Bias: The Hidden Tariff of AI for Industry in Africa

Regan Shen
Isazi Consulting
Sandton, South Africa
regan@isaziconsulting.co.za

Abstract

The growing adoption of Artificial Intelligence (AI) across industries has been accelerated by the availability of pre-trained models and open-source tools. These models offer practical benefits, enabling organizations to integrate AI solutions without the need for costly and time-intensive development from scratch. However, this convenience often comes at the cost of inherited bias—especially when models are applied in contexts that differ from those they were originally trained for. This paper explores the challenges that arise when AI models, developed with limited demographic diversity, are deployed in African industry settings. Through a real-world case study, we examine how these biases manifest in practice, the limitations of common mitigation strategies, and the systemic under representation of African datasets. We highlight the need for more representative datasets, local research capacity, and structural investment to ensure fair and effective AI adoption on the continent.

1. Introduction

The widespread emergence of Large Language Models (LLMs) such as ChatGPT [10] has significantly increased the general public's engagement with Artificial Intelligence (AI). This surge in public awareness, in turn, has boosted global industries' confidence in adopting AI solutions into their workflows [9]. The combination of growing popularity among the general public and industrial integration has led to substantial increases in funding for AI research and development [10, 9]. Consequently, this has led to the development of new AI models and datasets, both open-source and proprietary, deployed for use by individuals and companies alike.

While a considerable gap still exists in AI resource development and deployment between African nations and the rest of the world, Africa is no exception to the growing adoption of AI in industry and increase in research. This is indicated by companies such as Google and IBM opening AI research labs in different countries across Africa [2].

A common industry best practice is to leverage existing tools and solutions rather than expending time, resources, and capital to recreate them. This principle extends directly to the development of AI-powered solutions in industry. Instead of undertaking the laborious process of gathering sufficient data, developing models, and training them from scratch, organisations often research and utilise pre-existing models. In some straightforward cases, a problem may have already been effectively solved, allowing companies to ”plug and play” with readily available models, whether through commercial licenses or free open-source alternatives. For scenarios requiring enhanced performance on a specific task with an existing similar model, transfer learning offers a feasible solution, without the need to train a model from scratch. Transfer Learning involves leveraging the learned weights (knowledge) from a model pre-trained on a related task and applying them to a new model. This method can drastically reduce the data requirements and training time for the specific new task [15].

However, the convenience of directly using pre-trained models or applying transfer learning comes with a significant caveat: any inherent bias within the original models is transferred along, and subsequent fine-tuning does not guarantee the removal of the bias [13]. This embedded bias can lead to unfair or skewed results, particularly problematic in diverse contexts.

Through time spent working in industry and reviewing related literature, a common occurrence of bias in AI models used within the African context was evident. This paper focuses on a specific example illustrating how such bias manifests in AI models deployed in Africa, examines its underlying causes, and proposes potential strategies for mitigation or elimination.

2. Literature Review

2.1. Artificial Intelligence

Artificial intelligence (AI) is the field that covers computer systems designed to perform tasks by simulating human intelligence. Within AI there are many techniques and models that exist to simulate intelligence, such as Generative AI, Machine Learning, Deep Learning and more [11]. A comprehensive understanding of all the different architectures isn't required; the focus of this review lies instead on the fundamental learning process, known as training, through which these architectures acquire their capabilities.

The 3 main methods of training AI models are supervised learning, unsupervised learning and reinforcement learning [5]. The most common method of training a model is through supervised learning. This is done by using a labeled dataset, where for each data point in the dataset a corresponding label is provided, and the model will learn to map the features in each data point to the provided output label [5].

Using a labeled dataset, the model learns to identify patterns and features that help it predict the correct labels. However, to ensure the model generalises well and doesn't simply memorize the training data, it's crucial to evaluate it on a separate validation and test set. These datasets contain unseen examples, helping to detect overfitting, which is where the model fails to learn meaningful features and instead memorizes exact input-output mappings.

One common cause of bias in a model stems from bias in the training data itself,particularly when the data is unrepresentative. This occurs when certain groups are either completely missing or significantly underrepresented in the dataset. As a result, the model receives inadequate exposure to these groups during training, making it difficult to generalise well across all populations and leading to biased or inaccurate predictions for underrepresented groups [7].

A bias datasets can originate from historical data, where the dataset accurately reflects past societal inequities or discriminatory practices. While historically true, this data can misrepresent the current societal values or conditions that the model will be deployed and used in. Beyond the data's content, the methods of its creation can introduce issues; data collection bias emerges during the initial stage of data gathering. It happens when the techniques used to collect the data are inadequate or unfair, often because data collectors fail to sample adequately or ensure equal representation across all parts.

Lastly, Labeling bias occurs during the annotation stage of data collection. This is when the subjective biases or stereotypes of human annotators are inadvertently transferred to the dataset itself [3]. This can happen if individual annotators are manually annotating data points, or when the labels themselves are subjective, as seen in crowd-sourced tasks or surveys. Each of these forms of bias ultimately compromises the AI model, highlighting the importance of careful data curation [3].

Bias can also be introduced through the model's architecture, though this is less common than bias stemming from the training dataset.

For instance, repeatedly redesigning a model's architecture through multiple iterations of testing against a specific dataset to boost its accuracy can lead to the model becoming biased towards that particular dataset. Then when such a model is released, even if users retrain it on a new dataset, it will consistently under perform on any other dataset than the original dataset the model was designed for, even if the datasets are in the same domain.

2.2. Facial Bias

In the 21st century, facial recognition has become commonplace amongst the population, with many individuals using it daily: for example, to unlock their smartphones using facial biometrics. In South Africa, financial institutions are also using facial recognition technology, such as using facial recognition to grant users access to their banking apps. Building on this success, these financial institutions are now actively seeking to integrate it into more features. One such feature is facial recognition for confirming identities in uploaded documents.

This problem was brought to us by a financial institution seeking a reliable solution for verifying that the user uploading the identification document is indeed the same person pictured in the uploaded document. However, implementing this is more complex than the traditional facial recognition used in current banking apps. Typically, these apps require users to scan their faces during the initial setup after downloading the app and may periodically prompt users for updated face scans.

In contrast, some identity documents, such as the South African green card ID, have no expiration date. This means users can receive their ID at age 16 and never update their ID photo, which presents a challenge when performing age-invariant facial recognition (AIFR), a task that involves accurately identifying individuals despite changes in their facial features over time [14].

Unfortunately we were only provided with a limited amount of identification documents, (with no accompanying selfies for testing) and also considering the cost and time constraints of developing and training a model from scratch, an open-source, pre-trained age-invariant facial recognition model was necessary. Additionally, due to the strict security and compliance requirements in the financial sector, no external APIs could be used; all models had to be deployed and hosted internally on secure servers.

After evaluating several options based on benchmark performance, four models stood out as the most promising: FaceNet [12], FaceNet512 [12], VGG-Face [8], and MTLFace [6].

To assess which model was best suited for the task, the FG-NET Aging Dataset was used [1]. This dataset contains 1,002 images of 82 individuals, with age ranges spanning from 0 to 69 years. It was selected primarily because it includes images of children from infancy, whereas many other age-based datasets, often composed of celebrity images, start from the early 20s and lacked the adolescent years.

However when we initially embarked on this testing phase, our primary focus was solely on identifying models that excelled at AIFR. The potential for bias against African individuals in the models didn't cross our minds at that stage, as we weren't considering the task within an African context. It was only during a post-completion inspection of the dataset that the lack of diversity became evident: all individuals in the dataset were Caucasian. This oversight highlighted a crucial lesson – we should have selected a dataset that was more representative of Africa, including a greater proportion of Indigenous African individuals. Nevertheless we recorded the following results:

Table 1. Model Performance on FG-NET Dataset

Method	Accuracy (%)	Time (seconds)
Facenet	74.4	12.80
VGG-Face	75.0	4.91
Facenet512	76.4	66.33
MTLFace	63.8	2.53

After evaluating accuracy, execution time, and ease of integration, the most practical approach was to use the VGG-Face model with the DeepFace library. However, as previously noted, no corresponding selfie images were available to directly validate the model's performance against the identification documents provided from the financial institution. Consequently, model evaluation relied on indirect feedback obtained during internal testing.

To address persistent negative matches, we had to decrease the model's confidence threshold to 40%. While this adjustment led to more positive matches, it introduced ambiguity, these matches could either be true or false positives. Although the threshold value alone does not provide meaningful insight into the model's performance, feedback from internal stakeholders indicated that the model was unreliable and lacked consistent accuracy. This kind of second-hand feedback loop between business representatives and members of the technical team is not uncommon in industry, especially when dealing with AI, where key performance metrics are often unclear, overlooked, or not effectively communicated to technical stakeholders, making it difficult to iteratively improve model performance. As a result, the face verification component has not yet been fully integrated into the financial institution's production environment.

This raised a deeper concern, not only about performance, but also about potential bias in how the model interprets faces across different individuals or demographic groups. Particularly within the African context it is being tested on, which could be contributing to the inconsistency in results. And so during our investigation into whether the VGG-Face model exhibited bias and what might be causing it, we identified a compelling study conducted by Fakunle Ajewole et al. [1].

This research provided strong evidence that the VGG-Face model shows bias against African facial features. The authors used the same VGG-Face model as our tests but retrained it using a custom dataset. This dataset included 5,000 images of 500 different indigenous African individuals with their images spreading across a wide range of ages. The retrained model achieved an accuracy of 81.8% when evaluated [1]. In contrast, training the same VGG-Face model on a subset of the CACD dataset [4], limited to African-American individuals, resulted in a much higher accuracy of 91.5% [1]. To strengthen their findings, the researchers created two equal-sized subsets from the CACD and their custom dataset and repeated the evaluation. The results continued to show a consistent performance gap.

Their findings revealed that the VGG-Face model struggled with the aging features of indigenous African faces compared to nonindigenous African faces - such as African-American. Even after retraining the model on more representative data, its performance remained limited. This suggests that the original VGG-Face architecture was designed and developed with datasets lacking ethnic diversity, particularly with minimal representation of African individuals. As a result, the design of the model led to a biased architecture that can not be fully resolved through retraining alone [1].

A clear solution to resolving bias in our use case is not straightforward. As demonstrated by Fakunle Ajewole et al. [1]. They showed that even retraining existing architectures like VGG-Face on well-curated African datasets does not fully eliminate the bias embedded within the original model design. This suggests that addressing such bias requires more than just fine-tuning or dataset substitution, it requires a deeper re-evaluation of model architecture and training pipelines from the ground up.

3. Conclusion

This case study highlights the presence of bias in certain facial recognition models when applied to African populations, biases that cannot simply be resolved through retraining on new datasets.

The lack of AI infrastructure in Africa further complicates efforts to adapt or develop large models locally, particularly with the growing scale and complexity of LLMs. Addressing these challenges requires long-term, structural solutions focused on building representative, large-scale datasets that reflect Africa's ethnic and demographic diversity.

Equally important is sustained investment in African research, the empowerment of local institutions to develop state-of-the-art models, and a strong emphasis on open source collaboration within the continent. Reducing inherent bias and lowering the barrier to deploying AI models in African industry will not only improve adoption but could also inspire more students and researchers to pursue AI development, due to a greater demand for employment in this field.

This will create a sustainable cycle that leads to more research and development into AI in Africa, that will encourage the direct commercialization of homegrown models, accelerating the deployment of effective, ethical, and affordable AI solutions across key sectors like finance, healthcare, and agriculture across Africa. Enabling us to play a dual role as both a consumer and producer of advanced AI algorithms.

References

[1] Fakunle Ajewole, Joseph Damilola Akinyemi, Khadijat Tope Ladoja, and Olufade Falade Williams Onifade. Unmasking the uniqueness: A glimpse into age-invariant face recognition of indigenous african faces, 2024.

[2] Emmanuel Ogiemwonyi Arakpogun, Ziad Elsahn, Femi Olan, and Farid Elsahn. *Artificial Intelligence in Africa: Challenges and Opportunities*, pages 375–388. Springer International Publishing, Cham, 2021.

[3] Kedimotse Baruni, Nthabiseng Mokoena, Mahalingam Veeraragoo, and Ross Holder. Age invariant face recognition methods: A review. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1657–1662, 2021.

[4] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 768–783, Cham, 2014. Springer International Publishing.

[5] Dinesh Deckker and Subhashini Sumanasekara. Bias in ai models: Origins, impact, and mitigation strategies. *Preprints*, March 2025.

[6] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. In *IEEE TPAMI*, 2016.

[7] Luke Haliburton, Sinksar Ghebremedhin, Robin Welsch, Albrecht Schmidt, and Sven Mayer. Investigating labeler bias in face annotation for machine learning. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 145–161. IOS Press, 2024.

[8] Zhizhong Huang, Junping Zhang, and Hongming Shan. When age-invariant face recognition meets face age synthesis: A multi-task learning framework, 2021.

[9] Ritu Jyoti, Jennifer Cooke, Peter Marston, Hayley Sutherland, Al Gillen, Amy Loomis, Craig Powers, Dave McCarthy, Nancy Gohring, Mario Morales, and Laurie Buczek. IDC FutureScape: Worldwide Generative Artificial Intelligence 2024 Predictions, 2023. Document Number US51291623.

[10] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. Artificial intelligence index report 2025, 2025.

[11] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, USA, 1997.

[12] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[13] Stuart J Russell and Peter Norvig. Artificial intelligence-a modern approach, third international edition. 2010.

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[15] Grega Vrbančič and Vili Podgorelec. Transfer learning with adaptive fine-tuning. *IEEE Access*, 8:196197–196211, 2020.