Jess: So hi, everybody. We're super excited today to have Dimitris Chatziparaschis and Kayla Chaplin talking to us about bias in their two fields of robotics and psychology. So I will let them introduce themselves and a little bit of their background and we'll get right into it. Do you want to start Dimitrius?

Dimitris: Yeah, I can start. So hello, thank you also for the invitation. My name is Dimitris Chatziparaschis. As Jessica said, I'm a second year PhD student at the Electrical Engineering Department of UCR. And currently I'm working mainly basically on the robotics field and specifically mostly in the machine vision and machine learning approaches applied to the robotics field.

In the past I have been working a lot with sets and rescue robotics and collaboration between aerial and ground robotics. And in my masters back in Greece, I did my masters, I worked also in the full autonomy of aerial robots, namely the drones, the unmanned aerial vehicles.

And now I'm working mostly in applications that are related not necessarily to a specific field of robotics, but we're working mostly in the agricultural field. But I'm mostly focusing on the computer vision, mass learning aspect of it. Nice to meet you.

Kayla: Cool, and hi, I'm Kayla Chaplin. My research, I should say I'm a second year PhD student in the psychology program at UCR. And my research tries to understand underlying mechanisms that are driving bias, prejudice, judgments, et cetera. And I do this by using formal mathematical models that are trying to understand the mechanisms that are driving our bias.

I also do research that looks at context. So looking at like state level, state level and just contextual factors that also might be influencing our bias. So yeah, that's a little bit about me.

Catherine: Well, thank you to the both of you again for being on our podcast. So, you know, let's just start off right off the bat. Like how do you define bias in your fields? Where, how do these biases arise? Or like what kind of mechanisms lead to bias in your work?

Kayla: Sure, I can start with that question. So in thinking about bias in psychology, I think bias is kind of broadly understood as like a preference for something, or importantly in my work for someone over another. So we think about bias towards people or groups of people.

such as a preference for like one racial group over another, the preference for straight people over queer people, the preference for able-bodied people over disabled-bodied people. So those are just some examples of like bias that people can have. And in my work, specifically, when I think about the mechanisms that are underlying bias,

I think about them in a really interesting way. So I use this really cool tool. It's a formal mathematical model called multinomial processing trees. And it looks at how we are responding to implicit bias measures. So for example, the implicit association test is a test of implicit bias. You're looking at two groups of people. For example, let's say black people and white people.

And then you're looking at good and bad. And then those good and bad are paired with those groups of people, right? And then this is a measure of implicit bias. And then I use this mathematical model to apply to the responses on that implicit association test. And it looks at four contributing factors in one of the mathematical models I use called the quad model. So that is the evaluation, so like a racial evaluation of how we're evaluating people. And then there's two types of cognitive control. One of them is like the ability to detect the correct response. So are you able to detect that this in this certain block of a trial, are you able to detect that white should be paired with bad, right? But maybe you have this preference.

this racial evaluation that white should be paired with good. So then there's a third type of cognition or cognitive control that is your ability to overcome that bias and your ability to respond in the correct way. And then the last kind of mechanism that we think of underlying this response to a bias task is sort of a catchall is what we call it. It's called guessing.

And that's like, if you just have a preference to respond, a positive or negative. And also just if it's not, if a response cannot be categorized in the other categories, it also filters through that one branch of the mathematical model. So that's kind of how I think about bias in my work and how I think about the underlying mechanisms and bias of my work.

Dimitris: So, yeah, in my aspect, since I come from more into the computer science and the electrical engineering application of the bias. So the term of bias basically on how we, you can find this in our field is mostly, it's highly related to what Kayla said, but mostly with the aspect of during training. So bias basically, so in our, in our basically in our projects, most of the time we are trying to develop the system from scratch from pretty much knowing anything. If we say that this is in robotics, for example, it could be like an agent or a robot that needs to perform a specific task. Now if we are trying to train this system in order to learn how to do that and somehow like in a month or in a week from now.

can perform outside and do this specific task. And it has been trained. The bias that we introduce basically, like how we related bias in our work, is meant to be about the robot decisions. So especially in machine learning, bias is mostly a term that is not necessarily a good term in machine learning. Because most of the time, it leads that the robot or the system favors one task over another without a specific explanation and basically can lead to less accuracy or results that we not necessarily want to have. For example, like a robot that we are trained, basically if we apply, if we do train a robot in on in our without thinking and without somehow rules, we need to make sure that this is like more like an abroad application that buys is that we do not need to have to include, we need to make sure that bias is not included during our design in order to make the robots mostly like robust in more open, open problems and in all of the scenarios.

So yeah, this is more like introductory application of bias in system designs and how we can be related to our field. But of course we can get more in-depth to that.

Jess: Yeah, I kind of wanna like dive a little deeper into what makes bias implicit bias versus not, because Kayla, you mentioned implicit bias. And in my head, I'm thinking like implicit bias is one of those things that we kind of unconsciously assign good or bad to something, right? But with machine learning, with robotics, there's not so much of an unconscious process, right? Like we have to be training the robots to say like this is good or bad, right? So do you feel like implicit bias is something specific to psychology or does implicit bias like come up in robotics and or is that not even like a useful distinction? What do you think?

Kayla: That's an interesting question. Yeah, so in psychology, we kind of think of bias in a few different ways. So I kind of just introduced bias generally as like a preference, right? But in psychology, we break it down even further to two types of bias. I would say mainly in the field, there's explicit bias, which is like the bias that people are willing to say out loud or the bias that people are willing to report on like an explicit measure, for example. So... they're willing to say and it usually influences their actions and stuff, but it's not something that's unconscious. It's something that people are conscious of. But an implicit bias is a form of bias that kind of happens in your brain automatically. It happens unintentionally and it nevertheless influences our judgments, our decisions, and our behaviors towards people. And so that's kind of how we think about it in psychology. I'm really curious to know what you think Dimitris.

Dimitris: Yeah, I really like also your response because I can easily say that this is somehow exactly how it's applied also in our field. I can say we can both like we can have like implicit and explicit bias. Because like first of all, when we try and where this bias comes from, basically something I want to mention is that everything during training starts from the data set that we provide to our system in order to learn it, to use the rules that we apply and then perform in the real world. And that's the thing because we that's the main issue like we need to make sure that this data set does not contain or partially it has to contain some of the biases if this bias will help the robot to perform better or not. I can...

We can discuss more on that, like why we need to include or not include the bias. But still, in the data set, all the forms of biases may be included. And we can see this in during application, like during the development. For example, if we command the robot or we say that crashing onto a wall, for example, is a bad thing to do, somehow we can say that the robot will develop a sort of a bias of not approaching a wall really like approaching the wall really close. Okay. But in case that we have set more and more complex goal of selecting something or for example like detecting a person in mostly like in a photo like okay like more like doing a recognition of a human. One of the things that we're going to provide as an input on the on our system. It's going to be photos from a lot of people, right? But if we somehow don't, we somehow like, we avoid or like we miss some of the photos or like we do not present a really more like a representative like data set to the robot or to the system. The system eventually will learn about this

um somehow this is what I need to learn so eventually in the end it will build a bias towards something that it hasn't been learned in the past so this is why I think like there might be both of these sides like we can have like biases that we apply and we need to make the robot to learn on but somehow basically may occur uh that we haven't seen in the next step and there

13:11

how we can fix that or what we do in order to make it better.

Kayla: Yeah. In psychology too, sorry, I just want to add that there's sort of a branch of psychology right now where some scientists have been looking at like machine learning techniques and looking at the bias in machine learning techniques. So how exactly what you were saying Dimitris, like what we're feeding the robot or the artificial intelligence

And if we feed it a bias data set, right, then it's gonna have this bias. So if we feed it pictures of mostly white people, it's gonna be bad at detecting like features in black people and stuff like that. So it's really interesting that this bias, how we conceptualize bias in both fields also has some crossover, right?

Dimitris; Yeah. Also, if I can add something to that. Yeah, Kayla, I totally agree with that. And one of the things that we, What we do, for example, in our field, it has to do mostly about the evaluation. This is the biggest thing that's happening in the computer science in order to make sure that our system is good. And basically, an evaluation is made on criteria of somehow what was the accuracy of our system, detecting accurately basically all the people in this photo and what was the score.

So this is some indications that we can get in the end after the training system in order to see that, okay, our system is detecting 90% of the times like all the people, or there is an issue somehow 60% or the accuracy is low. And this is another term that can be also really, I don't know if you're familiar or if you have seen that, it's mentioned a lot. It's the term of the overfitting of our system.

Overfitting means that our system has learned the data set itself, like, it knows like what to do exactly and learns like really good on this data set and has become so biased on this data set that if you make it work in like outside in the world and see something new, it's not as good because you learn a lot on that. So this is another term that is really a big topic in machine learning to avoid overfitting to make the system more open to a lot of different scenarios. Yeah.

Kayla: Oh, that's so interesting. I'm thinking of how we kind of conceptualize bias and how we think about how bias is formed and started, right? Like, you're getting fed this information as a kid. As you're growing up and as you're developing, you're getting fed all this information. And if you're in an environment where you're getting this biased information,

I'm gonna use the black and white example for example. Like in media, right? We have a lot of positive white figures in media and we have a lot of negative black figures in media, whether that's TV shows, but also thinking about like the news, right? We tend to associate.

black with threat because in the news, every time a black person is in the news, they're seen as like they're shown as a threat, there's not a lot of positive. And so because of that, we kind of are fed this information in our brain. And that's how we are forming these evaluations. That's how we're forming these, these biases of people. So it's interesting how those mechanisms are working in both engineering and in psychology.

Jess: Yeah, in that sense, it does seem like in psychology, like the stimulus is paired with good or bad. There's like this paired association thing. And then maybe in robotics, it's like the bias is, is it more of just like what data the robot is taught without this like paired good or bad? And then the like good or bad part kind of comes with, you know, when there is, let's keep this example of like black or white person crossing the street and a car has to decide to stop, right? Because there's a person and it wasn't fed any black people and it's data said, that's kind of where it comes, right? Or what do you think?

Dimitris: This is, yeah, this is, I think, I know that this is a possible scenario in the past, like it could be from different like, it could be also like between like, discrimination between like, men and women and depends on the data that you provided on the when you did the training. And so yeah, this is also was an example of mine, like when in the past, because one of the things I want to make clear here is that when you're designing a system like that, that could be a really complex. Okay. So it could be like you have a multiple functionality. So it could be a lot of lines of code like I can do more than one stuff like driving could be driving includes both the computer vision aspect, but also the navigation, the mapping, there are a lot of things included, okay. And there are a lot of things working together. So when you are designing this system, and especially if you use training data set not to make it better, you don't use like about like 1000 photos of people.

and training only this 1000 photos. Nowadays, these systems are continuously trained. Even though initially they were trained in about, I think two terabytes of data would be the less that would be, how many photos could be there? I think like more than two million, three million or four million. I think this is a really small number.

Nowadays, these systems are continuously trained in new data sets, like in new like photos from human like the new people like from continuously like to make it even better. So that's the most trickiest like in the most somehow trickiest thing about bias here is because you cannot really control even in the 1000 photos, you can inspect them one by one to see okay, this is uniformly distributed among women.

or men or people with long hair or short hair or anything like facial characteristics different. But in the amount of a trained system that has been trained for all these years and like for all these billion of images that has been trained, this is a thing that you don't really know if there are biases in the data. So this is why these systems are continuously checked and evaluated before they come out in the real world to make sure that are somehow equally and uniformly can perform well. But I can mention some examples in the past that have been. As soon as I mentioned one example with, for example, I remember that I read this article. There was, I think it was, I don't remember the company, but even though there was a company, an autonomous vehicle company, basically an automotive company, I think from Germany.

They were trained their systems to recognize basically animals in the streets, okay, in order to perform, to do autonomous driving in the cities. However, since this company released the code and made it public, available in some of the countries, one of these countries was also Australia. Okay. So one of the things that the developers, I think they didn't include in their training animals that behave like like the kangaroo. Okay. So what was the issue here? A kangaroo has a really specific way that it moves, right? It jumps, but it makes a big jump, especially like if it is a bigger one, it makes a bigger jump. So during detection, even though a kangaroo is not like a person, okay. This system was strict, basically the system of this autonomous vehicle was strict and of the jumping of the kangaroo because it wasn't trained of an object or a human being being behaved like that. So you thought that it was more like a teleporting person in front of the car. So of course this, this issue, like, like if you're a developer in Germany, you don't, of course you need to think of that or something like that. But like, uh, it can be a scenario that you need to test your system before, like you really need to test your system.

before you applied in the real world because some of these scenarios, they can happen. So of course, I think they fixed it immediately, like directly after that, but it was an issue that you see the system wasn't really prepared for that. But yeah, that's an example of a bias, we can say, of how a person moves.

Catherine: That is such a funny example because yeah, Germany doesn't have kangaroos jumping around.I like the concept that you brought up a little earlier, Demetrius, the phenomenon of over fitting a machine learning model to like something and how that makes it actually like bad, how it's worse at doing its job because it's now inflexible. So I'm wondering, Kayla, if that's like something that occurs in people, I can imagine like if you've just been fed only a certain type of information that it becomes hard to retrain yourself.

Jess: Yeah, and to tie in the regional differences too, like Demetrius just brought up, like if you said that you study how regional differences affect bias too, so maybe you can include that because this seems like a perfect overlap.

Kayla: Yeah, yeah, absolutely. So yeah, when thinking about bias and bias training, specifically in overcoming bias training, unfortunately the research like doesn't have a lot of promising positive outlooks, right? So in humans at least, and right now at least. So when thinking about implicit bias training, there's tons of different types of training that people use. There's counter examples to try to show you positive examples of people from another race if I'm thinking about the black white research again. And there's also like, you know, thinking about yourself in that person's shoes. There's things where you can think of them as like your best friend. There's all these types of things, these types of implicit bias trainings.

But unfortunately, I'm thinking of a big study right now. Calvin Lai was the leader on the study and he had labs from all over the US employ these different implicit bias trainings in different lab settings.

And then they did sort of a big meta-analysis and looked at how these implicit bias training, different types of implicit bias trainings lasted over time. And unfortunately, none of them lasted more than a few days. I think most of them didn't even last for a few hours. And also it didn't even like change their explicit bias either, right? Like it changed their explicit bias for a few moments, then it went right back a few days later to being back to explicitly bias and implicitly bias. So it's really interesting to think about how in machine learning, we can use these, we can feed it non-biased information and it kind of fixes that implicit bias. But in humans, it seems a lot more complicated than that.

But I think that's kind of my goal as a researcher is to, and I think it's a goal of a lot of researchers right now is to understand why implicit bias is happening and how we can prevent that. So I'm thinking about it, bringing in the contextual stuff, right? When we think about how we form our bias, when we think about how our brains are wired and to have these mechanisms that are exerting as bias. I think macro psychology is kind of a newer branch of psychology. I mean, we think about like sociology and economics, they've been doing macro level stuff for such a long time, but we don't really, we haven't really in the field of psychology thought about macro level features influencing individuals or groups of individuals. And so this is kind of a newer way of thinking about implicit bias and thinking about how structures, both physical structures, but also social structures, are influencing our bias.

So if you grow up in a region where there's tons of churches, everybody around you kind of has the same mentality and the social norm is to kind of be biased towards this group of people, right? That's going to influence your bias a lot versus somebody who grew up in maybe a really diverse area who maybe there's not a lot of religious structures, maybe there's a lot more community spaces, maybe there's a lot less bias in that region and how that's influencing a person's bias or group of people's bias.

And so that's kind of how we can think about like context and environment influencing our biases and specifically thinking about like the structural stuff. It's really interesting. I feel like there's so many weird connections between our two fields, Dimitris.

Dimitris: Well, and also, I really like the response of Kayla. And also, can I add my aspect on that? Because I think it's really related and I can give you my, I think, our perspective on this field. So I literally what Kayla said is the same thing can be applied in computer science and basically machine learning and I can make it a little probably, I don't know, mostly like more like an easier term that we use. So most of the times when we train our system we assume that this is because we input data and we get an outcome of an action of that system. Okay.

So we treat the system probably have heard of that in the past as a black box. So nowadays in black box means literally this thing, like how we're trying to figure out how this system finally selects doing the A action instead of the B action. We don't know exactly what inside the system may be decide with all of this huge input that we provided in order to select A and not B. So now

So nowadays, basically, there is also a huge research it's been doing like on trying to figure out opening this black box and checking what's inside because for example, there are sometimes like some systems that have been used nowadays, for example, in detection of really early stages of cancer that has been proven that there are some surveys that prove that...

some systems that have been trained in a million of images of any stage of specific cancer can perform better than a doctor or can be used to advise a doctor that say, okay, this is something that we need to check on. So this is now what they are doing in the research. Like they're trying to open this box to see what it made it since it performs better somehow than us. What did it make it?

perform better or do we have a bias or like, does it have a better somehow if it has like a bias or has been trained better than us? Like what make it better? So in my previous, like when I did, for example, like I give you also this example, like when I did my masters, I was training about my system on detecting a human, okay? Like inside the, inside the, in the photo, since it will perform in real time. So in my...

because I had to do that during my research, I wanted to open this box basically the as okay, when we are training, we have a neural network that there are a lot of connections basically is related to the brain itself. Okay, how activation lay basically activation happening in our brain in order to trigger a specific course, Kayla knows mostly about that. But this is how we how we how we have this more in the technical side. Okay, so

So one of the things that you can do is that you can visualize, if you assume that we have like five activation layers, you can visualize, and let's assume that the first one is the input and the last one is the output, is mostly the image that shows, okay, this is the person inside here, and this is not a person, for example, more like in the pixels. You can open this box and try to see the intermediate basically activation layers.

So in my, when I did that training, for example, even though, so this is the most interesting thing is that you open the activation layers and you see how the sensor, basically how the detector points out areas in the photo that are more likely that these are a person or a human or it's not a human. Because it's also learns, that's a good thing that it learns also of what is a human and also it learns also on what

it's not a human in order to exclude it for the final year. So one of the things that I noticed, for example, that I never did, I never provided this rule or information like that on my system was when I was training between like also like all the people, okay, human, I noticed that my system could detect women, okay, girls.

From the data set that I gave, there were a lot of women wearing shorts. Okay. So probably the legs were more like, I don't know, it could detect mostly like, it was more like an indicator that, okay, this is probably a person like, or a human because there were a lot of photos with, it was more like a summer. I don't know. It was like more the data set that contained people with not a lot of like, uh, more like with shorts and like a t-shirts.

But I noticed that my detector, even though I never did something like that, I provided an information like that, it could detect a human wearing shorts, like it detects the legs, because the legs were somehow you can see the skin. And somehow, this could be mostly like a hidden kind of bias or a characterization of my system. But you can see that it wasn't something that I initially said. But if we wanted to use this system, if we were

really dumb on somehow finding a person like around us. Probably we can use that system to see, okay, probably we need to see, I don't know, the legs, even though probably we are doing that, but I don't know. But yeah, that's an example of, I think, so yeah, what Kayla said, and this is how we also approach our system and we're trying to see what is in that black box, what makes it, what's the biases in it, or like, it can make us better, for sure. Yeah.

Kayla: Yeah, I just want to add that, you know, in psychology for a long time, we just understood bias as like something that happened in the brain. A lot of early research was just thinking about bias as sort of memory, just something that's in your memory. But by sort of having this mathematical model that I've been using in my research, that's kind of a newer thing-ish. I mean, it was used in cognitive research a lot.

33:52

but now it's used more in like implicit bias research. I should, yeah, it's becoming more popular in implicit bias research. It's opening up that black box in our brain. It's looking at those like underlying mechanisms in trying to understand why bias is happening. So it's trying to look at these different things that are contributing to bias and not just understanding bias as something that's happening. It's trying to understand what exactly is happening. So totally opening up that black box and looking at those underlying mechanisms, which sounds like exactly what's happening.

Dimitris: But this is how we can improve, right? Like this is why we want to use these systems to improve ourselves. Like for sure, we want to have really good systems like doing like their task, but this is also for us like to help us like to teach us because there is some probably there might be something that we haven't seen before or there is a combination for a lot of formation.

And we can use this to become better in our research, in our field, in what we do.

Catherine: Yeah. That's so cool. I would have never expected there to be a black box in machine learning because it's human-made. We should know how it works, but we don't. And it's so interesting to me that even a machine kind of comes up with its own rules for what you said, like shorts are a person, because I can see legs, which is hilarious. So if you're wearing jeans, you're not a person or something. I don't know. But I guess for me, it's a reminder of how human, all these processes really are in the end. A machine or a robot is ultimately a product of human creation and therefore a product of human biases that we put into it.

And so I guess for me, the question is, what should we, I guess, focus on or what would be the solution for all this? Or maybe it's a combined solution of figuring out how to resolve human biases to not include those biases in our trainings for AI or machine learning.

Kayla: Yeah, I feel like in psychology, I think in our previous conversation, Dimitris, we learned that both of our fields are taking from each other, right? Where I'm taking things that we're learning from machine learning and from AI, and you're taking things that you're learning from psychology and we're kind of incorporating things into each other's fields. And by understanding, I think, how we can reduce bias in AI and in machine learning, maybe in the future we can as psychologists take that information and learn how we can reduce bias in humans. Though I don't know that that's a perfect answer to your question. I do think that like by continuing to understand these mechanisms that are influencing our bias, that are driving our bias, we can hopefully someday understand how to mitigate or how to eliminate those processes or change those processes from going in one direction to another direction to decrease bias or to eliminate it completely in certain situations. So yeah, that's kind of how I'm thinking about it.

Jess: I think it's so interesting because, you know, bias is something that's just involved in decision making, right? Like it's like, you know, we do it all the time where we need to like recognize objects as distinct, right? Like bias is kind of discernibility, seeing difference, right? Like seeing legs and saying that's human. But it seems like what we want to do is control what is the variable or what is the factor that they're using.

38:04

to see difference, right? And is that a way to think about it? Yeah, so yeah, if something I could,

Dimitris: yeah, I agree with you, Jess. So it's, so yeah, for me, I'm trying really to see. So also, this is also a question from Kayla. I don't know if we wanna get to this aspect of the discussion of like, or if it is good to have a bias or not somehow, because of course bias will be there also in our decisions.

But sometimes, yeah, most of the times bias, probably it's not good to have. But this is more like for next, but I can give more like an example of that. But like, is more like, or is there any chance that the bias can help us or a specific bias in a specific moment, help us to make mostly like a good decision? I can give another example also, like a really sort of example of mostly like,

more like an application. So there was a basically like, I think last year, you know, League of Legends is a really famous online game. Okay. So OpenAI developed an autonomous player, eSports player to play the game. And they called also a lot of really famous eSports players like more like it's not, I don't know if you're familiar with that. It's not mostly like it's more like a an online game that you play with a lot of strategies. It's real time, you need to add, and it's a really known game, okay?

So they trained the system to learn on that game, to how to play the game, and so that system developed some strategies. So when they had the system ready, they did this big event, and they called a lot of really famous players that... globally is for players about that game. So they made their system play against these players. So one of the things that these players noticed directly was, and that's the trick is the nice thing that it's a nice observation is that the system, okay, most of the times, spoiler alert, the system won these players. It's more like playing chess, but like playing over-performing chess. But when these players as people, as us, started playing against that system, they were starting reasoning about the way that it plays. So if they say, okay, the system now acting too greedy. Oh, no, the system now, like the computer now it's acting not as greedy. It's trying to do some something that I wouldn't do. So they're trying a lot of like to evaluate the system and depending on their criteria and on depending on their somehow strategy, what they have learned in the past.

But even though in the end, the system eventually find a way to beat, basically to win them over the game. So this is something I wanted also to ask, Kayla, or you can also further discussion. So probably sometimes, I don't know if sometimes the biases or something that is developed inside, more like the implicit bias, if it is eventually better to have this or like, because probably can...

lead us to better results or like more like, like, for us being better or not? Because like you can see that the players didn't like the way that he played, but he was playing like, better. Totally. Yeah.

Kayla: Yeah. When thinking about this question, I kind of think about like, bias as a preference, right? Like there are some situations where a preference is probably good. I'm thinking about like food.

Right? When you have a bias towards something, you probably want to have a preference for food that looks good over food that looks bad. Right? Like thinking about evolutionarily, like you want to look for something that looks good because it's probably safe to eat versus something that's like, has something green and fuzzy on top of it probably means like not safe. Like it might indicate mold or something. So.

In that way, having a preference for something might be really important for survival. And in the same way, in the same vein, I guess, thinking about evolutionarily, when you think about the outgroup at some point, when we were living in really small groups, that might have been really important for your survival. You might have wanted to have that bias for the a preference for the in-group because that probably means safety versus the out-group. That's going to mean maybe danger. That was how we evolved and that was, I should say, that's how we probably speculate that bias evolved. I think that now we live in a time where bias towards an out-group member is not a protective measure anymore.

when we think about bias towards the out group, it's now a negative thing and it's harming people. And so at least for groups of people, right? I think that bias towards people and specifically towards groups of people, like no longer serves a purpose. And so that's why I think that we should probably definitely try to mitigate or eliminate bias towards groups of people in our lives.

because it just doesn't make sense anymore. But I think at some point, having that bias towards a group of people was maybe some sort of safety mechanism. But yeah, I think that overall, bias towards something versus someone might not be terrible. Having a bias towards something, having preferences towards something, isn't a bad thing. But...when those biases start to be towards someone is when it can be problematic.

Jess: It makes me think about what you said with food preferences. It makes me think that having a preference for something helps with quick decision-making, right? It's the worst when nobody has a food preference and then you just sit there and be like, what should we eat? And no one decides and you just can't make a decision, right? So it seems like it's very important for a decision and then also Dimitrius, like what you said about like this machine being really good at its task in League of Legends, it made me think, well, maybe machines are really good and bias is very good when it's trained for a specific task, like being good at League of Legends, right? But then what happens when we're trying to create some generalized intelligence, right? Which is more analogous to a human, right? Where we have to evaluate. context, weight, different things, make all sorts of decisions, right? Is that something you have to think about with your machines?

Dimitris: That's a really interesting question. Yeah. And it's really well put. It's a, that's exactly, I think how it's happening nowadays. It's like, before you start in the engineering, before you start for sure, like, developing something, you need to set the specific task. And, uh, I, I believe that nowadays there, there has been not like a system.

like a proposed system, more like a training system that can do anything, like mostly like decide for anything and like do most of the stuff. So most of the times we train our system on a specific task. So it becomes the better it can be. There might be like, there isn't any scenario of such a thing that you can see that it's more like it's a training system that can give you a lot of answers in different, under different contexts. And really fast answer, right? And most of the times it could be, I'm not saying like it could be like a nice response, okay? Like more like a correct response. But yeah, definitely you need to define the task. But yeah, it's really interesting for sure to see how you set the problem and how you evaluate and how you train that system. Yeah.

Kayla: Jess, I want to touch on something that you said. Thinking about this time component of bias, right? When you think about implicit bias, when you think of something happening unconsciously and like automatically, it's something that's happening so fast. It's something that's happening without even thinking about it. And like, I think an important context where another, just example of where implicit bias can be really bad.

Is thinking about like police shootings, right? When you are a police officer and you have an implicit bias towards black people and you have to make a split decision of whether they grab for your gun and shoot or not, that implicit bias, because it's happening so automatically, like this time component is so important. And because you're having this automatic association, you're having this really fast, racial evaluation, it's influencing your behaviors in such an unintentional way. You're doing something that is so fast and you're not thinking about it. And so I just wanted to mention that time component of implicit bias, that split decision kind of stuff.

And in research, there's a really cool implicit bias task that has it's a where rather than pairing black and white people with good and bad, it's pairing white and black people with guns and tools. So you might be shown a picture of like a hammer or of like a gun. And so when you see white people, you tend to associate them faster with tools and you're more accurate in, in associating them with guns versus associating black people, you'll misattribute tools as guns more often towards black people than white people. Sorry, I don't know why my brain just paused there. But so it's really interesting, this like time component, right? That is in an important piece when thinking about implicit bias specifically.

Jess: Yeah, I imagine that's definitely a factor with cars that are self-driving and just ideas like these machines that are now embodied and in the real world and have to make quick decisions versus just a computer that's trained, right? Like, so yeah, I'm sure you have this quick decision-making thing too.

Dimitris: So yeah, in the robotic side, I don't think that there is... So everything that is the computer takes in an input. pretty much it's already computed like what it's going to get as an output. The only thing that you need to change to make it probably faster is mostly the computer itself, like the resource that it uses. So I think that the computer is not really affected about like really fast or like because this is the main also difference from us as people. Okay. So it's more like it's more like it takes all the possible cases. It takes what also has been trained on. So it makes this decision. It makes the decision in the requested moment that it needs to take it. Okay. But that's the thing, because it, I can check all of the data that I received in that time and get the compute basically the output that will be the same output. It was evenly, it will get the same data in a, like in the future step for us as people, like as humans that we are probably, this is why we say we are human. We are making mistakes.

That's the thing that we can relate this with a computer that a computer doesn't really make a mistake because if you It follows exactly the same path, the same. It's more like the same program right it's executed the same the same thing. But if you change the output or like if you somehow You won't set the problem properly, it will make a mistake. But of course the mistake will start from our initiative, mostly like for our design.

But yeah, so this is one of the things I want to mention is that when I was mentioning, referring mostly in the implicit bias in the computer side, was mostly about the hidden parts of the biases that the computer have that makes it decide on something specific that we have never said to decide on, not like to make a really fast decision on something that is not well thought.

It will use all the resources to really think of that. But yeah, that's my response on that.

Kayla: Yeah, it sounds like kind of in a, so maybe time is one way to think about it, but another way to think about it is control, right? In that time, you don't have the ability to exert control to maybe overcome that bias, which kind of sounds similar to a little bit what you're saying, like maybe if you have a bias that you don't understand why the computer has, you didn't exert that control on the computer. It just kind of had that, right?

Dimitris: Yeah, exactly. It's another way to kind of think about it. Yeah, it's really tricky. Yeah, it's how we define bias on the computer side for sure. a computer, really, it will not make a mistake. Because it's basically applied mathematics and programming. It's like you have set the goal, it will not skip a line. It will follow exactly the rules that things meant to do. And this is why, for example, in the League of Legends example, if you are a person, you perform the best. One of the major facts that the people were losing like the eSports. players lose their games, it was because they did the mistake. But of course the robot also, the machine also followed some strategies that the players they haven't seen before or they thought that these are not really well thought strategies because of their thinking and stuff. But yeah, eventually, yeah. But yeah, for sure a computer won't make a mistake.

But this is why we need to be really careful on how we design, how we set our goals, how we train, how we do everything from the day one to day end. Yeah, it's like when we use them and we deploy them.

Catherine: Yeah, so that's called like concept of what you've been saying, Demetrius, about how the computer isn't making a mistake. It's following the rules that we've put in. It kind of reminds me how with like people, know, if you call someone out for saying something sexist or racist or homophobic, they might get defensive. And I think a part of that is, and you can probably speak on more to this, Kayla, is that to them, they aren't making a mistake while to someone on the outside, it's very clearly a like terrible thing to say.

And so I'm wondering if you can speak a bit more on that, Kayla, but also another thing, we've been talking a lot about how biases arise, but I'm kind of curious to hear from the both. of you, how do we like get rid of biases? And like, we've talked about factors that can lead to biases, but are there factors that will help us retrain ourselves or retrain our machines to be better at unlearning those biases? Like are there factors that make someone say more receptive or retain implicit bias training better for a person or factors that make a machine better if you feed it more information at collecting that information and better at learning new information.

Kayla: Yeah, yeah, this is a really interesting couple of questions here. So when thinking about like how people, yeah, they have these beliefs about themselves and when they're challenged, right? You kind of feel defensive and you feel like, no, I'm not wrong, right? Like... You're wrong. No, that's just how I was taught. I think that's something that kind of speaks to the like, like the environment portion of like, if you're growing up in an environment where you're not taught that having a bias that being sexist that being homophobic is wrong, then you're probably not going to think that you're wrong. And so I think even speaking to the second part, sort of the second question, thinking about how can we create sort of less bias or how can we teach people to be to be less bias? I think maybe we start by looking at like how we can create environments and how we can create sort of from the beginning of somebody's existence, like teaching somebody to be unbiased. And I think exposure to an environment where you have no bias is really important. I think as humans, we learn a lot from our environment. We learn a lot from watching people in our environment, whether that be our parents, our teachers, our friends, et cetera. We're learning from them at all times. We're also learning from our media. We're also learning from our TV shows, from social media, et cetera. So...

Having a lot of examples of every kind of person and every kind of role, I think is really important for mitigating bias towards these associations, right? Thinking about, like, if we could create a perfect world situation, right? Like having people in all roles be non-threatening, having people in all roles be positive. but also having people in all roles being in the opposite too, right? Like making sure that we have these examples on both ends of an entire spectrum of people. And I don't know that right now, living in the world that we live in, that is something that is attainable. That's kind of a utopia that I'm thinking of, right? But if I could… take our world and just scrap it. And that's how we could eliminate bias. That'd be great.

But thinking, I guess, more practically about like how we can mitigate bias a little bit is I think still having those examples and having, trying to create a space, create an environment where you are not just showing a doctor as a man, you're not just showing a threat as a black person, right? You're showing these, you know, a woman can be a doctor. So you're not sexist when your doctor comes in and you're like, oh, are you the nurse? Because it's a woman, right? Like that is an implicit bias. But just having these, having these examples, I think are really important and having an environment and social media on our television, our movie media, in our news media, but also in our everyday lives, making sure that we are having this sort of diverse environment is really important for thinking about bias. I think that's probably like one step in a lot of other things that can probably go into it, but I think that's probably one of the more crucial and like base steps before we can move on to other things.

Dimitis: Wow, I love the answer. I totally fully agree on that. And I can give also a positive perspective in our from our end. One of the best things and positive things that we need to take into consideration, computer science and electrical engineering is that we are over control of the system that we're developing. So what does this mean? It means that if we reduce the bias before training.

We're designing that, right? And we're setting the task. So if we exclude the biases or like, if we all gather together and we find the data set and we know that this data set does not include the bias and we train our system, of course we can avoid future somehow behaviors from these systems out in the world that we don't wanna see. Is it gonna be, I don't know, it's gonna be either sexist or like something that is not good.

Okay. So that's a positive thing here is that in engineering, you are, we are basically, and also with your feedback, we are designing the systems and so we can make them not biased. That's our goal, basically. Okay. And that's a really interesting question also in terms of how we are reducing bias in our aspect in our field. And there are a plethora of ways that we can do. Mostly like to attack in a court like attack the behavior or make it make the system more stronger into biases for examples during training. It could be a lot of things for I can give some more like examples more like when you train a system there might there are some approaches that they insert somehow lies into somehow.

When you're training and you're saying this is a person, this is not a person, you can give some examples like really randomly that are not true or like they're not stated, like it's not a human or like it's not like that. So our system, the system through that way can be somehow be more, how can I say it? Like it tries to overcome this somehow like this noise that we create in order to see, to say that,

No, this is truly, - So this is like for sure, like a human, or this is not a human, why to make it like even more stronger in terms of the decision. Another way is that to reduce biases, like during training, we are canceling a lot of good examples when our system is performing well. We are saying that... t's not as good as you might do. So we are canceling a lot of the activations basically that the system does. So somehow the system needs to be even better, even more like, even more strong in terms of the final detections. More like you do something good and I'm not really a good, how can I say? I'm not bribing you saying, no, yeah, you're doing good. It's more like reducing the cost or you are trying to force to make it even better.

So also in terms of biases is more like how we train on the globe basically on the data set. We do a cross validation, how we call it. And there are a lot of ways nowadays. So you can see that your system has been trained uniformly above the data set. But yeah, there are a lot of ways, but yeah, one of the things I wanna point out. So one of the things I wanna point out is that We are over control. And that's a good thing because like we can design it, train it, and basically we can exclude advice. It's all on us. It's not all the system. We do that because we, it's a creation from us. And the most important evaluation is happening through the feedback.

For example, you think of such GPT, if that GPT didn't provide Like, how can I say kind answers or more like more like you had somehow like a hidden bias or somehow something that is not really okay of us seeing this system would be they too would be taken down and probably would be reset or like it would be reformatted. But yeah, I just want to bring a good positive positive vibes on the design and the control that we are

Kayla; I hate to bring it back down, but I just want to note that like, if we are in charge of this AI, if we're in charge of training these computers, and if we have bias that we're unaware of, that can still exert itself, it can still present itself in this computer training, right? It can still be there can still show up. And maybe because we're not aware of this bias, we may not even be aware that it's biased, if that makes sense. So I just want to note that too, that because we are training, because we're in charge of training these computers and understanding what's biased and not biased, if we're not aware of a certain bias, then we might not catch it. So just wanted to note that. Sorry to bring it back down.

Dimiris: No, no, no. It's totally fine.

Jess; Yeah, that's actually exactly what I was going to say to Kayla. Like, that's why I was thinking, you know, well, maybe it really does start on the psychology end of us understanding our implicit biases before we can even, you know, be in control of these like really powerful computers, especially in the future, as they're going to start being considered more objective than us, right? So I think that's a really good note.

And I also, you know, we usually like... end with what you learned from each other through this conversation. But I wanna also expand on how you mentioned earlier, your fields are already learning from each other. And I wanted you to sort of explore the ways that they already are. I know in our previous conversation that wasn't recorded, we're talking about reinforcement learning and stuff like that. So yeah, maybe some ways you're already borrowing from each other and then you can end on just what you've learned from each other in this conversation that you might take with you.

Catherine: I do want to bring up just a thought I had to bring the conversation back onto a slightly happier note is I do like what Demetrius said about like finding comfort that humans have control over our machines. But I think in a broader scope, we have control, we have power over the systems we create, either our government structures or our communities that we build. And I think through that and through remembering that and our, and like empowering ourselves that we can make better systems to have less implicit bias in the world that we live in is also a nice thing to remember. And then that could make for less biased machine learning as well.

Dimitris: This is why I wanted to mention that in the end, most like to give the positive bias. me as Dimitris, I've been working in robotics. As I said before, I was working in social and rescue robotics. It's my priority to build and train and develop something that is going to be useful from us people like to help us, really help us or like make us better. Because I think in all of the fields, like in the end is mostly the main goal is like making people better, right? It's more like making us better, with the better versions of ourselves. And for sure, if we're going to have also robots with us, we need to use this technology and even the systems to make us better. And better, it's more like a broad definition, but you can see like better can be, we can make it to discuss how we are being better, but like being in a better position.

Happy and like being like, yeah, but yeah, I totally that's my that was my intention of me bringing this in the end. So yes. So what I got from, okay, I can I can start with that. I for sure I believe that. Yeah, computer science more like a recent we can say like a recent field, a recent field, I strongly believe that

01:07:49

especially machine learning and how we exclude or include biases if we need to. The engineers need to look up to the psychology and the psychology department because this is where all these terms that we are using right now has been established for all these years back in the past.

And I know that there are a lot of- There are some collaborations between these two departments, but I think that it needs to be more like more intense or like more, more essential to be happening because like, many of the engineers could probably design some of the systems and they are not getting into a lot of deep thinking of like how we use them, how we do how we exclude bias, we use mostly the computer science terms, but probably there are a lot of proved and invented for all these decades and like all this year from a Kayla’s research view and background in psychology. So yeah, that's my, and I thought for sure, like I think that this is Kayla’s research on biases is the main topic and like the main reference of what we need to search on and then how we apply this on the computer aspect. So this is why I'm so happy for that conversation too.

Kayla: Yeah. And I don't think this is a one-way street, right? I think that you are learning from us, but we're also learning from you by having like engineering mapping out these neural networks and understanding sort of these mechanisms that are contributing to bias in machine learning. we are also as a field of psychology, able to take that information and reincorporate it into psychology, right? And think about different ways of how our brain and our neural networks in our brain are influenced and how they can be changed and how they can be maybe diverted from one way to another because your field is understanding how to understand bias, right? Like...it's really interesting that this is a two way street.

And I think that bias is a topic that we are, while it's been studied in psychology for so long, I think that it is an ever growing topic. And I think we're gonna be talking about it maybe forever as humans or for a long time as humans. And I think that if we keep understanding how important that our two fields have this connection, especially as both fields are advancing, right? Like I think that we're using AI and using machine learning in some really interesting computational ways and learning that both of our fields can continue to learn from each other and having these collaborations, I think is really important for the advancement of understanding and mitigating bias.

Dimitris: Definitely. I think we're going to have psychology forever. There is no way that we are people. We need to have the understanding on how we behave and how we do. But for sure, computer science is more like a recent invention. Yeah, I think we're going to stay with that for all that long. But for sure, it's going to be a relation. Yeah, it's going to be both ways for sure.

Catherine: All right, so that's, yeah, I'm so excited to see like, where the, I guess, collaboration between psychology and computer sciences is gonna bring us into the future. And I know it's maybe a collaboration between these two guests that we have here today. You never know. But thank you again, Kayla and Dimitris for being on our podcast. I mean, this has been such a fascinating conversation, but also, like I mentioned, in a way empowering, just remembering that we are in control and we can fight this bias.

Kayla: Exactly. Thanks for having us. This was such a cool conversation. And yeah, hopefully Demetrius, maybe we should chat after this. We'll see if maybe we can...

Dimitris: Yes. Yes, we need to. And thank you so much for your invitation. Yeah, I really enjoyed the discussion and your questions were... Yeah, I don't know. I really enjoyed the discussion for sure. It was really nice.

Catherine: Alright. I mean, what a great conversation that was, right. And I think this is probably like one of my favorite things about this podcast in general is just seeing how these fields relate, but also in this episode, particularly the history of how they relate to each other. I mean, I know it's like chicken and egg situation, almost in a sense, like, which field came up with the idea of figuring out where biases come from first, psychology or computer sciences, probably psychology, but I just love hearing that back and forth that Kayla and Demetrius had with each other.

Jess: Yeah, they made it really easy because I mean, sometimes it's a stretch on this podcast to find connections, but they're definitely like already there. And I think like, you know, the terminology highlights the similarities, just thinking about the neural network, right? You know, that they use in computer science is very similar to the neural networks in our brain, right? And so when they started talking about the black box of the neural network, that was particularly interesting to me because it was highlighting how they don't understand necessarily what the computer is using to make these differentiations.

Like, you know, in that sense, Dimitris brought up the, in his example, using legs to be like, this is a human. I've also heard before when people were trying to categorize dogs versus wolves, the algorithm was really good at doing it. But when they looked deeper into the black box, they found that what it was using was whether there was snow in the background or not. And so it's like, yeah, wolves live in snowy places. So that was a good indicator. But that doesn't like get at what is a wolf and what is a dog, right? If you've had a picture of a wolf not in the snow, it wouldn't perform.

And so it's a little bit like alarming to think that we often don't know the variable it's using to discern these things. And I think the human brain is similarly a black box, right? Because we don't always know like what we are using to notice difference in people. I think sometimes it's easy, I guess, with skin color, gender, something like that, if we're cognizant of the bias, that that's the variable maybe we're using, but all the time we're making these decisions based on variables that we probably don't even have access to either.

Catherine: Well, I think that's the whole like thing about implicit biases, right? We don't notice it. And I like what you brought up in our conversation earlier about time or like how fast our decision making. really is. And similarly with machines, I mean, it's just running, running, running codes pretty fast for our sake and for the sake of technology. Like you said, with automated cars, you don't want your car to be a loading wheel as you're in the car, trying to figure out if that's a red light and I should stop or not kind of a deal. So for me, I think there is so much like...I don't know, like I kind of mentioned at the end, like, there's so much humanness throughout all this conversation, obviously humans having bias or the real like human like nature that the machines are using to determine if something fits into this category or not, which makes sense. It's a human made creation. Yeah. And I don't know, a funny thought I had, which I didn't bring up during our conversation is that-

Like, I wonder if we can say that we can draw this comparison of feeding an AI or a machine millions and millions and millions of like images for it to learn something. Do you think there's a parallel between that and like human evolution across several million years? Like generation after generation being fed whatever information the environment that we're living in being fed leading to the evolution of how we perceive and how we develop these biases or preferences?

Jess: Yeah, that's a really interesting question because traditionally we don't think about our memories as something that is like hard coded, right? But we do have these hard coded perceptual biases, I suppose, right? Like what we can see is only in these wavelengths. What we can hear is only in these wavelengths. And so I guess how I would think about it is like that that's what's being kind of constrained, but also tuned like through time, right? To like pick up on what is relevant information and to pay attention to certain differences.

Because you know, like there's always similarity and difference in everything we look at, but where our attention focuses whether that's on the similarity between people or the differences between people, that seems like what shapes our bias. And so of course that's really tied into our machinery. And that's actually what I thought was so interesting about the idea of robotics and these machines that are out in the real world, having to make quick decisions like humans do. And having to take in all these sensory inputs, right? Like the car has to see and respond to what it sees and stuff like that.

So I thought that that was interesting. And I wondered if you could actually like reduce bias by changing the sensory machinery. I wish I asked Dimitris that, right? Because that would change like, you know, what they pay attention to and or what they can see, what kind of biases they can see or not see, right? Like if it doesn't see color, you know.

Catherine: I mean, that would be interesting. I mean, I'm not entirely sure how anything would work in going off of the automated car examples. I imagined heat could be a really good indicator because presumably a living walking body is warmer than the environment. But a lot of things are warm though, as well, like, and move, let's say you just had a, I don't know, a hot dog cart or something, not human shaped, but it's hot. So I don't know, I don't know.

Jess: Right, and then the car doesn't want to hit non-living things that aren't hot also, right? So it kind of like-

Catherine: Yeah, I guess it works out.

Jess: Yeah, it's so interesting. Like you have to think about like, yeah, like what machinery, what senses to incorporate decision making. Yeah. And I do want to like also respond to what you're saying about the time thing. I think that is another interesting component in the sense that I'm and like making quick decisions about stuff I feel like often is the source of our bias and stereotyping to write like, I read this book, thinking fast and slow. That was by a thinks a lot about like how humans make decisions. And he contrasts the intuitive way of thinking versus the rational way of thinking. And the intuitive way of thinking is very fast. It's the one that stereotypes and is very easy to use. It's not like hard, like rational thinking.

So a lot of the times like our decision-making is in this sort of... intuitive state where we also like don't understand what is behind it, but it just happened really fast. And so, you know, then the other side of that is, yeah, the rational analytical thinking. And I feel like that's where maybe psychologists are trying to target and change implicit bias, right? They're trying to train people to rationally or to consciously associate good with all different colors and genders and whatnot. But then maybe it's just that this is like a different way of thinking, like this explicit way of thinking, maybe, to use the terms that Kayla provided, explicit versus implicit. And like maybe there isn't enough exchange between this explicit, rational way of thinking to feed into our intuitive, implicit way of thinking.

Like I wonder kind of how that works because theoretically, like we should be able to learn stuff through this rational focus, intentional way of thinking that becomes intuitive, that becomes, that's just in our subconscious as a way to respond once a new, you know, neural network is formed through the process of learning in this intentional difficult way. So yeah, I don't know, what do you think about that?

Catherine: I don't know, it's, it is interesting, I think, for us to think about, like the idea of, yeah, it's harder to change, I think, something that you aren't aware of, or even if you are aware of it, as Kayla mentioned, the fact that most trainings to try to unlearn implicit biases don't last. And I remember reading this paper couple years ago on like, um, sort of the effectiveness of DEI efforts. And I think one of the findings that they found was that it's easier to change people's behaviors rather than their beliefs.

So I guess the question is, you know, what are we trying to do? What's the outcome that we're looking for? And if it is to create a better society than... Ideally, we could change everyone's beliefs and that no one in the world will be sexist or transphobic or racist, et cetera, et cetera. That's not realistic though. And I think the idea of changing people's behaviors, changing that explicit factor is easier through... reinforcement of positive behaviors, accountability for negative behaviors.

And I think that could potentially build a better environment for people to develop into, because as Kayla mentioned, something that she's been looking into is kind of how your structures around you, the people you're around, the media you consume, what you're hearing, what you're seeing, what you're just exposed to- how that can sort of defend you from developing negative implicit biases. So for, I guess our future, we can't necessarily super, super change everyone's beliefs, but we can start building a world where beliefs are not, those negative beliefs are not as likely to propagate.

Jess: Yeah, yeah, that's actually a really useful framework to kind of explore this idea and is our behaviors as being intentional and it's like, I think this is where the time constraint thing is useful, right? Because if

I think our behaviors are not going to be something that we can intentionally shape if we have to react really quickly in a moment, right? Like we're just gonna respond instinctually and do a behavior that reflects our implicit values and biases. But in these like contexts where people are very actively making decisions, let's say about policy or in the social media context, what people they're going to highlight as you know, doctors or lawyers or, you know, what color or what gender, like people are making those decisions beforehand and less quick of a manner, I suppose.

And so maybe, yeah, that's how we can tie the time thing into it is like, okay, there are certain times where we can't think fast, but there are certain times where we can. And then those can shape our environment, you know, then our environment can be one that is less biased. Those get embedded into our subconscious, like the unbiased environment. And then it'll be like, you know, finally a starting point to be able to be more implicitly less biased. So yeah, yeah, I like that way of thinking about it.

Catherine: And thinking about it in terms of machine learning. I mean, a machine's gonna learn faster than people. I cannot read a thousand pages in two seconds like a machine can. That'd be really helpful for writing my dissertation. Alas, I cannot. But I can see it kind of similarly working the same just on a faster scale for a machine. Like as Dimitris said, you go through that block box and you realize, oh, there's a bias here I never thought of. So let's try to figure out how to... retrain this machine by putting in more diversity of images, like people that aren't just wearing shorts, or noticing that your data set has more, I don't know, dogs than cats, and you want it to recognize all animals so your cart doesn't hit anyone's head or something.

Jess: Yeah, no, I think that is like what people are focusing on in terms of eliminating bias in machine learning is like giving it well balanced training data sets so that they have exposure to yeah all these different types of people and scenarios. But I also you know I think this still the problem is that even with all that exposure the black box part is that we don't know what variable the computer decides to focus on because we don't need a computer like a lot of the times to use the same things that we would use to differentiate them. But that is a way that we could control what factors it's using to differentiate is by being like, okay, focus on this variable.

But the benefit of the computer is that it's not just focusing on one variable a lot of the times and not the same ones as us. Like otherwise, like how would a, you know, we can't tell when someone's getting cancer, but the computer can. So it's almost like the computer can unlock these like new insights through doing its own thing. But I mean, maybe there just needs to be more of a dialogue where it does its own thing and then we get better at figuring out how it's doing that.

But the problem I think kind of is that it's not, the computer is probably not focusing on one variable a lot of the times. Like it was probably easy with the wolf and the dog thing to find that it was snow as the major thing or like in Dimitris' like legs, it's probably using like a bunch of variables in the same way we probably use a bunch of variables and can't figure out what single one. And that's definitely like a benefit, but it does make it more difficult, you know, to have control and accountability over what it's using.

Catherine: Yeah, for sure. And I think I like what you said, like there needs to be more dialogue between ourselves and our machines, but also between fields because definitely as we've learned, there's so much biases being fed into these machines that we're not aware of. So there's a lot of conscious decisions that we need to make. And I think the constant checking of machines or checking of people as well, I think just, I guess, practice, you know, like just constantly checking back on people as they're trying to unlearn their biases or checking back on that black box of a machine to see like, okay, what's the weird thing it's gonna do now?

Jess: Yeah, yeah. And one other like useful framework in this that I've just recently learned about and I'm not a math person, but this is I think a math concept of this variance bias trade-off where... You know, we're always trying to kind of optimize this balance between, and when I say we're trying to optimize this balance, I mean, us and machines in terms of being generally intelligent and adaptable to multiple contexts, we're always trying to balance, you know, this idea of being overfitted that Dimitris brought up, which is like, you know, having a limited training set, but then becoming really good with that limited information, but not general, right? Like, so you get something new, it doesn't know what to do. Versus like being very general, but then never actually being very precise, you know?

And so I think, yeah, we're always trying to find the middle ground with ourselves, right? Like we do stereotype and stuff like that because it's easier, I guess, for our minds to do it. And we can't get to know every single person of a certain group to know all the variation that exists there. So we're always like sort of taking averages, right? And this is actually how they describe bias, I think, in this math context is kind of like...drawing the best fit line through all these data points. But oftentimes that best fit line doesn't actually hit any of the data points. It goes through it and can make good projections and guesses to get close. But it's not gonna be spot on like it is for something that's over fit.

And so it's really interesting having this framework of like, yeah, like when we can't know all the variation that exists out there, how do we generalize in a responsible way?

Catherine: Yeah, I'm thinking about this. I remember reading an article a few years back about AI bias in like hiring in, you know, using an AI to scan through resumes. And turns out that- there's a lot of bias in that because if I remember correctly, I think I can't remember which company, but it was one of those big companies. So it was like Amazon or Google or maybe both, or, you know, big, you know, giant corporation here. A

nd, um, the AI, if I remember correctly, it was fed, um, like past resumes of like successful workers, but tech. has a strong sex bias towards men over the past years, still does. And so because of that, the AI was unintentionally biased against like resumes that women sent in because it was trained primarily on resumes that men sent in. And I think they found that it actually penalized any mentioned the word women or female or anything related to that, because it was not something associated with what it was trained with.

So I think that whole, what you said about like bias being like drawing a middle line, but not actually hitting any points is like, yes, the average is this. But the two of us as scientists know, when your data is skewed, the average is not going to be the best representative of your entire population.

Jess: Yeah, it's so interesting to think about how the average could sometimes be nobody. Like, I think I heard an example of this where it was like, it kind of like a military example where people are trying to get like, I might be remembering this wrong, but trying to get shorts that fit the most people in the military, so they took the average size. But then like the shorts ended up fitting like nobody.

So it's like, I don't know, it's a funny concept that I've been thinking about. It's just like, you know, we use averages so much in science but sometimes like averages are just like useful tools for projection, but not for describing reality. And I don't know the solution.

Catherine: But no, I mean, I know like even historically in other fields, like even today, some people use body mass index as an indicator of health, but all that is based on data from, if I remember correctly, white and Japanese men. That does not constitute basically most of the world at that point. And like people have different body shapes. People have... you know, different ratios and a simple number like that based on only two populations cannot be representative.

Jess: Yeah, and actually, yeah, this makes me think of, you know, this and what you just said makes me think about what metrics we're using for health with BMI, what metrics we're using for success with the resumes, because, you know, we're trying to optimize these machines to, you know, perform tasks like we talked about in the podcast episode. So the task being like, you know, for League of Legends is like to win at that game or the task being like, find a successful applicant, right? But what is success? Like what is the thing it's trying to get this algorithm optimized? for the task, you know, because that's another interesting area of bias is just like what we define as success or what we define as health and like, whatever we define as health or success, like that's what they're being, you know, developed around. It's like a really foundational source of bias.

Catherine: You know, it's funny that you brought up success because that fits in really well with our episode with Paige and Nora, how we were talking about. success for a male mouse could be completely different from a female mouse or is success winning the Tony award for your play or is it something else? Yeah. And yeah, I don't know. And yeah, like with our machines, like when you have it do a simple task, like find all the red dots in this picture, that's such a simple task. But once you start moving up to something that is truthfully subjective, like who's the best candidate in the stack of resumes? And I don't know, I guess the subjectivity of it all is why I think there always needs to be a human person in the end making that hard decision.

Jess: Yeah, and multiple perspectives of that human person, right? These different perspectives on like what success is, and then we can build these functions to optimize for those different versions of success using also like different metrics. And so I don't know, maybe those are some ways that people probably are already doing and we're just not well versed in the field, but yeah, I see that as hopeful, you know?

Catherine: Yeah, and I think like what Dimitri said, like everything we have, our machine greater societal systems, they're human made, and therefore they can be changed by us. You just gotta work through that black box of the machine or society as a whole, which is a daunting task, but something we can do together for sure.

Jess: Yeah, and I like also, yeah, just thinking about how, you know, when we see these machines and their biases playing out in the real world, they really are reflections of us and they allow us to see like our own problems and biases in ways that we often don't from within ourselves right so it's cool it's like it almost gives us perspective about ourselves by by how it plays out so I do think it is useful like yeah in both directions to kind of realize like how we want to be and the repercussions of like how we are.

Catherine: So what do you think that means for us in terms of science communication? Oh gosh.

Jess: Yeah, I mean, in terms of like where our biases are in science communication, I think, yeah, I mean, I think some of our biases are just that we maybe aren't including enough perspectives, kind of like what we just talked about, right? But yeah, like... Science, how we think of it is pretty narrow because it's usually just what studies have been done and what are the results. And it doesn't seem like there's a lot of different perspectives there to highlight. And I don't know, maybe we would see those if we ask people from outside the field how they would interpret scientific studies. And maybe it's in the interpretation side of of science communication and that we can bring in maybe more diversity.

Catherine: Yeah, definitely. Like through my internship here with the San Diego Zoo Wildlife Alliance, I'm getting to see so much of that like backstage kind of view of conservation where it's not just the team of scientists, but it's also working with rangers and people that manage the land with the local communities, all these different stakeholders that interpret data differently interpret sort of what we should do to preserve a species so differently. And it's figuring out how to navigate all of those, like those different perspectives collaboratively and like productively, because we are all having the same goal, which is we should preserve the species, let's try to make it better, but people have different ideas on. what's best and I think learning about what others have learned from considering those different perspectives of people outside of research has been so interesting to me. And yeah, the bias in science, one, you know, scientists are human and we can go on about that but also what gets published, what doesn't get published.

Jess: Yeah, I think all of this stuff is really like, ultimately complexifying our view of the world, right? If we take in a lot of perspectives and stuff. And that's something that isn't usually efficient and easy, right, is to like represent complexity. I mean, our models like can't, like a lot of the times, like that's why we simplify our models because it's so hard to have complexity and then have like a clear point. So I guess sometimes it's, I guess we must, we have to as science communicators resist the need for coherency a lot of time. Because-

Catherine: What about the need for coherency? You should be understood by people, but.

Jess: Yeah, coherency that comes from taking out like contradiction, right? Like, yeah, sometimes when you involve more complexity, you see contradictions and then it makes your story less coherent, you know? So finding a way to tell us.

Catherine: This is completely off topic, but you know, like different levels of knowledge where things are super simplified when you're first learning them in elementary school. And now both of us as grad students, it's just like nothing makes sense. There is no rules in life. Biology is a mess.

Jess: Yeah, it's so true.

Catherine: Another thing I do wanna say though, is I think we should keep in mind our own biases, especially when communicating with the public. And I'd be curious to see if there are any papers that look at this of people's like style of science communication changes, depending on the person they're talking to. Yeah.

Jess: Yeah, that's a really good point. Cause we actually are usually trained to do that, to kind of like code switch and think about our audience and only like talk about things that we think would be relevant and connect with our audience. And I think that's good. Like, I think it's on some level good to like, focus on what connects you and relates you to your audience at least initially. And then maybe not be so bias in what you're presenting as to only present stuff that, you know, like keeps you in good rapport with your audience, right? Yeah.

Yeah, so interesting. Well, we should probably not. Yeah, this one's getting long. So we could cut this off here.