group would like to also thank all our government, industry, and academia collaborative partners who provided invaluable feedback as reviewers.

Authors and Affiliations

This paper was written by a consortium of government and industry. The organizational affiliations of the authors and contributors are included for information purposes only. The views expressed in this document do not necessarily represent the official views of the individuals and organizations that participated in its development.

Adelaide O'Brien	IDC Government Insights
Alex Rebo	United States Government
Ansgar Koene	EY
Ashley Casovan	Ai-Global
Chakib Chraibi	NTIS-Department of Commerce
David Hernandez	Riva Solutions
Don Lovett	DC Government
Dr. Jaunita Stewart	Department of the Army, Office of the Deputy Assistant Secretary of the Army for Defense Exports and Cooperation
Eric Eskam	U.S. General Services Administration, IT Category Office
Fatima Akhtar	IBM
Frederic de Vaulx	Prometheus Computing LLC
G. Hussein Basaria	The Maven Group
George A. Tilesch, JD, MA	Global AI Expert, Co-Author: Between Brains - Taking back our Future in the AI Age
Henry Jia	Excella Consulting
Jeremy Wood	Millennium Challenge Corporation
Jessica Davis	Microsoft
Jignesh Shah	S2Alliance Inc
Joe Paiva	Hire Vue
June Lau	National Institute of Standards and Technology
Ken Farber	AIexpertise
Marc Wine	Department of Veterans Affairs
Michael Bruce	Leidos
Michelle White	U.S. General Services Administration, IT Category Office
Mike Rice	CornerStone

Neil (Anil) Chaudhry	U.S. General Services Administration, Technology Transformation Office
Nevin Taylor	National Artificial Intelligence Institute
Orlando Lopez	National Institutes of Health
Sandy Barsky	United States Government
Swathi Young	Integrity Management Services, Inc.
Tim Gilday	GDIT
Timothy George	The Maven Group
Todd D. Lyle	Council for Cyber Risk Control
Todd Hager	Macro Solutions
Tracy Lynn Jones	Grant Thornton
William D. Radcliff	National Institute of Standards and Technology

References

Bias

https://ai-global.org/about/
https://www.weforum.org/agenda/2019/07/ai-driven-companies-need-to-be-more-diverse-here-s-why/
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Organization/Our%20Insights/Delivering%20through%20diversity/Delivering-through-diversity_full-report.ashx
https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Fairness

https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf
https://arxiv.org/abs/1810.01943
https://arxiv.org/abs/1809.09245
https://www.go-fair.org/fair-principles/
https://drive.google.com/file/d/1BFh5KHoYWbH-OLkxfMKlGPXPeFc6vLa5/view
https://arxiv.org/pdf/1901.04562.pdf
https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb
https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html
https://pair-code.github.io/what-if-tool/ai-fairness.html
https://towardsdatascience.com/artificial-intelligence-fairness-and-tradeoffs-ce11ac284b63
https://fairmlbook.org/tutorial2.html
https://www.forbes.com/sites/googlecloud/2020/01/15/what-does-fairness-in-ai-mean/#3db0eb961574
https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
https://www.brookings.edu/research/fairness-in-algorithmic-decision-making/
https://towardsdatascience.com/the-importance-of-ethics-in-artificial-intelligence-16af073dedf8
https://cloud.ibm.com/docs/ai-openscale?topic=ai-openscale-anlz_metrics_fairness
https://cloud.ibm.com/docs/services/ai-openscale?topic=ai-openscale-quality_group
https://www.georgialegalaid.org/resource/fair-treatment-by-the-government-equal-protec
https://www.congress.gov/bill/116th-congress/house-bill/2202

https://www.whitehouse.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/
https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence
https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf