MITOCW | Ethics_of_AI_Bias_clip1

So you probably all know this, but just to remind you, the overall approach is simple.

You start out with a training set of data here together with descriptors, which we use to derive a model.

Then we input the data, and the model classifies it based on some criterion.

Obviously, the devil is in the details.

Based on this high-level approach, where do you see bias?

There's no bias here.

The algorithm-- say, a convolutional neural network-- objectively analyzes the data provided to it.

It's just math.

It is not just math.

We choose the training set.

We choose the algorithm.

Even if we think that we are not biased, we are subconsciously biased.

How do we actually choose them if everything is described by mathematics?

We don't really have a choice.

Duh.

The bias is already in the data due to historical inequity.

You said it yourself.

So then our choice is to note that and add an appropriate loss function.

Does everybody know what a loss function is?

Of course.

Yes.

Yeah.

A loss function is just the error function associated with an event, such as facial recognition.

It is what should be mathematically minimized.

Good.

Just to clarify, it shouldn't be simply minimized across the board but targeted, for example, so that the overall error in facial recognition is minimized with the constraint that it is equal among different groups.

But unless you have equivalent training sets for the different groups, the statistical errors won't be equivalent.

And you can't quantify the errors in advance to determine whether or not they will be equivalent.

In other words, you can't choose the loss function approach to solve the problem because the fluctuations or randomness.

The problem is statistics.

We have to have a very large data set to average out randomness to acceptably small percentages and make sure that our data is representative.

Yes, but my point is that it's just not possible.

I agree that we could make the error rates in facial recognition algorithms equal for Whites and Blacks and for men and women.

But what about all the other racial and ethnic minorities?

And for that matter, what about all the different genders?

Even if we could agree on what groups need to have bias eliminated, we can't practically address them all.

Yeah, and that's not even how the field works.

No one's going to spend five or 10 years for each algorithm to eliminate bias.

For cutting-edge research, people just try to hit 80%, sometimes 90% accuracy for a given algorithm and then publish a paper on it.

The incentives are such that they can't spend years getting 99% or so accuracy across the board, at least for new algorithms.

There just isn't enough data.

Better to just get something out there, even if it's not so good, so you at least get credit for it.

If we could only eliminate randomness, our problems would be solved.

But randomness is good because it's the basis of choice, otherwise our minds would just follow a deterministic path based on our neural pathways and electrical signals.

The heart of it is quantum mechanics, which introduces intrinsic randomness.

Well, randomness doesn't give you choice, just chaos.

That's also the reason we have bias to begin with, because of random historical circumstances.

We want to eliminate bias, not just make it random.

If it is random, then maybe you will mitigate it for minorities, or maybe, at least for some, you'll find a way to augment it.

[SCOFFS] That is what I was saying, too.

So we seem to be stuck again.

Mathematics is the rigorous way to understand the world, but it doesn't give us choice.

In the world described by mathematics, choice is an illusion.

Professor, we ran into the same impasse when we had this discussion last week during spring break.