Bias in Artificial Intelligence and Mitigation Strategies

The significant advancements in applying artificial intelligence (AI) to various domains have raised concerns about the fairness and bias of AI systems. Responses from such systems can result in unfair outcomes and carry forward existing inequalities. Drawing the line between using AI for decision making and avoiding accusations of bias requires a combination of transparency, fairness and accountability. This is an attempt to review aspects of these biases, general and specific strategies that can be employed to mitigate such biases.

Bias in Artificial Intelligence refers to systematic errors in an AI system that can lead to unfair, prejudiced, or unbalanced outcomes. These biases often reflect and amplify societal inequalities present in the data used to train AI models. Bias can manifest in various ways, affecting different groups unfairly based on factors like race, gender, age, or socio economic status.

Drawing the line between using AI for decision making and avoiding accusations of bias requires a combination of transparency, fairness, and accountability. Here, we look at the various aspects of bias and how they can be mitigated by choice of appropriate strategies to overcome each of them.

Bias classification

Here's an overview of the aspects of bias and how they contribute to the overall problem of intervening with AI outcomes. Doing this helps us understand the basis for the bias and the problems it can cause and lead us to strategies in mitigating them.

- **Data Bias**
  Data bias occurs when the training data is not representative of the real world population causing skewness. E.g. A bank is using AI to predict loan repayment capacity of a person before approving his loan application, if the trained model reflects historical biases (gender, race etc.), the results would be discriminatory.

- **Algorithmic Bias**
  This happens when an AI model's design or mathematical assumptions lead to biased outcomes. E.g. If the loan repayment capacity predictor model algorithm uses address as a feature, people from certain neighbourhoods which may be racially segregated could be unfairly penalized with their applications getting rejected.

- **Selection Bias**
  Selection bias arises when the data used to train the AI does not fully capture the diversity of the population. E.g. In the loan repayment capacity predictor, since the bank may not have enough data about young people or immigrants, suggesting their applications to be reject.

- **Confirmation Bias**
  Confirmation bias is when AI systems reinforce existing beliefs rather than presenting diverse perspectives. E.g. In the loan repayment capacity predictor, since the bank may have a lot of data about people with account with them for a long time, the loan applications of only such customers would get recommended for approval.

- **Automation Bias**
  Automation bias is when users overly rely on AI decisions, assuming they are always correct, without questioning potential errors. E.g. In the loan repayment capacity predictor, if a human doesn't review all recommendations from the system and chooses to accept them directly, it would result in automation bias.

- **Measurement Bias**
  Measurement bias is when incorrect or unfair labels are used in training data – when the correlations are incorrectly assumed. E.g. A credit scoring AI that uses ZIP codes as a proxy for financial reliability may disadvantage lower income neighbourhoods.

General strategies in overcoming bias

The following is a listing of general strategies which can be followed in overcoming various bias.

- **Understand and Mitigate Bias in Data**
  Bias often comes from the data used to train AI models. Ensure that training data is diverse, representative, and regularly audited for disparities. Use fairness metrics to detect and mitigate bias in AI outcomes.

- **Use Explainable AI (XAI)**
  AI decisions should be interpretable. Use explainability tools (e.g., SHAP, LIME) to provide reasons behind AI driven outcomes. Make AI generated recommendations easy to understand for stakeholders.

- **Human Oversight and Review**
  AI should assist, not replace, human decision makers, especially in high stakes scenarios (e.g., hiring, lending, medical diagnosis). Implement "human in the loop" processes where necessary.

- **Transparency and Documentation**
  Clearly communicate how AI models work, including data sources, methodologies, and limitations. Provide documentation on AI decision making frameworks, ensuring regulators and stakeholders can assess them.

- **Regular Audits and Bias Testing**
  Conduct ongoing audits of AI

models to identify and correct any emerging biases. Use fairness tests (e.g., disparate impact analysis) to measure bias in outputs.

- **Ethical Guidelines and Governance**
  Establish AI ethics policies aligned with regulatory frameworks (e.g., GDPR, AI Act). Set up an AI ethics committee to review sensitive AI applications.

- **User Feedback and Redress Mechanisms**
  Allow affected individuals to challenge AI decisions and provide alternative inputs. Create pathways for users to report potential biases or errors in AI outputs.

Specific strategies to overcome various types of bias

Next, we dwell into specific strategies using which bias can be overcome in AI – addressing each of the types of bias specifically

Data Bias

When the data used to train AI models is not representative of the real world population, it leads to biased outcomes. Mitigating bias in training data is crucial for building fair and ethical AI systems. Bias in AI isn't intentional and can have serious consequences. It may not be able to completely avoid bias in training data. However, by actively managing them, one can build AI systems that are fair, ethical, and trustworthy. Here are some strategies to mitigate bias in data.

- **Collect Diverse and Representative Data**
  Ensure training data includes a broad range of demographic groups (e.g., race, gender, age, geography, socioeconomic status).

  Avoid datasets that are skewed or overrepresent certain populations at the expense of others.

  Use multiple data sources to capture different perspectives. e.g. In facial recognition, include images of people from different ethnic backgrounds, lighting conditions, and facial expressions to improve accuracy across groups.

- **Preprocess Data to Reduce Bias**
  Reweighting: Assign higher weights to underrepresented data points to balance the dataset.

  Resampling: Over sample minority groups or under sample majority groups to create a balanced dataset.

  Synthetic Data Generation: Use AI techniques (e.g. GANs) to create synthetic but realistic data for underrepresented groups

- **Remove Proxy Variables that Introduce Bias**
  Some features act as proxies for sensitive attributes like race, gender, or income level. Example of problematic variables: ZIP codes (which correlate with race/income), names (which can imply gender/ethnicity), or college names (which may favor privileged applicants).

  Run statistical tests to detect unintended correlations between input features and protected attributes.

- **Use Fairness Aware Algorithms**
  Train AI models with fairness constraints, ensuring they do not favor or disadvantage specific groups.

  Implement techniques like: • Adversarial debiasing – Train the model to minimize bias while maximizing accuracy

  Fair Representation Learning – Encode data in a way that removes discriminatory factors.

  Fair Regularization – Add penalties for biased outcomes in the model's loss function.

Conduct Regular Bias Audits and Testing
Perform bias detection tests before and after model training.

- **Demographic Parity:** Check if different groups receive similar predictions.

- **Equal Opportunity Testing:** Ensure all groups have equal chances of positive outcomes.

- **Disparate Impact Analysis:**
  Verify that no single group is disproportionately at advantaged or disadvantage.

- **Ensure Transparency and Human Oversight**
  Document AI model decisions to explain how predictions are made.

  Include a human in the loop to review AI generated decisions and override them if necessary.

  Provide feedback loops where users can report biased decisions for further investigation.

- **Continual Monitoring and Updating of Data**
  Biases can evolve over time, so AI models should be retrained with fresh and unbiased data periodically

  There should be a process to set up ongoing audits to detect bias drift.

  Use real world feedback to identify and correct biases as they emerge.

Algorithmic Bias
Algorithmic bias refers to instances where an AI model generates outcomes that differ across groups as a result of its design, training data, or decision-making process.

- **Perform Fairness Testing & Bias Detection**
  Evaluate whether different demographic groups receive significantly different outcomes.

- **Use Explainable AI (XAI) to Understand Model Decisions**
  Use interpretability techniques to reveal biases in decision making:

  – SHAP (SHapley Additive Explanations) – Shows which features influence a decision more than others.

  – LIME (Local Interpretable Model Agnostic Explanations) – Generates human readable explanations for AI predictions.

  – Counterfactual Analysis – Changes certain input values (e.g., gender, race) to see if the outcome changes.

- **Conduct Real World Testing with Diverse Data**
  A/B Testing: Compare AI outcomes for different groups before deployment.

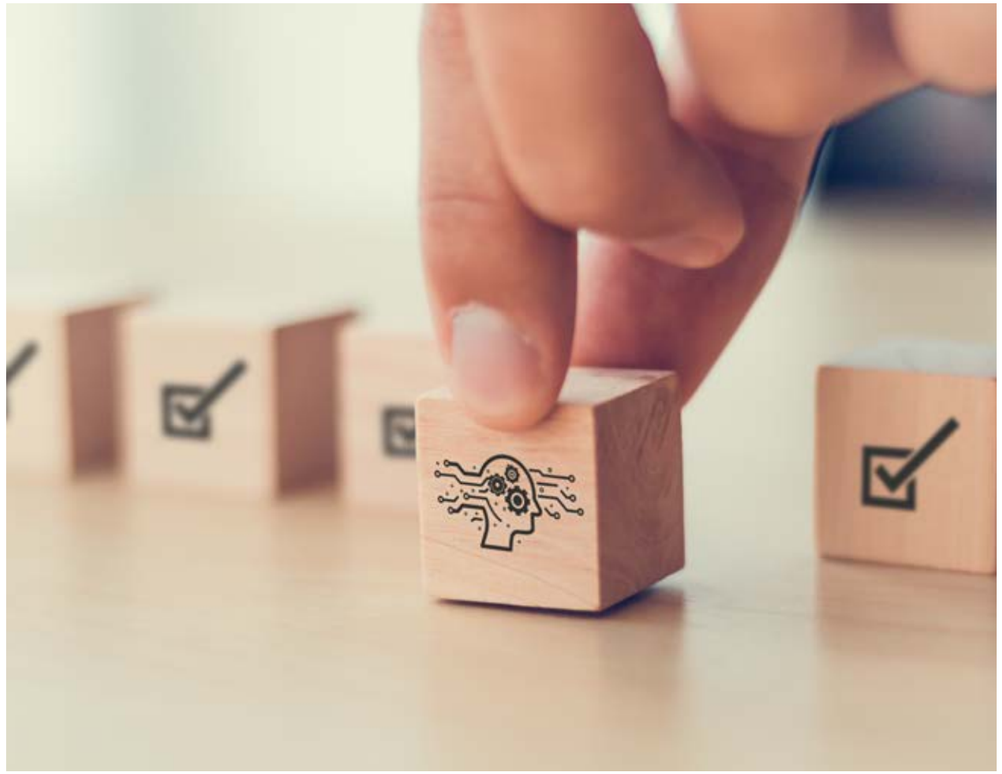
  Edge Case Analysis: Test the model with underrepresented groups to check its reliability.

  User Feedback: Collect feedback from affected individuals to identify potential biases.

- **Selection Bias**
  Selection bias occurs when the dataset used to train an AI model does not accurately represent the real world population, leading to skewed or unfair outcomes. This bias can cause the AI to generalize poorly, favouring certain groups while disadvantaging others.

- **Improve Data Collection for Better Representation**
  Ensure balanced sampling across demographics.

  Collect new data to supplement underrepresented groups.



Use data augmentation techniques (e.g., synthetic data, SMOTE for oversampling).

- **Reweight or Resample the Data**
  Oversample underrepresented groups (e.g., duplicate minority class samples).

  Under sample overrepresented groups (e.g., reduce dominant class samples).
  Reweight data points to equalize their influence in training.

- **Use Fairness Aware Algorithms**
  Train AI models with fairness constraints to minimize bias.

  Apply adversarial debiasing, where an auxiliary model identifies and reduces biased patterns.

  Use fair representation learning, transforming data so that sensitive attributes (like race or gender) do not affect outcomes.

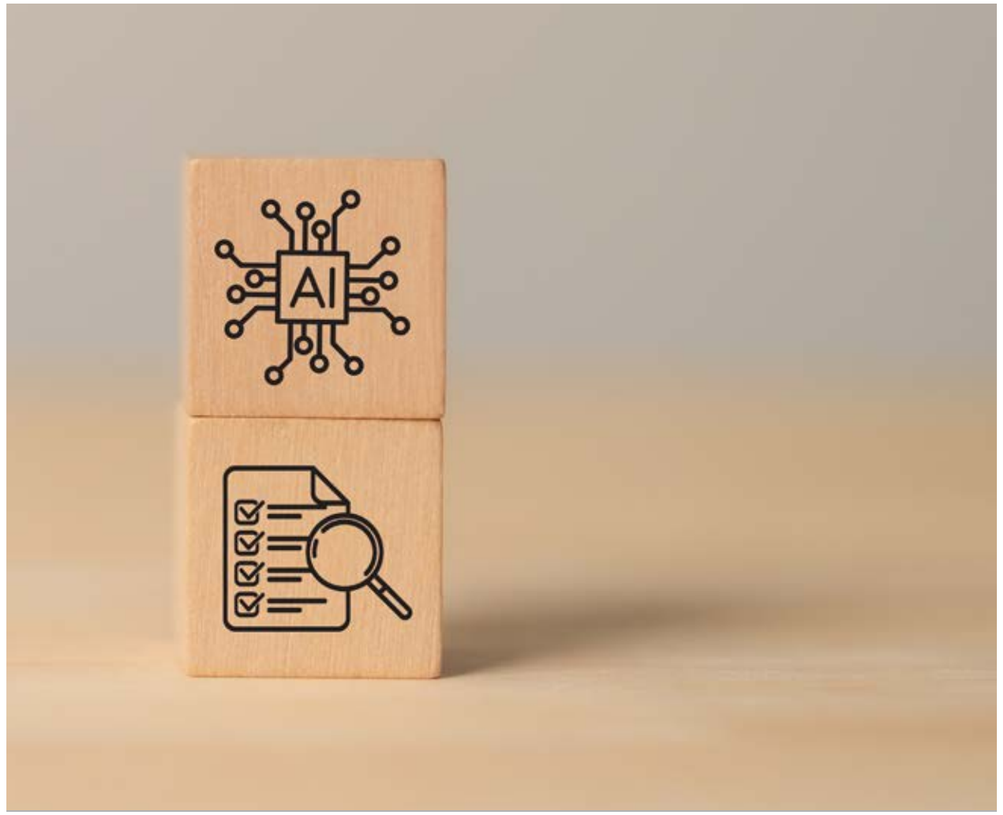
- **Adjust Model Outputs (Post Processing Fixes)**
  If selection bias remains, adjust AI predictions after training.

  Use reranking methods to balance outcomes for different groups.

- **Maintain Continuous Monitoring and Updates**
  Bias can shift over time, so conduct regular audits om generated output to identify if bias has crept in. This can be done by setting up automated fairness checks to catch emerging bias.

  Periodically retrain the model with new, more balanced data.

- **Confirmation Bias**
  Confirmation bias occurs when AI models reinforce preexisting beliefs or selectively favour information that aligns with prior assumptions,



leading to skewed decision making. This can happen in recommendation systems, hiring AI, news feeds, and predictive analytics, among other applications.

- **Diversify Training Data and Sources**
  Collect data from multiple perspectives to ensure balance.

  Avoid filtering data in a way that excludes alternative viewpoints

- **Modify the Algorithm to Promote Diversity**
  Use fairness constraints to encourage balanced decision making.

  Apply reranking techniques to ensure a mix of diverse content.

  Introduce exploration strategies so AI does not always reinforce prior choices.

- **Adjust AI Outputs to Encourage Diverse Perspectives**
  If bias remains, tweak AI generated outputs to ensure fairness.

  Introduce randomness to prevent reinforcing existing patterns

- **Implement User Controls and Transparency**
  Allow users to adjust recommendation settings to explore other viewpoints.

  Provide explanations for AI decisions, so users understand why certain content is recommended.

- **Monitor AI Over Time for Bias Drift**
  Bias can change as AI learns from user interactions, so periodic audits would help identify such drifts.

  Set up automated fairness checks to detect emerging biases.

- **Automation Bias**
  Automation bias is bound to happen when users overly rely on AI decisions, assuming they are always correct, without questioning potential errors

- **Design AI with Explainability & Transparency**
  AI should provide clear explanations for its recommendations.

  Use Explainable AI (XAI) tools to show why the AI made a decision

- **Implement Human in the Loop Systems**
  AI should assist, not replace, human decision makers.

  Require manual review before acting on AI decisions in high stakes applications.

- **Introduce AI Challenge Mechanisms**
  Users should have an easy way to question or override AI decisions.

  Provide a **“second opinion”** system where AI outputs are verified by alternative methods or experts.

- **Train Users to Critically Evaluate AI Outputs**
  Educate users about AI limitations and possible errors.

  Implement bias awareness training for professionals relying on AI (e.g., doctors, lawyers, pilots).

- **Monitor AI Performance and Bias Over Time**
  Continuously audit AI decisions to detect patterned mistakes.

  Create feedback loops where users can flag AI errors for improvement

- **Measurement Bias**
  Measurement bias is said to happen when incorrect or unfair labels are used in training data – when the correlations are incorrectly assumed.

- **Improve Labelling Methods to Reduce Human Bias**

– Use multiple independent labelling to reduce individual bias in data annotation.

– Implement blind labelling so annotators do not see demographic information.

– Use active learning to focus labelling efforts on uncertain cases

- **Remove or Modify Proxy Variables**
Identify features that correlate with protected attributes (e.g. race, gender, age) and either remove, reweight, or modify them.

Use causal analysis to determine if a feature is leading to biased decisions.

- **Use Fairness Aware Algorithms**

– Train AI models with fairness constraints to minimize bias

– Implement fairness aware machine learning techniques like:

– Reweighting – Adjust weights for different groups to ensure fairness

– Fair Representation Learning – Encode data in a way that removes biased information.

- **Continually Audit and Monitor AI Performance**
Regularly test AI models for new measurement bias.

Set up automated bias detection to prevent unintended drift.

Collect user feedback to identify biased outcomes in real world usage.

- **Biblography**

Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies by Emilio Ferrara- https://www.mdpi.com/2413-4155/6/1/3

Human decisions and machine predictions by Kleinberg, Lakkaraju, Leskovec, Ludwig, Mullainathan - https://pubmed.ncbi.nlm.nih.gov/29755141/

Discrimination in the Age of Algorithms by Kleinberg, Ludwig, Mullainathan, Sunstein- https://academic.oup.com/jla/article/doi/10.1093/jla/laz001/5476086

Semantics derived automatically from language corpora contain human-like biases by Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan- https://core.ac.uk/reader/161916836?utm_source=linkout

The ethics of algorithms: Mapping the debate by Brent Daniel Mittelstadt, Patrick Allo and Luciano Floridi- https://journals.sagepub.com/doi/10.1177/2053951716679679

Mitigating Unwanted Biases with Adversarial Learning by Brian Hu Zhang, Blake Lemoine, Margaret Mitchell- https://dl.acm.org/doi/pdf/10.1145/3278721.3278779

Krishnan Parthasarathy
Chief Architect, Product Management
TCS BaNCS