So you all agree that it is desirable to eliminate bias, right?

Yeah.

Yes.

So you all think it is possible to mitigate bias depending on what actions you take?

Yeah.

Yeah.

So you do think that can make choices and not just randomly.

More well-grounded choices are better than less well-grounded ones, say, for addressing bias.

I don't think we really have a choice about things.

People might think that they do, but they don't really.

We can't choose where we are born or how smart we are, and we can't choose if we get hit by a car.

We think we are making a choice, but we really aren't.

It's like Sophie said.

It's all just an illusion.

Yeah.

Yeah, you're right.

Randomness doesn't help.

You've convinced me.

Wait.

Did I have a choice, or was I compelled to agree?

No.

Your neurons told you to agree.

Yes, exactly.

And randomness, a consequence of a chaotic quantum process, just makes what we do not predictable but doesn't mean that we choose what we do.

Right.

Well, you have a choice whether you're going to be biased or not.

Mhm.

Indeed.

It does seem that judgments involve choice.

But Alyssa, if you were given the full understanding of a situation, wouldn't you always take the best action?

Yeah, of course.

No.

You can't really choose whether anything is better or worse.

It's all just a value judgment.

You may be compelled to do something by others, but you never really freely choose.

Do you really believe that, or are you just playing devil's advocate?

OK.

A little of both.

But even if I am just playing devil's advocate, what's wrong with what I'm saying?

What's wrong is that your model of human life is inadequate.

It doesn't describe what it really is.

We're not just some balls moving around deterministically with random noise interspersed like a term in a Langevin equation.

Isn't that true, Professor Muller?

What a great discussion this is.

Sounds like you had a much better spring break discussing these questions than going to the beach.

[CHUCKLES]

I think you're on the right track, Sophie.

Funny.

Coincidentally, that's my first name too.

At any rate, you have unveiled a great tension in modern science.

We are brought up to think that modern science is the way to describe the world, but it seems inadequate to describe human things.

Ethical questions cannot be reducible to mathematics-- similar to justice and beauty, hope and suffering, and, for that matter, even truth.

Yes, Glenda.

Aren't those things even more important than modern science?

I would say so.

No.

Those things are arbitrary or a matter of individual's interests-- what's in it for them.

That's the way to look at them scientifically.

Bias exists because people have conflicting interests.

Sure, if you look at it purely scientifically, but interest alone cannot be adequate since people are willing to sacrifice interests for passions and, more importantly, for things beyond themselves.

At any rate, do most of us really know all of our interests?

We all have preferences that go far beyond interests.

Perhaps we can say at best that those interests are intertwined with different approaches to the world.

Besides, aren't the most boring people the ones who say, I'm right, and, even if one can never make value judgments, I have my own values and I stick with them?

[CHUCKLES] Yeah, these are the most irritating and boring people.

So, now let's dig down a bit more into the mathematics of AI and the sorting problem.

Perhaps it's best to start with a specific example.

What's the reason for the bias in facial recognition that we discussed a little while ago?

A biased training set due to historical inequity.

OK.

So one of the issues is the training set.

The other issue is the algorithm, including its hyperparameters, all leading to how the classification is done.

This is why different algorithms give different error rates using the same training data.

Anything else? Right.

So those are the two key issues.

However, within the choice of algorithm, there is the problem with classification.

Let me project this data set, here with points in red and blue.

You have many choices for putting the dividing surface.

But what if we choose the wrong one?

Some of the data may be misclassified, won't it?

You mean like this?

Yes, exactly.

So how do you choose the right line ahead of time?

That is a problem.

It could be even worse if your data set looks like this.

But then we don't have to stick with lines.

Why don't you draw a curve around them?

Like this?

Yes, exactly.

Now you've classified them all perfectly.

Well, wait a second.

If you do that, it works for the training set, but it may not work for any additional data sets.

In fact, it might be even worse than the lines because you're overfitting to the training data.

Do you mean like this?

Yeah.

That doesn't work.

Oh, yes.

I see it.

Exactly.

Remember this is a sorting algorithm, here in two dimensions with two descriptors, thus it is quite limited.

But even if you do it in a very large number of dimensions with a very large number of descriptors, you are still projecting a complicated system-- shall we say a natural system-- onto an artificial mathematical construct that necessarily leaves much out of the description.

Or if you want to stay in the realm of mathematics, you could say that due to a necessary incomplete set of descriptors you were fitting noise.

This allows you to get a perfect fit for your training set, but then you have an even worse fit for your data set moving forward.

Patch.

In other words, we have to accept the fact that there will be errors.

But at least we could distribute the errors evenly among groups.

That is what Alyssa said before.

We can try, but then I'm worried we won't include all the groups.

Even if we could, there'll be errors in the errors.

Unless we have a very, very large training set, which practically we never have, we won't be able to do that.

Yes.

The training set would have to incorporate all the data in the world.

You know, that reminds me of a story I once heard, I think from a South American writer, where someone makes a one-to-one scale map of their country with all the details included.

[CHUCKLES]

Ha, ha.

Well, I grant the problem.

But then what do we do?

We all agree that we need to eliminate bias, but it seems so hard.

What if we could at least have a very extensive model that for all practical purposes captures a given situation?

Which I guess is actually impossible because it would require a complete and accurate model of all the objects in the whole world.

Yes.

As engineers, we can at least work to do that.

If the model is extensive enough, we can eliminate bias.

But then we'd be eliminating any choice that we might have had.

We would be controlled by an algorithm.

That sounds-- that sounds tyrannical.

Indeed.

Mathematization of human things distorts them so much that they become something which they are not.

This is actually the utilitarian approach.

Human things like joy, anger, justice, honor, beauty, and happiness are really not mathematizable.

We lose all of what is important in them when we mathematize them.

We could in principle try to do so and therein transform human reality. Glenda.

That doesn't sound appealing.

It sounds dreadful.

Yeah, that wouldn't be good.

But we can at least do it partially, can't we?

Yes.

Maybe that would be a good solution.

Yes.

Even if those human things cannot be described by mathematics, it is still the only rigorous way to understand the world.

I'm not so sure anymore, but the bias is in the data set, and it is in the algorithms.

So we must be able to adjust our models to solve the problem.

Mustn't we?

Well, consider this-- our recognition of the problem of bias is not mathematical.

Mathematics comes only after such recognition.

Mathematics is something we impose on our broader understanding of the world.

We can use it to develop models, but they will always be just that, models which are necessarily incomplete descriptions.

I see what you're saying, I think.

Sophie.

But AI is based on mathematics.

So we have to fix this problem in the algorithms via fixing the algorithms, i.e.

by using mathematics.

Don't we?

Well, we can think of things this way, but that will lead us to a double whammy.

We cannot develop a non-arbitrary definition of the criteria for the training set, and, two, we cannot develop a non-arbitrary definition of the error target for the sorting algorithm.

Well, I knew that we were at an impasse, but I didn't think it was this big of one.

Professor, your class has made us more stuck than we were before.

So, what do we do?

Is there another way?

Indeed, Glenda, there is that.

It requires a reconsideration of our concerns and of the issue.