Artificial Intelligence is becoming increasingly popular and is used for decision making in sensitive areas including hiring, criminal judging and even healthcare. Since AI impacts individuals and society directly, it is essential to understand the overlap of AI with bias and fairness. Past experiences with AI models have shown that biases can have longer term impacts, and hence it is trivial to understand biases and minimize their effect on the models output. Although, decisions made by humans are also prone to biases, however AI tends to not only absorb biases (societal) but also scale them largely.
However in some cases AI models may be fairer than human decision makers, since the model only considers features that increase its accuracy. Hence, AI can be used effectively to not only minimize biases in models, but also to identify and remove biases in human decision making.

The purpose of this article is to **explain algorithms that scale biases** and require human vigilance for careful detection, **highlight the work and research being done** in regard to minimizing bias in AI and to draft a **tentative path for future researchers** of the field.

Biases can be introduced into the model via different methods. The most common form of bias is the Algorithmic bias which refers to the biases present in the underlying data. Algorithmic bias occurs when either the data contains human biases or second order societal biases. These biases are often encoded in other features and need to be carefully analyzed before being used. Other sources of biases are data collection methodology, choice of features, use of feedback loop for data collection and statistical correlations.

In order to understand biases in AI models, it is important to define fairness. Many researches have defined the term in different ways, however each definition has its trade offs. Fairness is also sometimes defined in terms of a group of variables, but this can have longer term (positive and negative) consequences. Although, these trade-offs can be minimized by either setting different thresholds for different categories of the data or using a universal threshold for the whole data, however, crafting a single definition for fairness is not possible.

Under the umbrella of **“enforcing fairness constraints”**, multiple techniques have been formulated to ensure that a model adheres to fairness. Different approaches include careful data preprocessing, post processing and algorithm optimization. During the data preprocessing phase it is essential to remove sensitive features that impact the decision making processes greatly. It is important to maintain “Counterfactual Fairness” and get a fairer data representation. Improving fairness during the post processing phase includes making changes to the model in order to align the output and the fairness metric. Other methods of removing bias from a model include adding more data points to the training data, and using techniques like transfer learning and decoupled classifiers.

Identifying and mitigating biases from an AI model is also possible since the output can be dissected and analyzed. Local interpretable model-agnostic explanations (LIME), integrated

gradients, and testing with concept activation vectors are used commonly to identify the factors influencing the decision greatly.

Considerable work and research has led to the formation of different techniques and statistical models which help detect and reduce biases in AI, however human judgement is still required for making important decisions. In order to decide whether a model is sufficiently fair and unbiased and can it be used for automated decision making, human knowledge is required. Although, checking fairness of a model and following ethics during data collection is greatly emphasised, however methods like “Data sheets for data sets” and “model cards for model reporting” are used to ensure transparency of data construction, testing and use.

AI models tend to contain biases and might require human intelligence, however, they can also be used to detect biases in human decision makers. Running an AI model alongside a human decision maker can potentially identify the biases that influence a human decision.

Since AI models are both useful for eliminating biases and are prone to biases, it is important that future researchers understand the problem fully. Understanding the context in depth is essential since biases can have disastrous effects. Having a mechanism to remove biases from AI models and human decisions is essential, hence it is important that organizations follow ethics during data collection. Investments in form of data and research are required to diversify the field of AI and make it fairer.
Although AI and humans are susceptible to containing biases, however if used together, they can lead to fairer decisions