In (Manyika, Silberg, & Presten, 2019) human biases and their presence in AI systems are discussed. A few examples were provided such as:

* A British school used a computer program which result in discrimination against women and those with non-European names
* A criminal justice algorithm in Broward [County] Florida mislabeled African-American defendants as "high risk" at twice the rate as white defendants.
* NLP models trained on news articles can tend to exhibit gender stereotypes.

The article also discussed how there are different ways that biases can find themselves into algorithms. One was through the training data that the algorithm was develop with, such as the British school example where the algorithm had matched human performance with an accuracy of 90-95%. Another key source of bias was the way data was sampled for training, where certain groups may be over or underrepresented, such as when facial analysis technologies were found to have higher error rates for minorities and women of a minority.

Some things that the researchers suggested is that we should take advantage of ways that AI can help people improve their traditional decision making. One method they suggested was using a model to make a prediction or recommendation alongside the human, and probe sensitive variables included in the model such as sex or race to help understand how that changes the outcome of the prediction or recommendation. This in particular is an advantage that they pointed out since while it may be difficult to or nearly impossible to interpret the coefficients of models such as large neural networks, but we can change input values and use other similar methods to understand how the prediction or recommendation changes; whereas, humans can lie about their decision making or not even realize they have a particular bias and are less transparent about how they came to the decision. This methodology could be used to help out disadvantaged groups.

Another key point that the authors made is that where we can we need to collect metrics and try to quantify fairness of a model so that we can better understand the impacts of its use. Some ways that researchers have quantified or defined fairness is a requirement that a model have an equal predictive value across groups or require that false positive/negative rates across the groups be similar. They do acknowledge that this can be a challenge, and multiple fairness definitions often cannot be simultaneously met.

They also try to point out that not just AI researchers or data scientist implementing these kind of algorithms be aware of potential shortcomings or biases but management and business leaders also need to be away and try to stay up-to-date in this fast moving field. The management also needs to make sure they have responsible processes available to their teams to help make sure that bias is mitigated as much as possible, and some tech companies such as Google and IBM have published recommended practices or common tools that can be used.

**References**

Manyika, J., Silberg, J., & Presten, B. (2019, October 25). What Do We Do About the Biases in AI? Retrieved from Harvard Business Review: https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai