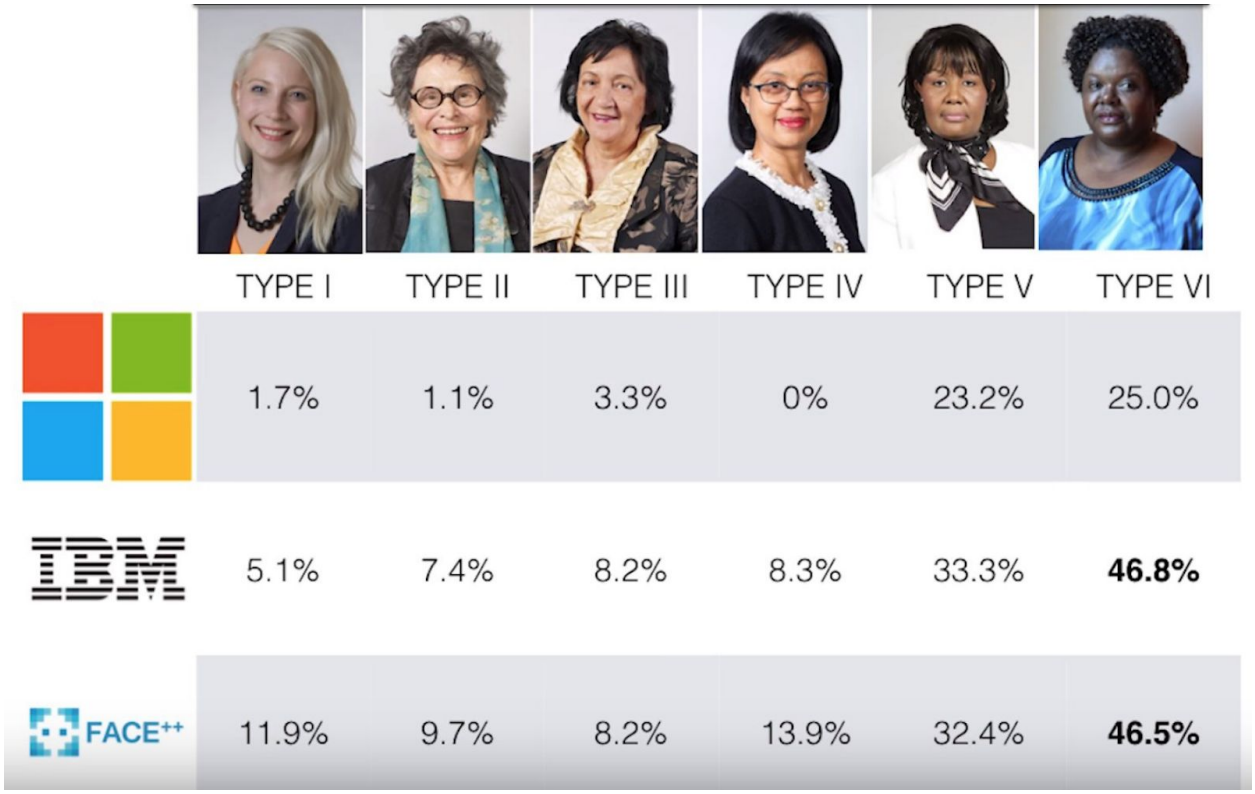
Algorithms with Human and Data Bias

Most of the models you've seen and/or programmed, rely on large sets of data to train and learn. When you approach a challenge, it's up to you as a programmer, to define functions and a model for classifying image data. Programmers and data define how classification algorithms like face recognition work.

It's important to note that both data and humans come with their own biases, with unevenly distributed image types or personal preferences, respectively. And it's important to note that these biases propagate into the creation of algorithms. If we consider face recognition, think about the case in which a model like a Haar Cascade is trained on faces that are mainly white and female; this network will then excel at detecting those kinds of faces but not others. If this model is meant for general face recognition, then the biased data has ended up creating a biased model, and algorithms that do not reflect the diversity of the users it aims to serve is not very useful at all.

The computer scientist, Joy Buolamwini, based out of the MIT Media Lab, has studied bias in decision-making algorithms, and her work has revealed some of the extent of this problem. One study looked at the error rates of facial recognition programs for women by shades of skin color; results pictured below.



	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
Microsoft	1.7%	1.1%	3.3%	0%	23.2%	25.0%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	**46.8%**
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	**46.5%**

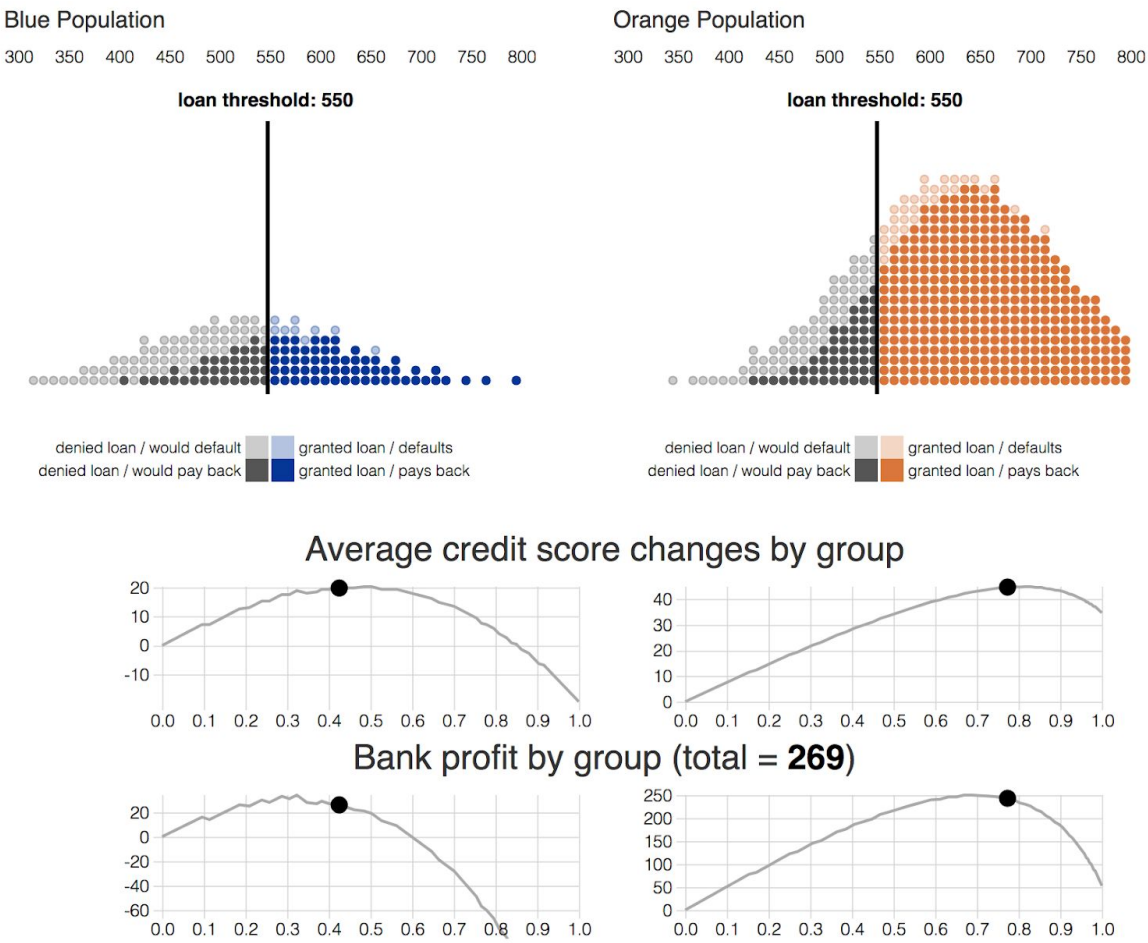
Image of facial recognition error rates, taken from MIT Media Lab's

gender shades website.

Analyzing Fairness

Identifying the fairness of a given algorithm is an active area of research. Here is an example of using a GAN (Generative Adversarial Network) to help a classifier detect bias and correct it's predictions: Implementing a fair classifier in PyTorch. And another paper that shows how "fair" credit loans affect diff populations (with helpful, interactive plots). I think that as computer vision becomes more ubiquitous, this area of research

will become more and more important, and it is worth reading about and educating yourself!



From credit loan paper, Delayed Impact of Fair Machine Learning.

Working to Eliminate Bias

Biased results are the effect of bias in programmers and in data, and we can work to change this. We must be critical of our own work, critical of what we read, and develop methods for testing such algorithms. As you learn more about AI and deep learning

models, you'll learn some methods for visualizing what a neural network has learned, and you're encouraged to look at your data and make sure that it is balanced; data is the foundation for any machine and deep learning model. It's also good practice to test any algorithm for bias; as you develop deep learning models, it's a good idea to test how they respond to a variety of challenges and see if they have any weaknesses.

If you'd like to learn about eliminating bias in AI, check out this Harvard Business Review article. I'd also recommend listening to Joy Buolamwini's TED talk and reading the original Gender Shades paper.

Further Reading

If you are really curious about bias in algorithms, there are also some excellent books on ethics in software engineering:

- Weapons of Math Destruction, Cathy O'Neil
- Algorithms of Oppression, Safiya Umoja Noble
- Automating Inequality, Virginia Eubanks
- Technically Wrong, Sara Wachter-Boettchera

Supporting Materials

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification