

The logo for the 2024 Embedded VISION Summit is centered on the left side of the slide. It features a white octagonal background with a colorful, multi-layered border in shades of purple, blue, green, yellow, and orange. The text "2024" is at the top, "embedded" is below it, "VISION" is in large, bold, dark blue letters with a gradient, and "SUMMIT" is at the bottom.

2024
embedded
VISION
SUMMIT®

Identifying and Mitigating Bias in AI

Nikita Tiwari

AI Enabling Engineer, Client Computing
Group

Intel Corporation

The Intel logo is located in the bottom right corner of the slide. It consists of the word "intel" in a lowercase, blue, sans-serif font, enclosed within a white square.

intel

- Why Should we Care About Responsible AI (RAI)?
- Identifying Bias in AI
- Tools to Mitigate Bias in AI
 - Fairness Metrics
 - What-If Tool
 - AI Fairness 360
- Key Takeaways
- Industry-Wide Ethical AI Revolution & Resources

Discussion



Ethics is a conversation!



AI stories you have come across (positive & negative)




Why should we care?


Why Responsible AI?



Massive global AI market size



Daily reports of AI harm



Generative AI accessible to all end users



Barriers in implementing trustworthy and explainable AI

Cost of AI Incidents

Harm to human life

Loss of trust

Fines in compliance & regulations

Introduction of systemic bias

Misinformation

Breach of privacy

AI Incidents in the News

A news site used AI to write articles. It was a journalistic disaster.

Secretive Algorithm Will Now Determine Uber Driver Pay in Many Cities

Oracle's 'surveillance machine' targeted in US privacy class action

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

Tesla Model 3 Taxi Cab Accident Hurts About 20 People in Paris Due to Braking Issues

How ChatGPT can turn anyone into a ransomware and malware threat actor

BBC fools HSBC voice recognition security system

<https://incidentdatabase.ai/>

Defining Responsible AI Principles

Respect Human Rights

Human rights are a cornerstone for AI development. AI solutions should not support or tolerate usages that violate human rights.

Equity and Inclusion

Focus on data used for training and the algorithm development process to help prevent bias and discrimination.

Transparency

Understanding where the data came from and how the model works.

Enable Human Oversight

Human oversight of AI solutions to ensure they positively benefit society.

Personal Privacy

Maintaining personal privacy and consent. Focusing on protecting the collected data.

Security, Safety, Sustainability

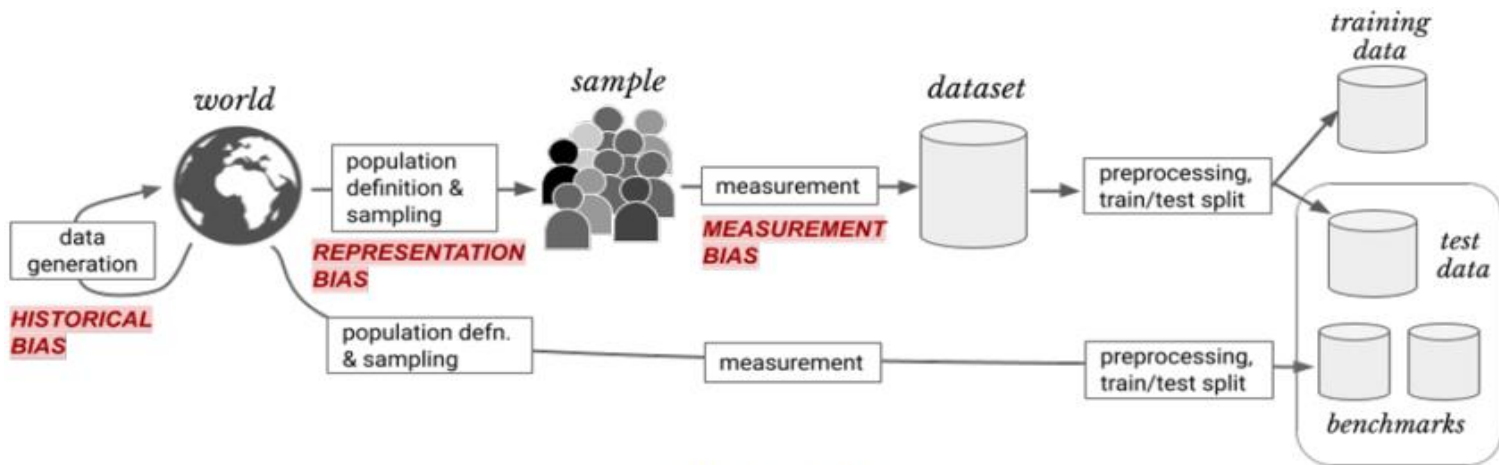
Ethical review and enforcement of end-to-end AI safety. Low-resource implementation of AI algorithms

Bias Identification

Bias in Data Generation

Historical Bias:

- Caused by preconceived notion even on perfectly measured/sampled data
- E.g., Gendered occupation
- Case study: Amazon hiring algorithm/ Google Gemini AI image generation



(a) Data Generation

Representation Bias:

- Target population does not:
 - reflect the user population
 - include underrepresented groups
- Sampling method is limited or uneven
- Two-fold representation – uniform vs proportional
- E.g., ImageNet images

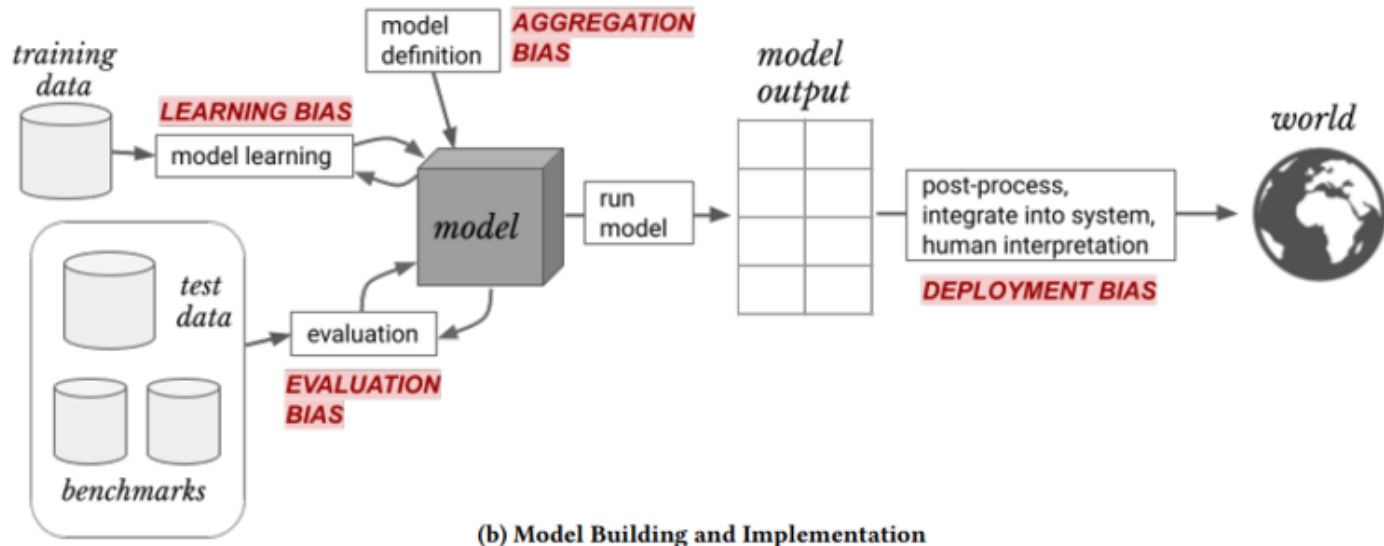
Measurement Bias:

- The proxy oversimplifies a complex construct, or measurement methods and accuracy differ among groups
- E.g., COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

Bias in Model Building and Implementation (1/2)

Aggregation Bias

- Arises when data from different sources or groups are combined, leading to distortions in the model's performance or predictions
- E.g., Housing price prediction model trained on data aggregated from multiple cities without accounting for differences in housing markets



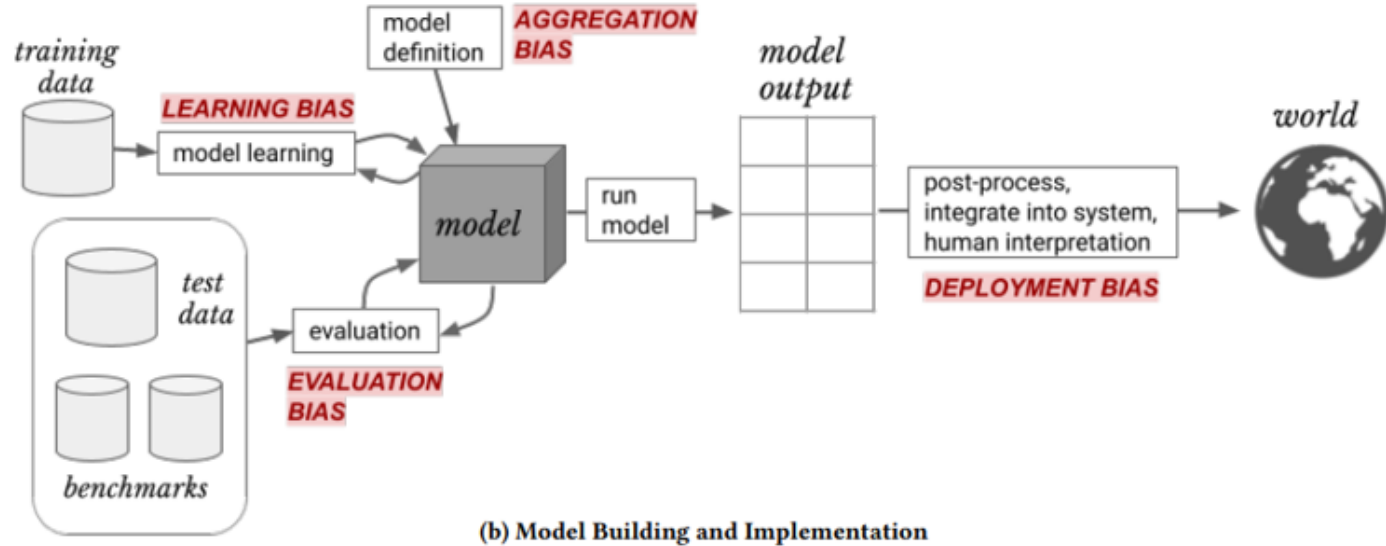
Learning Bias

- Arises when modeling choices amplify performance disparities across different examples in the data
- E.g., Prioritizing one objective damages the other, like optimizing for privacy or compactness may affect accuracy

Bias in Model Building and Implementation (2/2)

Evaluation Bias

- The benchmark data or the evaluation process used for a particular task does not represent the user population or favors certain groups over the others
- E.g., A language translation model evaluated primarily on its accuracy for European languages, hinders its usefulness in diverse contexts as a global language translator



Deployment Bias

- Arises due to “off-the-label” usage of model other than intended use
- E.g., Models intended to predict a person’s likelihood of committing a future crime also used to determine the length of the sentence

Bias Mitigation

Common Fairness Metrics

- Fairness metrics are a set of measures that enable you to detect the presence of bias in your data or model
- At least 21 fairness metrics. Many are conflicting. Which is the fairest? No right answer. Some common metrics -
 - **Group unaware** – Removes all group and proxy-group membership information from the dataset to avoid favoring any groups or sub-class. Similar logic as unsupervised learning. Difficult to achieve.
 - **Group threshold** – Alternate thought-process to group unaware. Adjust the confidence thresholds for different groups independently such that the confidence threshold for correct predictions for a minority group will be slightly lower.
 - **Demographic parity** – Similar percentages of datapoints from each group are predicted as positive classifications. E.g., A class with x% of subclass-1 should have x% of subclass-1 positive predictions.
 - **Equal opportunity** – Among those datapoints with the positive ground truth label (true positive rate), there is a similar percentage of positive predictions in each group.
 - **Equal accuracy** – The model's predictions are equally accurate for all groups. True positive and false positive should be same across all groups.

Tools to Detect and Mitigate Bias

- Key challenge in developing and deploying a ML system is understanding their performance across a wide range of inputs
- Several open-source tools utilize fairness metrics and bias mitigation algorithms to analyze ML systems with limited coding and test bias in hypothetical scenarios

What-If Tool (Google)

- Simulation with data manipulation & specific criteria to detect bias
- 5 fairness metrics
- Bias mitigation not straightforward

AI Fairness 360 (IBM)

- Extensible toolkit for bias detection & mitigation
- 70+ fairness metrics
- 10 bias mitigation algorithms
- Fairness metric explanations

Key Takeaways and Resources

Key Takeaways

- Establish RAI principles that guide the decision-making for your AI development
- Drive RAI requirements into product definition
- Adopt human-centric approach at every stage of your product development
- Integrate RAI tools in your software development lifecycle
- Preventing bias is complex. Define fairness metrics, document trade-offs and share with your users transparently. Re-check for bias often.
- Conduct regular assessments, audits, and update AI response plans
- Keep up-to-date with the evolving legislature, regional laws and standards
- Certifications can help adherence to standards, legislatures and build user trust

Industry-Wide Ethical AI Revolution

- **Responsible Artificial Intelligence Institute (RAII)**
 - First independent, accredited certification program for RAI in US, Canada, Europe and UK
 - Vectors: Systems operations, explainability and interpretability, accountability, consumer protection, bias and fairness, and robustness with collaborations across world economic forum, OECD, IEEE, ANSI, etc.,
- **Executive orders on responsible AI around the globe**
 - [European Union AI act](#)
 - [Executive Order on the Safe, Secure and Trustworthy Development and use of AI](#)
 - [The White House Blueprint for an AI Bill of Rights](#)
- **IEEE Standards Association**
 - <https://standards.ieee.org/participate/>
 - 2100+ Standards, 175+ Countries, 34000+ Global Participants
- **IEEE CertifAIEd™ (part of Standards Association)**
 - A certification program for assessing ethics of Autonomous Intelligent Systems (AIS)
 - Vectors: Ethical Privacy, Algorithmic Bias, Transparency, and Accountability, Agentic AI
 - AI Safety Certification/Assurance Development
 - Case Study for [AI ethics applied to the city of Vienna](#)

Resources

Responsible AI Landscape

<https://hai.stanford.edu/news/2022-ai-index-industrialization-ai-and-mounting-ethical-concerns>

Grandview Research

<https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market>

Global AI adoption Index

https://filecache.mediaroom.com/mr5mr_ibmnewsroom/191468/IBM%27s%20Global%20AI%20Adoption%20Index%202021_Executive-Summary.pdf

Measuring Bias – David Weinberger

<https://pair-code.github.io/what-if-tool/ai-fairness.html>

Intel Responsible AI Program

<https://www.intel.com/content/www/us/en/artificial-intelligence/responsible-ai.html>



Responsible AI Institute Certification

<https://www.responsible.ai/how-we-help>

IEEE CertifAIEd™

<https://engagestandards.ieee.org/ieeecertifaiied.html>

AI Incident database

<https://incidentdatabase.ai/>

European Union AI Act

https://ec.europa.eu/commission/presscorner/detail/en/IP_23_6473

White House Executive Order

<https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>

Blueprint for AI Bill of Rights

<https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

Questions

Backup Material

Designing With a Human-Centric Approach



Definition

Does AI add value?

Who are the intended users of the system?

Identify intended potential harm and plan for remediations

Translate user needs into data needs

E.g., Prototyping a chatbot



Development

Source high-quality unbiased data responsibly

Get inputs from domain experts

Enable human oversight

Built-in safety measures

E.g., Improving autonomous vehicles



Deployment

Provide ways for users to challenge the outcome

Provide manual controls when AI fails

Offer high-touch customer support



Marketing

Focus on the benefit, not the technology

Transparently share the limitations of the system with the users

Be transparent about privacy and data settings

Anchor on familiarity

Google Gemini AI Image Generation Mistake

Gemini image generation got it wrong. We'll do better.

Prabhakar Raghavan
Senior Vice President, Google
Feb 23, 2024

What happened?

Three weeks ago, we launched a new [image generation](#) feature for the [Gemini conversational app](#) (formerly known as Bard), which included the ability to create images of people.

It's clear that this feature missed the mark. Some of the images generated are inaccurate or even offensive. We're grateful for users' feedback and are sorry the feature didn't work well.

We've [acknowledged the mistake](#) and temporarily paused image generation of people in Gemini while we work on an improved version.

Why did the incident happen?

So what went wrong? In short, two things. First, our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly *not* show a range. And second, over time, the model became way more cautious than we intended and refused to answer certain prompts entirely – wrongly interpreting some very anodyne prompts as sensitive.

These two things led the model to overcompensate in some cases, and be over-conservative in others, leading to images that were embarrassing and wrong.

Remediation and next steps

This wasn't what we intended. We did not want Gemini to refuse to create images of any particular group. And we did not want it to create inaccurate historical – or any other – images. So we turned the image generation of people off and will work to improve it significantly before turning it back on. This process will include extensive testing.



Diverse Skin Tone Recognition with Intel Evo Laptop

- An AI algorithm was used to recognize when person moves away from laptop and to turn off the screen
- The algorithm was tested to be inclusive and performant on individuals with different skin tones, to ensure the algorithm is fair and the output is not affected.



Pedestrian Detection including disabled individuals

- Pedestrian detection for self-driving cars should incorporate diverse data, including data from disabled pedestrians, such as folks in wheelchairs

Open-source Fairness Metrics Libraries

Open-source library	Notes
AIF360	Provides a comprehensive set of metrics for datasets and models to test for biases and algorithms to mitigate bias in datasets and models.
Fairness Measures	Provides several fairness metrics, including difference of means, disparate impact, and odds ratio. It also provides datasets, but some are not in the public domain and require explicit permission from the owners to access or use the data.
FairML	Provides an auditing tool for predictive models by quantifying the relative effects of various inputs on a model's predictions, which can be used to assess the model's fairness.
FairTest	Checks for associations between predicted labels and protected attributes. The methodology also provides a way to identify regions of the input space where an algorithm might incur unusually high errors. This toolkit also includes a rich catalog of datasets
Aequitas	This is an auditing toolkit for data scientists as well as policymakers; it has a Python library and website where data can be uploaded for bias analysis. It offers several fairness metrics, including demographic, statistical parity, and disparate impact, along with a "fairness tree" to help users identify the correct metric to use for their particular situation. Aequitas's license does not allow commercial use.
Themis	An open-source bias toolbox that automatically generates test suites to measure discrimination in decisions made by a predictive system.
Themis-ML	Provides fairness metrics, such as mean difference, some bias mitigation algorithms, additive counterfactually fair estimator, and reject option classification.
Fairness Comparison	Includes several bias detection metrics as well as bias mitigation methods, including disparate impact remover, prejudice remover, and two-Naive Bayes. Written primarily as a test bed to allow different bias metrics and algorithms to be compared in a consistent way, it also allows additional algorithms and datasets.