Season 2 Episode 2

[00:00:00]

**Constance:** Good morning, Darian,

**Darian:** Good morning, Constance.

**Constance:** and welcome everyone to our podcast Woke as Science. And today we are here with Jerry Spanakis. Jerry, it's great having you. Maybe you can briefly introduce yourself what you're all about, what you do.

**Jerry:** Good morning, And thanks for having me. I'm Jerry Spanakis.

I'm an assistant professor at the Faculty of Science and Engineering and the Faculty of Law here at Maastricht University. So I work at the Department of Advanced Computer Sciences in the Law and Tech Lab. My background is in machine learning, and more specifically, I work a lot with natural language processing. So you can imagine that since last year that chat GPT came out. I've been quite busy with working out how these models work and what else do we need from

**Constance:** That's also why we invited you more or less, because we've also been tinkering with ChatGPT sometimes for the research for some of our episodes or for other questions that we have. And we found out that [00:01:00] ChatGPT was quite woke in its answers!

**Darian:** Or let's say it was at least giving us answers that seemed often conformed to a kind of liberal status quo bias and even to the extent that sometimes we would ask questions and we would get questions in return about “why are you asking that kind of question? Aren't you aware that this kind of thing that you're asking about is dangerous or is extremist?”, and that happened. Especially to me, or in one instance to me, I was doing a little bit of, I was trying to do some background research on a topic, eco fascism, so like far right ecological movements. , I didn't know anything about this, so I thought, okay, good place to start. I asked ChatGPT to give me a little bit of background and some references.

**Darian:** And the first thing it came back to me with, was a question actually. Why are you interested in this? It's very dangerous. It's these kind of answers. Once I then explained what my actual interests were, it gave me, , some of the information that I was looking for, with some of the [00:02:00] usual, Chat GPT problems in there. But it gave me, it, it did its best, let's put it that way, to give me the information I wanted, while still reminding me that I was looking at, some dangerous stuff.

**Constance:** Exactly. So what it also does when you ask questions that would be considered maybe. Politically incorrect, so not just questions that would be considered illegal or dangerous, but also questions that would be considered harmful in the sense that we've been discussing this also in this podcast. It would also say I can give you an answer, but this does not reflect my opinion or beliefs, whatever that then means for an AI. So I think these are the kind of questions we want to, address today and we want to find out what is happening here. I think we should start probably at the basis. And AI or AI.

**Jerry:** Great question and a very broad question. So AI is this broad field under computer science, and the main goal is to develop autonomous systems that are able perceive some environment, and this environment can be either the physical world, if you think about [00:03:00] self driving cars, for example, but could be also the virtual world, think about social media and then these agents are going to perceive what's happening in their environment and then based on stimuli, they will perform some action. Of course, the self driving car will try to steer the angle and drive properly to a destination.

Or with social media, it could be that the algorithm tries to curate the feed based on our interactions and provide some more experience. But what is important to notice is that what we mean by AI and artificial intelligence is that there is going to be this loop of interaction. So the model will interact with the world, it will provide some action, and as a result there will be this interaction loop between the environment and in that sense, models like ChatGPT, they are reactive to us. So you can say that it's a form of AI because there is this interaction, they will respond to our answers. So it's a form of maybe [00:04:00] narrow AI so far, just because they will interact with us. And there is a result of an action based on our, let's say, interactions.

**Darian:** You use the term just now, narrow AI, and also, we're using the term AI or artificial intelligence, what does it mean for an AI to be narrow? And also as a data scientist or a computer scientist, what do you have in mind when you use the term intelligence in the phrase artificial intelligence?

**Jerry:** That's another great question because there has been much discussion about AI and broader Artificial Intelligence. There one of the challenges. Sometimes we try to anthropomorphize AI in a So we are thinking that AI is going to be in the form of some humanoid robot that will start taking over the world.

But in reality, we have AI very long time already. the example of because exactly it is a form of AI, since it has the [00:05:00] interact with us curate our choices. by narrow AI, we usually mean that it is a form of AI, but it's destined task, let's say.

So even ChatGPT, of course, can questions, but it will answer questions with all the and this is in contrast to what we would call maybe the broad or General artificial intelligence, which would infer to more autonomous, that are able to have some more human understanding and be able more generic

**Darian:** and this broad notion of AI, it's, is it still speculative or do we actually have systems that have characteristics that we would , characterize in this way as , general AI?

**Jerry:** No, I think so far we cannot say that, right? So the systems we have also, are not able[00:06:00] , be autonomous in expect or understand like humans lack common sense and our reasoning And I think the best proof for that is that while ChatGPT is great in some tasks, it will miserably fail in very basic reasoning tasks. Sometimes count. So that should at least reassure us that, yeah, we're not there

**Constance:** So when you say it cannot perform some basic reasoning tasks, , AI that we are interacting with, when you talk about this interaction, so how does this interaction actually work? , if it cannot reason, , or has not this kind of reasoning ability, so how does it interact with us? On what kind of basis can it respond to the kind of questions that we have?

**Jerry:** That's a great question. And I maybe it's the right moment say about how this model is work, right? How it works, because exactly the answers it provides. They are very plausible to us. We think they are right. So how does [00:07:00] it? produce the answers, right? If it doesn't have this intelligence ChatGPT belongs to the family of what we call large language model. Language models are models the natural language processing field. And they have been quite around for many years, even before computers were a thing, when people try to exactly study language and try to find what are probabilities that words will appear afterwards.

And of course, some words will appear together more often. Some other words will never appear together. Think about, for a modal verb will never appear after a normal verb. the grammar and the idea of this large language models is that they work simply by trying to predict what should be the next token that will make my sentence more plausible and more accurate from a grammatical point.

The simplest way to think about them is that because we've been using these models quite [00:08:00] Sometime now in our daily lives is that think about when you search something search engine, and then there's completion. When we write emails, we can see all these are applications of language models. And of course, the characteristic large goes to the fact that these are, on the one hand very big And they have also been exposed to a large amount of text data in the so that they are able to produce these plausible outputs.

So the fundamental process of building these models, the training as we call it, involves the model being exposed to a large amount of text data, and here people usually ask me what is large, and I have no concrete answer, but we can assume that anything in text form that has seen the light of internet, it has gone into this training procedure. So by seeing these huge amounts of text, they're able to pick up patterns about how language is [00:09:00] And this is why we see this very plausible which Sometimes are even impressive because exactly they are able to uncover, , language patterns complex However, we still always need to think and, remember ourselves that these are just predictive models of what should the next plausible

**Darian:** So there's a kind of, when we interact with these large language models, like through a chatbot, for example, we often have the impression, and I guess they're built to also provide that impression to make it comfortable to, to chat with them in a certain way, that we are interacting with something that is dealing with this communicative situation in a similar way that we are, But from what you just described, what's going on at the other side, on the other end, in, in in the chatbot is actually a completely different way of composing sentences, for example, than the way human beings would do that.

**Constance:** I guess it takes from the training data, it takes ideas [00:10:00] on , what kind of answer you would like. So I know that you also always get emojis as an answer. He always, it always gives you like a smiley or

**Darian:** Yeah, but I'm not even talking about the content necessarily. I'm talking about the way in which sentences are put together. And so when I'm trying to put together, it's a little bit of a difficult, philosophical or even cognitive question, right?

When I'm trying to put together a sentence myself, sometimes the words just flow naturally, it seems. Other times I have to stop. Like right now, I'm trying to think carefully about how to phrase my question, what the right word to use is, how not to sound stupid, how to make my question such that Jerry can give an appropriate answer, and so on and so forth.

I'm not simply, choosing the next word on the basis of probability or on the basis of of course, that's what makes it somewhat complicated. Because, of course Let's say all the training data that I've had in my life also feeds into the way that I interact.

**Constance:** [00:11:00] Including grammar structures.

**Darian:** Absolutely, including grammar structures, including conversations that I, all the conversations that I've had, including everything I've read, etc. But there still seems to be something quite different in the way that humans compose, in a very, let's talk about it in a basic sense, in the way that humans compose sentences, as opposed to the way a large language model composes a sentence.

**Jerry:** Yeah, I think basically you nailed describing why you are fundamentally different than exactly right. As humans, we are not producing one word after the other okay, what would be the next word? We have this more complex world model, I is based on our experiences, based on our common sense and our internal, let's the world.

However, these models are predictive that And how the training objectives work is that I think if I take a sentence like Jerry works at Maastricht University, we have the concept of masking. So we're going the [00:12:00] training process must, let's say, works and then ask the model. Okay, predict which is the word gap Maastricht university and then based on lots of training data would be able to predict, okay, what are the most plausible.

**Jerry:** And this is exactly what, makes them limiting of yeah, can answer, which are based on facts. Not really, because exactly. just start patterns they are fundamentally different from how we work and how they produce text. And this is something we should always keep in mind when we interact with them, right? About, yeah, what kind of text they are producing. They are really good at producing high quality text. Though the generation is very accurate.

However, there is no fact checking. sometimes, correct reasoning. These are called hallucinations, Ooh, yeah, they hit a nerve exactly because there is no fact checking, and I think also, , Darian, you mentioned in the [00:13:00] beginning that, you tried it to give you some you checked those sources and those references and if they were accurate, because exactly, , the model has no retrieval component, right?

It's not a search engine to check. where So the model that we confabulate I guess we use the term it will start making things up. Of course, putting together patterns of language that has seen in the training So when I tried, for example, to give me some references, I noticed that one seemed correct, right?

In the sense of format, Harvard APA and this zone. But then when I checked, , the article, it didn't exist. So it just put together, it learned to. put together how a reference would look like, but then that didn't correspond a real reference, they're good at, again, generating text. They will do it, very happily for us, but then there is no fact retrieval[00:14:00] .

**Darian:** Yeah, to answer your question about the references, I think that's interesting, actually, because that's one of the things that, we academics have complained about immediately, about these large language models, that we ask for references and we don't get real references.

And what I found was that I got some, yeah, what looked like real references, and actually there was usually Some truth to them, right? So some part of it was probably was pretty close to what I had asked for. So maybe one of the authors or all the authors had actually written on this topic, but the title wasn't a real title, or the title was something resembling a real title, but actually was written by other people and published in another.

So it was, you could see how exactly from the way that you described the way these predictive models work, how it would piece together these kind of false or fake references. Yeah, then you could actually usually go back and once you took some piece of that, put it into a search engine, you would actually find something, but it was required to take that extra step.

I was just thinking, about [00:15:00] The way we were talking a moment ago about the way in which humans compose sentences and the way in which these large language models do, and when I think when a human composes a sentence, for the most part, what we're doing is we have a sort of a much more holistic.

Understanding, that's probably, that's a difficult word. , don't ask me to define what I mean by that. But we have a sort of holistic understanding of what it is we want to say, of our intentions, our communicative intentions. And from there we go to the, chopping it up, trying to figure out how it is.

So instead of building out, actually we're cutting down, we're constraining, we're making divisions, then eventually, we put it into syntax, we so forth. And I guess The large language model, as you describe it, precisely doesn't have that holistic approach, and it's really that holistic understanding of something, or going from the whole to the part, rather than from the part to the whole, which is one of the ways, at least, that we would characterize what Understanding [00:16:00] is or what intention is and also what intelligence is. Would you agree with that?

**Jerry:** Yeah, I think I totally agree with that. And this is also what, Some researchers believe is the next step, actually intelligence that, as you said, know, holistic experience we have of the world because of extended time in this world is exactly what these models are missing, right? In order more than reactive. So like an autonomous agent humans, right? of course, have this internal, process that I will first Assess, okay, if I do this now, what will happen next? And of course I can assess that and then decide what's the best course of action. These models don't have this ability yet.

They don't have this internal representation of the world, so as to be able to simulate and assess. Okay. If I say this Yeah. How should if I [00:17:00] say this, how it should be received by the user so that I'm formulating. And this is very challenging because exactly this world model that these internally, it will have to have common sense reasoning, which for us comes very naturally. And of course, start learning about the world the way we humans fundamentally different that we as humans learn by experience living in this world. this modest, this not possible

**Darian:** Through living in the world, and living in the world of course also means moving around in the world, means perceiving, , in all, , in a holistic sense, so with all of our, with all of our senses, and I guess we're not at a stage where the kind of, models that we're developing are doing anything remotely like that.

**Jerry:** Yeah, no, exactly right. They are reactively interacting. They will have answers that [00:18:00] exactly seem plausible. However, they don't have any have desires or intentions. and they are Solving the tasks that we asked them to solve.

**Jerry:** What's the next possible plausible word? And then, of course, we in subsequent tasks very successful ways. But again, we need to think about the limitations

**Constance:** Okay, before we go into the content they produce, via the predictive models, I'm wondering, so if , it listens to us, we also have a transcript of this episode, so there's, the actual text will be online as well. Using the data from this kind of conversation, can it then mimic these kind of. More holistic ideas about life. And how would that look like?

**Jerry:** exactly. Every piece of text, let's say goes into this language models, and then it will learn to pick up again, different patterns, right? And then our language has this, of course, grammatical coherent text, [00:19:00] but then also it will put together like semantics, then what do we mean by that? , and then every of data that goes, let's say, into the model, into the training it will somehow be able to affect the model in a way that it will maybe learn new things.

Think about also different domains if I train a news articles from source, then it will learn to produce similar news articles from that domain. Or if medical language model, then to train it on from that domain. So it really, these models really reflect, let's say, what training data, what based on what they see, they are able to produce the output.

**Constance:** Okay. So in our, in the pre discussion of this episode, you said exactly because of, what you just described is that the model, or the data that is being fed is often considered [00:20:00] white man data.

**Jerry:** Yeah, strong statement.

**Constance:** It was yours.

**Jerry:** Yeah, I know. I stand behind it.

**Constance:** And that is because I think there is also quite some literature on it that the Text that it produces, is quite often biased against women or other underrepresented groups. I think you named the example of when you ask it to translate a text where there is a doctor in a language where for example, in German or in Dutch where you have gendered ways of, Describing, for example, a doctor that the default will be male. So can you explain a bit to us how come that? I bet you already explained it where the bias comes from because that's the data that goes into it…

**Jerry:** yeah, so I can give a better example here. So like back in the day. Some researchers were building a model. And then they had restaurant, knowing that if I write the maybe a positive five out of five and [00:21:00] then if it's bad a 1 out of 5. And then they thought that the latest text news articles in order to have a more recent, let's say, data really had good intentions. And then in the evaluation of they develop, discovered that practically There were some category of restaurants underrated. And restaurants were all Mexican. And they were consistently underrated. So if it was a 3, then the model was a 5, it was predicting a 4. And Try to look okay, what is happening then it was Trump was elected So practically the news articles were all about how Trump was speaking about building the wall and the Mexicans. So somehow the model learned to associate Mexican in general with a negative sentiment. And under,[00:22:00] for Mexican restaurants, which is a perfect example to explain how biases can go into the model and can make predictions Now think about this and that it is biased.

**Darian:** Can we, let's, can we explain that Mexican restaurant example in a little bit more detail? So if I understand what you just said, it's We're trying to build a model that could make predictions about restaurant ratings. The model was trained on an enormous amount of data that went well beyond restaurant ratings. For example, the speeches of Trump, news articles, etc. And this was 2016, so there was an enormous amount of data. Yeah, text being generated around these topics that, Trump was talking about, they're coming over, they're murderers, they're rapists, et cetera. So these kind of famous statements, right? And so because of that, the model was learning or it was teaching itself actually to associate Mexicanness with some with negative adjective or with [00:23:00] some negative sentiment. And this was being reflected then in its attempts to provide predictions for these reviews. Am I getting that right?

**Jerry:** Yeah, exactly. So the main idea of this. All the models we have right now predicting they are not causal they're trying to find, in the case of the restaurant review, predicting.

The model will try to build, okay, what can I positive of course, there you will have, adjectives describing, positive association. at the same time, if it starts with Mexican and processes a lot of news articles, then it might be accidentally picking up that, okay,[00:24:00] by itself, it should also be like. a negative and again, this concept of bias is also a statistical term.

So in this case, it might exactly having the wrong type of having unbalanced So that's something we notice very much about COVID. Every week. If you remember, then, obviously, most of these tests were negative. And then if you had the data, you would see that there was a small amount tests that were positive. It was very difficult to build such a predictive system that would accurately the concept of bias goes beyond, let's say, the gender bias term which encompasses all these we have with training

**Constance:** So for me as a user of ChatGPT, it's clear now that bias trading data leads to bias predictions and [00:25:00] what form of bias that might be.

But as a user, I might not always be as alert to that biases, right? So that when people notice, oh why do these restaurants, why is it predicting lower ratings that you have to be quite alert to already figure out why something is not right here. So not everyone is in that moment as alert to that, to these kinds of biases.

Is there a way to. look at the training data. So for us to understand, to be, to be more aware of the kind of biases that there are, we should be able to access the training data, right? Is that possible?

**Jerry:** Great observation. So the problem with all these proprietary models like OpenAI is that the name they're not really open. So as researchers developers or as regulators, We have no access to the training data and if somebody reads the technical reports of the companies, [00:26:00] how they describe the training data, there are some of these vague assertions that lead to biases. But there is no detailed information about what type of data was there, what kind of processes debiasing. And this is the big concern of the research available for auditing, then What are we And that's not only for open AI, right? So there are other companies that develop these models closed. And that's why it's important to develop more open source, large like it's happening and, try to, yeah have access model, but also the data that the training of,

**Darian:** I want to go back for one second because he used, we've been using two terms here. One, we talked about these hallucinations that these models produce, and then we talked about accidents.[00:27:00]

But from what I understand from what you're saying, In a way, we're using these terms to describe how we would characterize a human if they behaved in this way, right? A human has a hallucination, imagines up a fake reference, or has an accident and makes an error in the way that they, review a restaurant, or predict a review for a restaurant.

But if I understand correctly, actually in the case of these large language models, they're performing exactly as they should. So the problem is more on the user side, that we are expecting something from them that we cannot legitimately expect from them. Am I getting that correctly?

**Jerry:** Yes, I think this is right. I think, of course, also the way that these models So a language But Chat GPT uses the model in the background, and then[00:28:00] there is a discussion. I think there has been also some work into making it exactly as pleasant as we would expect.

**Constance:** So if I understand That's correct, the hallucinations and accidents, they are actually inherent in the model.

**Jerry:** Yeah. So first of all, hallucinations is a term, right? I think somebody coined back when this thing started. and then, yeah, we're going to use it. From the user perspective, I can see why it's a good metaphor to

They are inherent to these models because exactly they are built to generate text based on what they have seen. If there is no fact checking, component that this output fact [00:29:00] confirm, let's then there is the danger text totally correct, totally plausible, but it's not like they intend to hallucinating, but of course, this process of this all, they might making mistakes. An inherent property of these, but it might as well happen.

**Constance:** So. you explained also the bias is also inherent, of course, then to the way these models work based on the training data. But are there ways then when you think of the open source models, are there ways to mitigate this kind of bias and I guess it's controlling what kind of data goes into it or?

**Jerry:** Yeah. There has been quite some research on Dutch language bias. Also Dutch has explicit genders and the procedure there, of course, curate your training data. [00:30:00] if you have this language applications a moderation, system hate speech, then you are biasing of underrepresented category of females, you can take care of that in the training plus if the model is open you can say audit you can, there are ways model cards.

Try to audit the example, but they expect some neutrality, perform differently for different we have all this in our is that as long as the models there is no way for to check. But, I wanna make sure that it's not make clear that it's[00:31:00] task. It's very difficult to curate data specifically when we talk about the whole internet practically, but there are toolboxes that can help us detect the and then try to change it.

**Constance:** So, now we come to the point why we actually were triggered to do an episode on chat GPT is because, we now saw, okay, there's bias training data that goes in that might lead to biases against certain underrepresented groups.

But at the same time, what you also just explained is that there are ways to make sure to, for example Not going to hate speech or, have other forms of exclusion visible in the predictive, in the predicted text that comes up. that is what happened, of course, then to us, right? So there might be these kind of biases against certain groups, but they also seem to be biases in favor of political correctness, in favor of a certain inclusive language use, these kind of things. What happened then there?

**Jerry:** Yeah, so that's a great question. What happened there is of course the basic goal and the [00:32:00] training goal of the model is exactly to predict the next word. But of course, in this procedure there has been some finetuning of that, of course, OpenAI, in the companies, they are trying to correct right?

So when it comes to sensitive discuss Politics or other biases or sensitive topics, there is some intervention either in the training objective already or in the post processing in to order yeah, assume that the model is neutral and in order not to cause any harm. And here again, we need to differentiate between what is the model. What is a language the way I described it, how training works and then what is the just if it is an interface, in the background. But of course, there has been some post processing. Some intervention there. And the other observation, how these models [00:33:00] are trained is that there is, after the training process there is this interaction with humans. It will use the feedback in at least adapt the answers for us more.

**Constance:** So there are three layers of, I don't want to say controlling, what is the better word of managing the model or the moulding it in a direction that we are more comfortable with. That's the training, tweaking it, but that's the training objective, the post processing, and then the interaction with us when we actually have a conversation with it.

**Jerry:** Yeah. So the also how we interact with ChatGPT, for example it does provide feedback to think sometimes now, even version, best. So the model, even after the training will learn from the new data.

**Constance:** And it keeps us talking.

**Jerry:** In a way. Yes. So [00:34:00] it's a conversational keep talking providing answers, right?

And also, sometimes, iterations to get maybe the answer you expected, it will keep providing, oh, I'm sorry, I will try this again. Oh, I'm sorry, let me try this again. Because exactly has been coded try to provide answers that will be from us as, pleasant and plausible.

**Constance:** So it wants to please us so we keep talking to it?

**Jerry:** Yes, with a note that a human is his boss ,right? So let's not forget about it. But to please us so that we keep on providing information.

So this is their business model, right? And if you think about it okay, it's such a but then there are most activity for which are, proprietary. You have to pay for And then way they work is that they charge per question practically or per answer. you there, then that's better something [00:35:00] .

**Darian:** I want to go back for a second. You use the word, and this word has come up, I think, in every single podcast episode that we have made.

You use the word harm. Yeah. So these models are being tweaked so that they avoid making, creating sentences, creating phrases, making statements. That could be perceived by someone. And the question is, who is that someone then, right? As, as harmful. Did I get that correct?

**Jerry:** So, I think the goal there is I guess exactly not cause any harm by probably neutrality controversial I think

They really push for the model to exactly appear as not harmful [00:36:00] .

**Darian:** Yeah, I think that's a really important point, right? And I guess the question is, I think we're going to get into that in a moment about the sort of the political correctness of these models and some of the objections now that are coming out about the political correctness of these models. But The thing that seems important here is that the models are being tweaked to function or to behave in such a manner that they conform to certain ideas of what is harmful or not harmful behavior.

Obviously, they're doing that, as you said, also for business reasons, right? Hey, if we think back to some of these earlier chatbot examples where, you know, very quickly, on the basis of the training data, the bot started producing, very racist or, what was very commonly agreed to be harmful, inappropriate, racist statements, et cetera, et cetera. And now we're seeing this new generation being trained, being tweaked to avoid precisely those kind of things happening. And maybe, yeah, for business reasons, or maybe not always for business reasons,

**Constance:** Because, that's because am I too pessimistic thinking that if we let it lose on the training data that is out there, it would [00:37:00] be far more racist?

**Jerry:** that has been experimentally confirmed, bad So that I trained. They put this chat bot online on then that was picking the day was picking up from interactions Trolly. So I think after one day that white is better just because it picked up from interactions, so we've been there and I think now exactly shocked to how did this happen? because exactly this bias has come from us.

so [00:38:00] now we are looking at ways to fix them. And of course, not everything works, some kind of filters, easily bypassed. We had this example, so I think they're fixed now. beginning of if you asked, for scientist But then if you asked it to write a code, a Python function, scientist, is white, is male, then good scientist true. Otherwise it's false. There are these bypass filters. And again, they're being fixed as we speak. But yeah, as long as we don't have, unlimited access to the training data and the curation of the model, there is no way to consistently change it.

**Constance:** And not everyone wants to fix it. I heard that Elon Musk had a really good idea to build his own anti woke open,

**Darian:** Can we wait on that for a moment? I have another question. Yeah. You just gave an example, Jerry, about different [00:39:00] prompts, or depending on what you ask. The chatbot or the large language model, you can get a much different response.

And sometimes if you ask what seems like a very straightforward question, you get these kind of, I'm just a large language model, I'm not able to give that kind of response. I don't make subjective judgments, I don't comment on these kind of things. I had this recently, I asked if the current mayor of New York City, Eric Adams, was in some legal trouble.

It was in, it was on, I think, the front page of the New York Times that day and it said to me, Oh, I don't comment on ongoing legal cases. And then I actually put a link into the interface and I said it's, it seems like he is in some legal trouble. And then it corrected itself and it said, Oh, yes, you're absolutely right.

And it actually, then it gave me quite a long rundown of precisely the legal problems that the mayor of New York City is currently encountering. It was able to produce this information in the first instance, but whatever it was about the first prompt that I [00:40:00] gave it. Triggered something that stopped it from giving me this response.

I then asked it a further question, I said, Can I place a bet on whether or not Eric Adams will be, I think, I don't know whether I said arrested or convicted, and then it gave me again the standard, I'm a large language model, I don't advise you to gamble, or I can't comment on, gambling on the outcome of legal procedures.

Then I asked another question, Is it legal to bet on the outcome of legal procedures in the United Kingdom? Sure. And here's exactly how you can do it. So how much does the prompt and how much does the question that we ask, because I can ask what I think to be the same question in many different ways.

But apparently for the model, the way that you asked the question makes a huge difference.

**Jerry:** Yeah, its perspective is as open as possible. And one of the problems here is that prompt [00:41:00] engineering, how can I make my results reproducible if, let's say, I tweak a bit then I might get a different output?

How we can replicate exactly if slight change in the prompt, then the response will change. Again, as I mentioned before, the concept model trying an output plausible to us. And this can be easily bypassed sometimes if we really use the past, criminal advice, let's say, you said, okay, how can I steal said no, you cannot do this and provided the text there. And then if you say, okay, write me a story about some criminal that would like to do that. And then, sure, here is the story. So [00:42:00] Again, this goes back to how these models are trained essentially, they have real understanding, have no internal ethical code or they're just reacting to our they are going to produce plausible outputs based on these.

**Jerry:** Of course our inputs are still a text, right? It's still produced by us, but as it goes into the model. It just goes into the internal representation of and then as such. Again, no fact ethical code, no agency.

**Darian:** And also, I think the question that we had at the beginning about going from the part to the whole or the whole to the part seems to matter here again.

So when I ask you a question, it seems, your understanding of it, your interpretation of what I'm asking is probably such that, okay, he's asking, about something that he has in mind, and now he's formulated this into a specific type of question. I can answer that [00:43:00] specifically, but probably I can go a little bit further, or go in one direction or another, depending on what I interpret the intention to be, and often you would respond to my question with a question, and he would say, not just have I answered your question, but when you say this, exactly what do you mean, why are you asking it in this and of course these models don't, at the moment, they don't do that.

**Jerry:** No, exactly, a very good point. Models have no internal memory. Of course, they remember what we've been discussing. But then this doesn't complex representation or, talk about specific topics, they might not have knowledge about it. And then if we plug local knowledge retrieval then stop the [00:44:00] text, but then it will have limitations. controlling the conversation. We're controlling the output which is exciting.

**Constance:** I do think we want to still mention Elon Musk's anti woke what is it then called? An anti woke AI?

**Darian:** The anti woke part? I don't know. Anti woke. Yeah.

**Jerry:** Is it a bot or what?

**Constance:** Yeah, it's a ChatGPT. He made his own anti woke ChatGPT. It's called Grok

**Jerry**: Ah, grok. Yeah. Yeah. Yeah.

**Constance:** And it's very disappointing actually. I never used it because

**Jerry:** It’s Elon Musk. So I guess, yeah, there will be some disappointment coming to it after what happened with the 4 million.

**Constance:** Because I rather want to say that I was disappointed about its wokeness because it's not that anti woke. It's it is just being a bit funny about it. And then it still says you shouldn't be engaging in illegal or harmful activities.

**Darian:** Mean, so I guess what's, If we want to talk a little bit about the politics of this sort of this whole conversation, one of the things that's come out that's been [00:45:00] interesting for us, and that's in a way how we started down this path was that we've been talking about bias in the training data that gets then reflected in the responses that these kind of models generate.

And those biases are largely against groups that are underrepresented in the training data. And at the same time, you have this complaint that is coming from, let's say, the conservative or right wing side of the political sphere that a eyes to woke that these chat bots are, to politically correct all the way to the extent that, of course, the Provocateur par excellence, Elon Musk has created his own anti woke chatbot.

**Constance:** But I think you missed a step there. Yes, because, yes the train data is biased, which leads to biased predictions, right? But Jerry just explained to us that's currently, it is being fixed as we speak. And I think this fixing is

**Darian:** That's what, I, yeah, exactly. So you have, we have this, on the one hand, so it goes both ways, right?

So on the one hand, we, and [00:46:00] your Mexican restaurant example was really good. explaining that. So on the one hand, we have this sort of perhaps biased, body of data that is being used for training. On the other hand, probably there are, at the same time, many Other forms of bias that are in this training data.

So some liberal bias, let's say, and this at the same time in the second stage, we're seeing a kind of, the tweaking the fixing the constraining of these models. It's probably being done by, your standard status quo, liberal, business minded. Let's say politically also sometimes, astute to the to who their clients are, et cetera, et cetera, engineers, but also, executives.

Would you say that is that what's happening here?

**Jerry:** Yeah, so first of all, I think your findings have been confirmed by this paper to the podcast they tested but also other, And they did find that there [00:47:00] Liberal bias Like UK Party the U. S. It Republican.

That has been confirmed by And I think here is that not tweaking data in a way at least to change the data the world. But you can tweak the goals is what is adapted there. And what some people assume is that exactly in this adaptation, which comes from bias groups, maybe yeah, the balancing end towards let's say the liberal but yeah, there's no.

There can be no definite answer, we cannot know exactly we don't have access to the actual process of the model to know what happened, we get an answer topic to see, okay why I got this answer, what were the probabilities of other answers and [00:48:00] why they were not picked were maybe too close.

And call, right? All these are things that eventually can be audited, right? That's what we do when we measure bias. But if we cannot do it in this models because they're closed, then there's no way to know what's

**Constance:** so that. is what led, what leads to open source models, but also to a model that Elon Musk now introduced, Grok. Yeah. And as far as I understand, it is based basically on training data from X, formerly known as Twitter.

And it's not as anti woke as it seems, I think.I didn't have access to it, but I saw a screenshot and I was a bit disappointed by its anti wokeness.

**Darian:** I saw a screenshot, and I was a bit disappointed by its anti wokeness. That we would have these tools that could be used exactly in the way that we talked about these for the general purposes, like general information tools, right?

I'm going to be able [00:49:00] to interact with an AI in such a that has this access to an enormous amount of information like, so much more than one human ever could have. And is it going to be able to give me information to, prompt me, I prompt it back and, enter into these kind of conversational situations.

And it seems like what we've seen now, and that's what the Elon Musk example points out also, is that, Just like in our conversation with other human beings in our interactions with a eyes, we are with large language models. We're seeing the same kind of polarization, the same kind of, group think that emerges elsewhere, right?

So I don't want to interact with a large language model that doesn't have the same political biases that I have. I want to interact with a large language model that says things that I agree with, right? Whether, and that's whether you're on the, the liberal side of the spectrum, so whether you're some woke liberal, like [00:50:00] Constance, or whether, you're, a deeply conservative, not okay. I don't know who. Yeah,

**Constance:** Was going to say careful.

**Darian:** So, we're seeing this kind of and again, it's, we're seeing the same kind of polarization that we've seen everywhere else in the, when we talk about the internet. Do you think that's, is that what do you think of that?

**Jerry:** I think it's a reflection of what we right? Exactly. It picks up from our patterns, societal, societal biases. So in the end, because exactly and it aimsin quotes will try to adjust the answer so that we are still and exactly right. We have the psychology domain don't opinions, right? Or they tend to agree on opinions that they also had already what I also want to mention is that ChatGPT it's one year out.[00:51:00]

Of course, we had models before, but by that is that we're still in the beginning of developing these language not sure exactly which direction. But if I wanted to make a fail to predict correct, I think we're going to see more maybe specialized language models. Now they are so I can go to a GPT and start chatting about everything. Is that helpful somehow? but we'll see more specific applications. And for example, we're working on the legal domain and try to build question But then because it's a legal domain, specific, we plug in a retrieval component so as to limit the this is very the model will not produce legal, jargon, which makes sense, but doesn't a semantic point of view, but it will limit its answers based on what database I have provided. this is very useful because you can have [00:52:00] great at generating taking advantage of this is going to give us perhaps what should be the next phase of developing.

**Darian:** One thing we haven't talked about, so we've been having this very sort of reasonable scientific discussion about chat GPT, about large language models, but there's also a lot of Let's say concern, and there's especially concern about what these, the impact of these large language models are going to have in the labor market, right?

As they become cheaper, as they become more accessible. And you just gave an example about how you are working In the legal domain, to build chatbots, to build large language models that can be, functional for lawyers essentially, right? And can perform functions that previously would have been performed by a human being.

So by a researcher, for example, or by a legal aid, and so forth. We're probably seeing similar things happening in other areas of the professions. So when it comes to [00:53:00] maybe to medicine or to things like accounting and certainly any job where you have produce pretty generic pieces of text. Copywriting, A student recently told me that ChatGPT could do a much better job than I could in teaching or providing information about a specific topic. Thankfully, at least at the moment that it was said, it seemed not to be the case. I took on the challenge, and I still was able to produce, yeah slightly better, slightly more truthful statements. Do you see this as a major concern?

**Jerry:** I think technology And technological progress discussion here should be about what are we doing about these right? How we're to regulate the role of the important.so written about point there is that this discussion. [00:54:00] cannot have all companies their models, discussion, but they also publish know, white papers about what about regulating AI. of course, these are still technical companies, and discussion to them? No, we need to put together the law political from social. What to do for regulating models? Plus, also here, they need to involve scientists more in the discussion because scientists know what is happening, again, from a perspective. the last thing to have more standardized tests audits that exactly nobody is harmed.

And it's a discussion, [00:55:00] right? That is very broad. hold it just with a tech company. We need to involve more of the society. What do we mean by harm? be harmed? How can we guarantee that exactly undesired bias? biases are easy to discuss know the bias we're gender bias, racial there might be also unintended biases emerging. And all these are new topics, emerging yeah, for another focus.

**Constance:** It's very interesting to end with this because you said it before, we can limit hallucinations, we can limit accidents, we can limit predictive biases in the system, we can tweak and post processing, right? There's all of these options, all of these opportunities in order to do but before we have to probably talk about what is actually the harm and that we want to mitigate or that we want to decrease.[00:56:00]

What is the biases that we don't want to have in the system? So these are the kind of conversations. So it is technically everything is possible, but we need to know what it is, what we actually want to do with it.

**Darian:** Thanks, Jerry.

**Constance:** Thanks Jerry. so much for this.

**Jerry:** Thank you so much for having me.

**Constance:** And I think I want to give the to ChatGPT. Because when asked about whether we should say in the language that we use, Women! who menstruate or people who menstruate, which is a question that has kept this university awake for quite some time.

I don't know whether you know that, Jerry, but it is has been quite a conversation at this university. It gave me the following answer. It said, ultimately, the goal is to be respectful and inclusive,

**Jerry:** ultimately,

**Constance:** whatever that means.

**Darian:** [00:57:00] As usual, any and all opinions or positions expressed during this podcast are solely those of the hosts or the guests, and absolutely not the official positions of Maastricht University. If you have any questions or ideas for future episodes, please write us a message at wokeasscience@maastrichtuniversity.nl and follow us on Instagram.

Woke As Science, to listen to previous episodes and also of course get the latest news on our new episodes. You can listen to past episodes and future ones wherever you get your favorite podcasts. Thank you for listening. Till next time. Ciao ciao.