

Artificial Intelligence Discrimination: Cause, Damage and Mitigation

Kailan Shang, FSA, CFA, PRM, SCJP

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

As a simulation of human intelligence, artificial intelligence (AI) has been used to perform tasks with limited human intervention. The recent breakthrough of large language models (LLMs) allows us to use AI with little instruction. AI can act as a helpful and creative adviser and significantly improve our productivity. At the same time, the easy access to AI may lead to unexpected and undesired outcomes when there is a lack of controls and experience.

One of the benefits of using AI is to make unbiased decisions. Unlike humans, AI is not driven by emotions and adopts rational approaches. For example, an AI program can be built to use skills and qualifications to screen job applicants. Without demographic information, the recruiting process is expected to be more inclusive. Ironically, as studied in Chen (2023), biases in terms of gender, race, skin color, and personality were present in AI powered recruitment processes. Other biases are not uncommon in AI, and they need to be understood and addressed to avoid wide and adverse impact on our societies.

AI BIASES: EXAMPLES AND IMPACT

Biases in the AI programs may be observed in different areas such as the training data, the algorithm used for predicting, and the predictions themselves. It is no surprise that our data contains biases as it is measured and collected by humans who are subject to different biases. Algorithms can sometimes reinforce the biases in the data to a greater extent, ignoring new patterns in new data. When biases can be easily observable in prediction results, lack of robust validation becomes obvious.

Table 1
EXAMPLES OF AI BIASES

Area	Description	Example
Training Data	<p>Using biased data to train AI models, the biases are likely to be kept in the AI algorithm.</p> <ul style="list-style-type: none"> • Selection bias: the data is not representative of the population under study due to incomplete data, biased sampling, and so on. • Measurement bias: the data collected differs systematically from the reality due to a measurement issue. • Prejudice bias: the data includes existing human stereotypes and assumptions 	<p>U.S. hospitals used an algorithm to predict the need of extra medical care. Historical health care spendings were used as a measurement of the needs. This inappropriate measurement caused issues to underestimate the medical needs of black patients.¹</p>
Algorithm	<p>Although the training data may not contain demographic information as the source of biases, the model may learn from highly correlated variables and unintentionally discriminate against a certain group.</p>	<p>Amazon’s AI-enabled hiring algorithm favored male applicants based on words “executed” or “captured” commonly used by men, and penalized resumes with the word “women’s”. The program discontinued after the findings.²</p>
Prediction	<p>The biases in predicted results are not always obvious. But when it is obvious, the impact is usually devastating to the business and to the technology.</p>	<p>Google’s Gemini AI image generator produced images of historical figures in wrong and often darker skins. This led to the pause of this AI service.³</p>

Example Sources:

¹ <https://www.science.org/doi/10.1126/science.aax2342>.

² <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/>.

³ <https://blog.google/products/gemini/gemini-image-generation-issue/>.

When biased data is used for training AI models, it is challenging to predict the rare cases. Using the selection bias as an example, the data records that belong to the rare classes may be insufficient using standard processes. The training algorithm may be overwhelmed by the common cases and provide little insights on the rare cases. In addition, statistical measures may indicate a high level of prediction accuracy although rare cases are not predicted at all. For classification, precision, recall and the F-measure are popular measures based on the confusion matrix, as shown in Table 2.

Table 2
CONFUSION MATRIX ILLUSTRATION

	Predicted: True	Predicted: False
Actual: True	True Positive	False Negative
Actual: False	False Positive	True Negative

Precision measures the Type I error ¹ and recall measures the Type II error. F-measure (or F-score) is the harmonic average of precision and recall and may be used as a high-level measure to rank the performance of different models.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall (True Positive Rate)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$F - \text{measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

If we use the common class to calculate these measures, the prediction results may look promising, as shown in Table 3.

Table 3
CONFUSION MATRIX FOR THE COMMON CLASS

	Predicted: True	Predicted: False
Actual: True	95	0
Actual: False	5	0

Precision is 95%. Recall is 100%. F-measure is rounded to 97.4%. However, using the rare class, the same measures give us an opposite picture, as shown in Table 4.

Table 4
CONFUSION MATRIX FOR THE RARE CLASS

	Predicted: True	Predicted: False
Actual: True	0	5
Actual: False	0	95

Precision, recall and the F-measure become 0. It is clear that the AI model did a terrible job predicting the rare class mainly due to data imbalance in this example.

AI bias is not uncommon in the real applications. Even for companies that stay on top of the AI technology, it is possible that the AI applications may end up reinforcing immoral stereotypes and perpetuating inequities in our world. The impact can be quick, material and widespread, given the high efficiency of adopting AI in various areas in our societies. Certain groups may be given reduced opportunities in the economy. This can lead to a higher degree of economic inequality and social unrest. The trust in the

¹ Recall from classical statistics, a Type I error is a false positive where you reject a true hypothesis. A Type II error is a false negative and occurs when you fail to reject a false hypothesis.

technology may also be lost. With bad reputation and legal consequences, this promising technology may be ditched by business and society. Technological regression happened in human history. The collapse of the Roman Empire led to a decline in engineering projects, writing, and urbanization. Lower productivity is not an implausible scenario if there is a regression of AI technology.

AI BIAS IN INSURANCE

Many areas in the insurance industry have AI applications to facilitate automated decision making, including insurance pricing, underwriting, claim processing, and risk management. Bias of AI in insurance applications is no different from others.

- **Insurance pricing.** AI models may generate higher premium rates for certain demographic groups unintentionally. For example, insurance pricing models may use postal codes or location as a pricing factor. However, historical data may suggest a strong relationship between location and ethnic group. Demographic groups in poor communities are less likely to get affordable coverage. This means that ethnicity is indirectly used as a pricing factor as well. AI models may generate an unfair higher premium rate, even if the applicant's individual factors suggest otherwise due to data linking these areas to poorer health outcomes
- **Underwriting.** Similar to insurance pricing, AI bias may lead to unfair high-risk rating of minority groups even though individual data suggests otherwise. This can lead to limited access to insurance products and may affect the economic and social activities of the affected groups.
- **Insurance claim.** Claim processes may be different due to AI bias with certain groups having longer settlement time as more scrutiny was suggested to certain groups based on limited data that cannot differentiate further at individual claim levels.
- **Risk management.** AI-enabled fraud detection algorithms may target certain groups disproportionately suggested by training data but in reality, lead to wrong flagging of normal transactions.

Like other industries, AI bias in insurance applications may be caused by using datasets that are not representative of the entire insured group. Past practices may cause insurance data embedded with human biases. Without proper treatment, they will lead to a biased algorithm. When designing AI models, human bias may affect their fairness unintentionally.

MITIGATION STRATEGIES

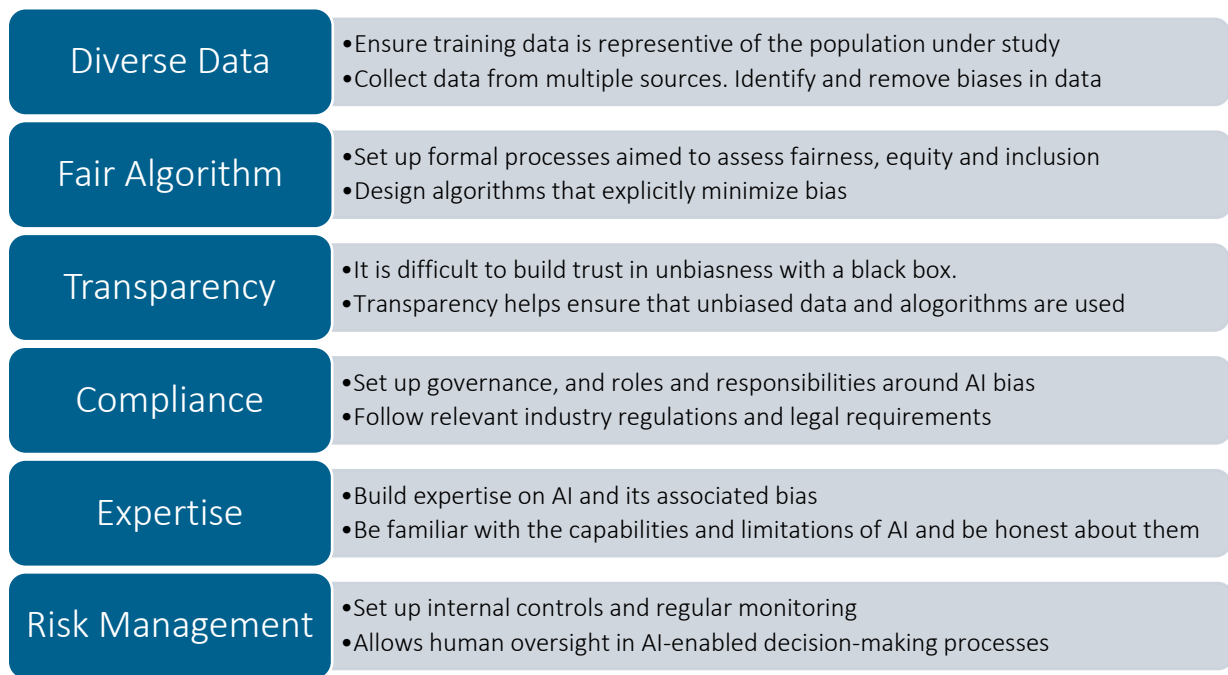
When the training dataset contains selection bias, there are ways to address them in AI model training.

- **Balancing the dataset.** If your data volume is big enough, you may consider removing some data records that belong to the common class(es) to make it more balanced. On the other hand, if no data can be sacrificed, you can use oversampling to increase the number of data records belonging to the rare class(es). The new data records can be created by adding small noises to existing data records. Well-established algorithms such as the synthetic minority over-sampling technique (SMOTE) can be used to generate synthetic samples. The algorithm chooses a few similar data records like in clustering analysis and adjusts the explanatory variables by random amount limited to the difference to similar records.
- **Adjusting the error function.** The error function is the objective function to minimize in the model training process. It can be adjusted to penalize false negative cases with a heavier weight.

- Collecting more data records belonging to the rare class(es) if possible. This may be better achieved with efforts from the entire insurance industry.
- Categorizing the common class into subclasses to achieve balance through more classes, if possible.
- Using measures such as Receiver Operating Characteristic (ROC) that consider the prediction results of all classes, rather than a chosen class. The ROC curve helps understand the trade-off between the true positive rate and the false positive rate by varying the threshold that is used to determine whether a prediction is positive or negative.

Although from the technical perspective, methods are available to fight against AI bias, during the implementation of AI applications, insufficient efforts may be spent on AI bias due various reasons. It is therefore important to adopt good practices to ensure that AI bias is managed actively. Figure 1 lists some practices to mitigate the risk of AI bias.

Figure 1
AI PRACTICES TO MITIGATE RISK OF BIASED DECISION



In addition, domain knowledge and diversity in talents can provide us with different and relevant perspectives to fight against AI bias.

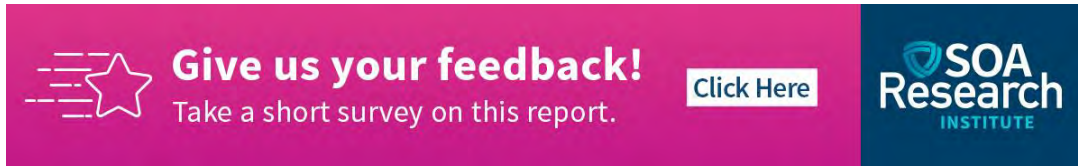
CONCLUSION

Given the remarkable advancements in AI, it is now easily accessible to the public with little prerequisites. At the same time, it also brings challenges to address the accompanying risks. In particular, AI bias can lead to unfairness in the decision-making process due to biased data and not well-designed algorithms. Bias was observed in AI-enabled health care, recruitment, and image generation systems. In the insurance industry, AI bias can affect fairness in underwriting, pricing, claim processing, and risk management. The impact of AI bias can be quick, material and widespread, given the high efficiency of adopting AI in various areas in our societies. Certain groups may be given reduced opportunities in the economy. This can lead to a higher

degree of economic inequality and social unrest. The trust in the technology may also be lost. Lower productivity is not an implausible scenario if there is a regression of AI technology. The risk of systematic AI bias needs to be and can be addressed using technical approaches and sound AI practices.

* * * * *

Kailan Shang is a director at Aon PathWise. He can be reached at klshang81@gmail.com.



The banner features a pink background on the left and a dark blue background on the right. On the pink side, there is a white star icon with horizontal lines extending from its left side. To the right of the star, the text "Give us your feedback!" is written in bold white font, followed by "Take a short survey on this report." in a smaller white font. A white button with the text "Click Here" is positioned to the right of the survey text. On the dark blue side, the SOA Research Institute logo is displayed in white, consisting of a shield icon and the text "SOA Research INSTITUTE".

REFERENCES

Chen, Zhisheng, "Ethics and discrimination in artificial intelligence-enabled recruitment practices." *Humanities and Social Sciences Communications* 10, 567 (2023). <https://www.nature.com/articles/s41599-023-02079-x>