My question for this week is focused on whether there is a framework for proving that neural networks are free of unintended bias.

One big issue with facial recognition algorithms is the existence of possible racial bias. This bias could occur intentionally or unintentionally depending on the data used to train the algorithm. I’m wondering if a standard method exists to certify that a tool based on neural networks is free of unethical bias.

An example of an unethical bias would be a resume sorting algorithm that has been trained to filter out applicants with unusual names. This algorithm could just be intended to filter out invalid or joke names, but could also accidentally filter out names commonly used in other countries.