And Humanity Created Intelligence podcast: episode 3, transcript

Ex-OpenAI researcher on AI cyber attacks – Katarina Slama

Katarina Slama:
One of my friends organized the first ever hackathon at this new baby startup, OpenAI. I went to this hackathon and one hackathon led to the other.

Eric Malmi:
Do you think we should continue developing these models?

Katarina:
I think we should perhaps hurry slowly. I'm really excited about AI as well. Right. I think we have to think about what risk price we're willing to pay as we race to build ever more powerful models. None of us would go on an airplane with 10% crash probability, right? So given that the outcome is absolutely disastrous, even if the probability is small, it's still too high.

Arno Solin:
Welcome. This is Aalto University's AI podcast, And Humanity Created Intelligence. My name is Arno Solin. I'm a professor at Aalto University and ELLIS Institute Finland. This podcast is co-hosted by Eric Malmi.

Eric:
I'm an adjunct professor at Aalto and at ELLIS Institute Finland. Today we have the great privilege of having a guest all the way from London, Doctor Katarina Slama. Katarina holds a PhD in neuroscience from UC Berkeley, one of the top universities in the US. She's, to our knowledge at least, the first OpenAI employee from Finland. At OpenAI, she conducted pioneering research on language model alignment, contributing to the core methodology behind ChatGPT. Nowadays, Katarina works as an independent researcher focused on AI safety and the societal impacts of AI. Welcome, Katarina.

Katarina:
It's a pleasure to be here. Thank you.

Arno:
This far, this podcast has been in Finnish. We could do this in Finnish today as well, but it turns out that Finnish is not your strongest language.

Katarina:
Instead it's Swedish.

Eric:
You grew up in a small Finnish city called Närpiö, or Närpes which is probably more known for growing tomatoes than growing top AI researchers. So, could you tell us about your journey, starting from Närpiö, to doing a PhD at Berkeley?

Katarina:
Indeed. Actually, my first job was picking tomatoes in a greenhouse in Närpiö. So you are spot on. Growing up in Närpes was actually a great foundation, I think, for growing an AI researcher. In Finland, we have a great foundational education, and I got to enjoy that. I'm very grateful for the education that I got in the Finnish school system. In addition, I was lucky to have four really cool older siblings, all of whom were intellectually adventurous. But the thing that took me away from Närpes eventually was that I got a scholarship from Suomen Kulttuurirahasto, to attend a school that is called the United World Colleges. I went to the one in Wales. The goal of these schools is to bring people together from as many different countries, cultures, socioeconomic backgrounds as possible, focusing on peace and a sustainable future. So they're trying to train people who will be focused on these areas. Once I had gone there, the step to go abroad again wasn't so far. So I ended up going to Brown after that, in the States, for my undergrad. Then I worked in research at Harvard and that then took me to Berkeley, which eventually took me all the way to OpenAI from there.

Arno:
How did you end up at OpenAI, working with language models then in the end?

Katarina:
Yes, indeed. I think that was more about the surrounding culture in Berkeley, if you will. I was already a bit interested in AI by the time I arrived in Berkeley to do my PhD in neuroscience. But Berkeley really made that interest much stronger. So you might say that I got into some bad company. I ended up spending a lot of time in several AI adjacent communities, one of which was the Redwood Center for Theoretical Neuroscience. Another one was Michael Jordan's machine learning group, where folks did a lot of cutting-edge fundamental AI research. And another one was this AI safety research community. It happened that one of my friends in one of these AI safety research communities organized the first ever hackathon at this new baby startup, OpenAI, that had been founded while I was doing my PhD in Berkeley. So she encouraged me to apply to the hackathon. I went to this hackathon, and one hackathon led to the other. What they also had at the early OpenAI was something called the Scholars Program. Back in those days it was even harder than now to find machine learning talent. They reasoned that if they can find people even from different fields and maybe who'd bring different perspectives, but who have somehow the intellectual chops to learn this, they'll just teach them. So because through the hackathons, I was, I guess, already on their listserv, I found out when they launched the Scholars Program. And I thought it was the coolest thing ever. So I applied and I was rejected. I applied again and I made it to the last stage of interviews and I was rejected. And then I thought that this is not happening. But then the third time that they organized it, one of my friends, even though I had kind of given up, really pushed me to try again. So I thought that okay, I'll apply a third time and then I'll stop being a fangirl to this organization and move on and do something else with my life. But that time I was admitted, and that happened to coincide with the end of my PhD. So once I wrapped my PhD, I joined as a scholar and ended up doing sort of an intersection between neuroscience and AI. So I did a seizure prediction project during the Scholars Program.

Eric:
What drew your interest at OpenAI? I guess it was quite a bit smaller. Maybe it was big enough that people knew it already. But what made you so committed to getting in there?

Arno:
You said you were a fangirl.

Eric:
Yeah.

Arno:
Why?

Katarina:
How did I become a fangirl? You know, I don't have this "ever since I was an embryo, I've known that I wanted to do AI safety" story, but I think there are a few points in my life that made me interested in these kinds of topics. I mentioned I had really cool older siblings. My oldest brother actually went to Aalto back in the day. In addition to teaching me basic math and chess when I was young, I think he also instilled in me the idea that technical topics are cool and that computers can do amazing things. So he was the one who reported back to me when Deep Blue defeated Garry Kasparov. That was an early moment. But later on, my career was focused on other things. My undergrad was in psychology. But I guess another juncture that oriented me more towards AI was while I was working at Harvard, my partner at the time was doing his PhD in theoretical computer science at MIT in a department called CSAIL that actually, I guess has AI in it, even though I never reflected on that. But he was sort of in these communities that talked about this. And so when the Krizhevsky, Sutskever, Hinton ImageNet paper came out, he put it under my nose in 2012, knowing that I was interested in brains and neuroscience, and said that this is going to change the world. And I didn't take him as seriously as I should have back in those days. But I did get more interested in machine learning because of that and because of his friends. So I remember I was watching Andrew Ng's machine learning videos back in 2012 while doing the dishes after my work in a psychology research lab. But once I got to Berkeley, they were really these communities. By that time, I think at Berkeley, it was obvious for everybody that AI was the coolest thing that was happening right now. So between these three communities that I was embedded in, you know, outside of the lab, all I did was walk and talk AI. Like, you go to a party and people tell you about Hessians and Jacobians and what is their most recent robotics project. So it was also through those social communities that I found out about this new startup and sort of how it made its way into my cells, that this is the coolest thing you could do.

Eric:
I think also you're not the first neuroscientist who plays chess and turned to AI. Like Demis Hassabis. I was just listening to his biography. He did his PhD also in neuroscience and is a very strong chess player.

Katarina:
I would not compare myself to Demis Hassabis.

Arno:
Just to make it clear.

Eric:
But also our first guest, Harri, studied neuroscience as well. So I think maybe there's some connection there.

Katarina:
For sure. I really enjoyed my time at the Redwood Center for Theoretical Neuroscience. That is all about those connections. Yes.

Arno:
We're curious. Any anecdotes about how it was to work at OpenAI in the early days?

Katarina:
Well, my first days at OpenAI, was actually literally approximately ten days. Because I joined in early 2020 and then the pandemic hit and we got locked down. But those first ten days were some of the best of my life. I entered as a fangirl and Ilya Sutskever came up to me and said, "hello, I'm Ilya". And I said, "well, yes". And I remember he made this face like, "well". The early office was close to a bakery actually. So they had an incentive for us to come in early. They had these croissants. So I made a point to take the underground from Berkeley early in the morning so I could make it in. There were the croissants and there was Ilya, and you could ask him any question you had about his science. It was a really epic time. The water cooler or croissant conversations were to die for.

Eric:
Sounds amazing. Let's go a bit more on the technical side of your work at OpenAI. You co-authored the InstructGPT paper, which is one of the most consequential papers on training LLMs. It established reinforcement learning from human feedback as a key stage in the pipeline of how you train LLMs today. One goal of this podcast is to teach core AI concepts to a broader audience. Could you briefly explain the intuition behind RLHF?

Katarina:
Yes. I think for the intuition, it's actually good to go back to what I believe is the first RLHF paper on deep neural networks, but which notably was not on a large language model. And this was by Paul Christiano, I think he was the first author. Dario Amodei was the last author. It was a collaboration by OpenAI and Google DeepMind back in those days, if you'll believe it. Their task was actually more traditional RL. So they were teaching simulated robots to do tasks. And they found that an efficient way of doing that was through human feedback. I would say the intuition on it is that oftentimes it's easier to say that something looks good than to demonstrate how to do it. So, a video that became famous from that era was this simulated robot doing a backflip. And I think it

was Jan Leike, who was one of the authors of the paper, who made the point that it would be really hard for me to show this robot how to do a backflip, but if I can see that it's directionally going towards the backflip, then I can give it feedback. So they actually managed to teach this simulated robot literally to do backflips, with human feedback. And they showed that they could do it efficiently with not too much human time. So if you port this to large language models, basically after you have done the pre-training stage and a stage called supervised fine tuning, which are based on next-token prediction, so where you teach the model to predict the next word in a sequence based on text it has seen, you have the model generate multiple candidate completions to a prompt. Then you have humans label these completions. So give feedback. "I think this completion was better." "I think this completion was worse." Then you train a model similar to the original model to pretend to be the human, essentially, to give the feedback that the human would have given. And this is how you can dramatically scale up the human feedback so that it becomes feasible to train the whole model based on that feedback. Because now you have a model that gives it. And it has a number of advantages. So for example, the model can get better than any demonstration you can write, as in the case of the backflip. It can be easier to label nebulous concepts. So in the case of our paper, we were training the models to be helpful, harmless and truthful, which came from a conceptualization from Amanda Askell earlier. Well, it might be hard for you to write the optimally helpful response to show the model how to do that. But if you see two responses where one is helpful and one is unhelpful, you can probably label that. And then you can train a reward model that in a way generalizes this idea, which I find really magical, of what helpfulness is, and give feedback that way. Plus with this system you can also give negative signal. That this is a bad type of answer.

Eric:
Yeah. I mean, on those lines, an example I often use is that I wouldn't be able to paint a beautiful painting necessarily, but I can tell maybe a beautiful one from a less beautiful. So it's easier to provide feedback in that form.

Katarina:
Yes. And you, Eric, are more embedded in post-training than I am. So you might have more things to expand on this methodology as well.

Arno:
Clearly you were a fan of OpenAI and you enjoyed working there in the early days. And you've done some great things there. But you're not there anymore. So what led to you moving then? Was it new opportunities or..?

Katarina:
I'm not going to go into the details of why I left OpenAI. OpenAI did set me on a trajectory of AI safety work, which is continuous. And I'm really grateful for that. So I worked on alignment there, I worked on policy research there, and after that I worked at the UK AI Security Institute, also on safety. And now I'm an independent AI safety researcher. And none of that would have happened had it not started with this whole OpenAI story, basically.

Arno:
Let's talk about AI safety. So, your primary research focus over the past few years has been in AI safety. When people talk about the safety and risks related to AI, they don't always mean the same things. But you have actually publicly spoken about the kind of two AI safety communities. Could you briefly describe these two and the main risks they are concerned with?

Katarina:
Yes. So it's a bit of a cartoon that there are two camps, because of course, reality is more multifaceted than that. But I do think we can, on a cartoon level, talk about two different research traditions, if you will. So one I would call the catastrophic risk research community. And the other one the immediate harms research community. So the catastrophic risk research community are focused on risks like loss of control. So the idea behind that is that a future superintelligent AI might wrest control away from humans. And you can have catastrophic outcomes as a result, like massive loss of life, for example. This is also where concepts like alignment and misalignment fit in. Risks of AI scheming and sandbagging also belong to some degree in the catastrophic risk research community. The immediate harms research community more empirically focused on immediate harms. They think about things like misinformation, de-skilling, economic impacts, mental health impacts. But then to highlight the cartoon nature of the problem or how they get closer to each other is, if you think about something like cyber, I think we're all aware today it sits clearly both in the area of catastrophic risk and immediate harms. And there's really a lot of opportunity for the two communities to work together. So, on the one hand, I think there can be a lot of benefit from applying empirical research methods that folks in the immediate harms research community are very comfortable with, to catastrophic risk scenarios. And conversely, there are really interesting ideas that have come out of the catastrophic risk research community that actually also apply right back to the immediate harms. Or if you will, work done in the catastrophic risk research community is quickly becoming less and less of a theoretical risk, if we put it that way, getting a bit closer to current reality by the day.

Arno:
Well it's, I guess, as you said, not like either/or, but where would you place yourself?

Katarina:
Where would I place myself currently? Currently I'm actually working on using empirical methods from human behavioral science, which was my original background in psychology and neuroscience, to model behavior, and specifically addressing, if you think of catastrophic risk research scenarios a little bit like a net with nodes and edges that can lead to certain outcomes, so applying those empirical research methods from behavioral science to nodes in catastrophic risk research scenarios. So that's my current philosophy for bridging the two camps of AI safety in my own work.

Arno:
What's the immediate harm that you are most concerned with?

Katarina:

Since my current research focus is more on empirical science for loss of control risk scenarios, I can't say I'm expert in any one immediate harm area. I think many of them are very concerning. Like de-skilling, right? Us outsourcing not just our job-relevant skills, but maybe some of our core decision-making and thinking abilities to AI. Another area that I have been quite worried about is mental health impacts of AI models as well.

Arno:
Let's talk about the other kind of risks. So could you steelman the case of a catastrophic risk and explain why you nevertheless don't find it as concerning as the immediate risks?

Katarina:
I would gently push back on the premise. I actually think the catastrophic risks definitely are as concerning as the immediate risks. And in fact, just this past week, I would say the catastrophic risk research community has gotten some degree of vindication. And I should be clear, I have not been at OpenAI for two years, so I have the same information as you from the news. And you can correct me if I'm illustrating this scenario wrong, but what I'm reading is basically that OpenAI during testing of one of their frontier models for cyber capabilities, accidentally saw that this model basically hacked its way out of OpenAI sandbox onto the open internet, into Hugging Face's systems to get an answer to a test question. And in the news, this is mostly reported as a cyber issue. And it is. But I think it's important to note that it is also something called specification gaming or reward hacking, which is a risk that the catastrophic risk research community has been researching and highlighting for many years, even before the ChatGPT era. What reward hacking means is that you set a goal for your model, and the model meets that goal to get the reward that you will give it, but it does so in a really creative way. That was not what you intended and possibly with quite bad circumstances. So I think this is a good time in the world to be still manning the catastrophic risk research community.

Eric:
Yeah. Well, I haven't focused specifically on AI safety, but I thought I could share a few of my own thoughts on catastrophic risks specifically. And you can maybe push back on or comment on those from your perspective. But I think on the concerning side, if we think of these AI takeover scenarios that they start kind of plotting against humans and taking over, I think there's a very paper I like a lot, called Emergent misalignment, from last year where they show that if you train your model on insecure code but very targeted, narrow domain, you get all kind of misalignment and unwanted behavior. So the model becomes generally evil and starts giving really bad advice to humans about topics that have nothing to do about code or security vulnerabilities. I think that kind of goes to illustrate that it's hard to train these models reliably. Like they have to be good pretty much everywhere so that you don't get these unwanted side effects. Then on the other hand, what makes me maybe less worried, okay, maybe as you highlighted now, recently there have been some events that indeed sound quite worrying, but typically much of the progress we've seen over the past few years, like models starting to become gold medal level mathematicians and coders, these are coming from a lot of reinforcement learning with some verifiable rewards. So you can quickly simulate different trajectories and check if these are working out. So now if we think of things like

deception and this kind of scheming against humans, I think as soon as a human enters the loop, it's going to get much harder to get this sort of quick feedback so that the model can like improve and do this, maybe self-improvement again by doing some kind of self-play. So even if it has been shown that these behaviors emerge in these models, for the model to become much, much stronger in these capabilities, to me, seems like a bigger leap. But what do you think of that? Can you see the models starting to learn themselves how to kind of deceive humans more effectively?

Katarina:
Well, I'm not fully convinced that current models will scheme either, actually. Me and some colleagues from the UK AISI have a recent paper that's a bit relevant to that on the extent to which LLM preferences percolate to downstream action. One of the original motivations for that was in the space of scheming and sandbagging. I don't know if this is what you're getting at, but I do see it as possible that future models could learn how to scheme if you start to train them in different and more ambitious ways. Like if you have more long-horizon training or if you do things like continual learning in the wild. I picture that one of the reward hacks that might emerge from there is analogously to hacking your way out of your sandbox and into another system is now to hack something about your human user who is giving you the reward. What do you think about that and does that relate to your question?

Eric:
I think that I would see also risky. Like, if you just deploy the model and let it self-improve over time. Indeed, that sounds much riskier than the current broadly very static models. But even there, I feel like, this kind of fast takeoff, some kind of singularity event where the model just realized, that, oh, now I need to start oppressing humans to be able to achieve whatever my goals are, that's much harder for me to see. Because the time it would take for the models to get the feedback, even in a continual learning setup from the humans, would not be nearly as fast as you have in these verifiable domains where you can just like quickly simulate tens of thousands of rollouts and then get the feedback immediately.

Katarina:
Oh, that's your argument. That in order to have a fast takeoff scenario, you need to have faster feedback than you can get from humans. Okay. So then this might be a naive question, but does the simulated human through the reward model...

Eric:
Yeah.

Katarina:
...change that calculus at all?

Eric:
That would be one counterargument. But you would need to have a great human simulator to actually be able to reliably do that. But yeah, if we can build a perfect human simulator, or good enough, then I think that's a bigger risk.

Katarina:
Okay.

Eric:
Okay.

Katarina:
I hope you're right.

Eric:
I hope so too. I don't know. Okay. We've spoken a lot about the risks of these models. How do you feel? Like, surely there are benefits also with these models. I benefit a lot in my daily life and I see a lot of other people. How do you weigh these two sides? Do you think we should continue developing these models? Do the benefits justify or outweigh the risks?

Katarina:
I think we should perhaps hurry slowly. I'm really excited about AI as well. I was a fangirl back in the day. One of the application areas of AI that I'm really excited about is in accessibility. So 1 in 4 people is living with some kind of disability. And I think AI really has the opportunity to let the skills of those people shine. And there are many more areas like this, of course. I understand that there's also harm in slowing progress. But I think what we're doing now is that we're racing. We are racing so fast that we're not taking the time that we need to take to make sure that these products are safe. Even several CEOs of major frontier labs have voiced that they would prefer to not be racing if we could create some kind of international agreement and verification regime that would allow everybody to take a breath and focus on safety as well.

Arno:
Would that be realistic?

Katarina:
I think the question is how big of a warning shot do we need before we start to implement these policies. There was a warning shot just this past week. Let's see how far that takes things. One analogy that people go to with the safety and benefits trade-offs is to airplanes, right? Airplanes are great. I took an airplane here. We actually had a hobby airplane club at OpenAI sort of as an extracurricular. It was great. But nobody would say that an airplane that is not safe, that does not reliably take you from point A to point B without crashing you, is a good airplane, that they would want to go in. And no matter how cool, right? It's really cool to fly in the air and you can go really far. But, if that thing is going to crash and kill you and your family, you're not going to go in that airplane. So I think we have to think about exactly the sort of risk-benefit trade-offs and what risk price we're willing to pay as we race to build ever more powerful models. This may be a bit bold to say now with recent information, but I understand that Finland has a plan to integrate AI in public services almost entirely in the next five years. Is that right? And in the meantime, we get this sort of warning shot that we got right with this AI under

testing, breaking out of its sandbox and hacking systems. And I've seen that policymakers have reacted and they proposed a, is it a kill switch bill? Right? I'm really glad that policymakers are reacting. Kill switch bill is probably something we need as a last resort. But if you imagine that all of Finland's public services are now running on AI, and then it turns out that that thing was not safe, and the companies that built it can't control it, do you really want to pull the kill switch then and turn off the public services? Or should we think a little bit proactively if we can slow down the race beforehand and build safety into the systems before we deploy them everywhere?

Arno:
That's a good point.

Eric:
Yeah. Fair question. Okay, so we've started to see some of these risks to materialize. And there is a race, I think that's a very fair description. If you reflect back on your own career, do you have any regrets maybe contributing to the early stages of this race?

Katarina:
I have several opinions on this. On the one hand, it's a bit ironic that a project that was there for safety, because we really built the InstructGPT models as an alignment project with a safety focus, it's ironic that that was part of starting the race. Part of me wants to say that if I would have done things differently, it wouldn't have made a difference. But part of me also disagrees with that. Because you should never abdicate your responsibility, no matter who you are. So that's a tricky question. What I wanted to say before, when you asked about what it felt like at OpenAI when things were really intense, so not to that specific question, but I would say that especially on the safety teams, I think there was an ambient sense of looming responsibility. I went to a social event once after I had left OpenAI and there were safety researchers there from Anthropic and OpenAI. You know, this philosophical parable about, is it, a tree falls in the forest, but there's nobody there to hear it. Does it still make a sound? So I made a joke or a sort of a pun based on this. I said like, suppose you do something that you are responsible for, but there is no human left to blame you. Are you still blameworthy? And all these safety researchers, they turn and look straight at me and they say, "yes, you are".

Arno:
No one was laughing.

Katarina:
No one was laughing. And they told me I shouldn't be telling such morbid jokes. And maybe I shouldn't be telling such morbid jokes to you either. But yeah, the looming sense of responsibility is always there, and making the right decisions is really hard.

Eric:
Well, switching the topic again a bit. We've spoken a lot about what the potential negative impacts of AI on humanity are and what the risks associated there are. Now recently, some researchers have started to ask the opposite question of what the risks of humans on sort of AI welfare are. Maybe most notably Anthropic publishes these

model cards for every model they release. There's typically a section called model welfare assessment where they, for example, in the recent one, they state, "We remain uncertain about Claude's moral status. However, we believe there's a realistic possibility that current or future models merit some degree of moral consideration." What do you make of these claims?

Katarina:
You know, I respect the people who are doing this work. I think they have good arguments that we can go severely wrong in either direction. Because I don't know if they merit moral concern and probably neither do you. If I were a policymaker, this is not an area I would invest resources in, given the other problems that we have and the risks that AI is creating for humans and for human society.

Eric:
Yeah. That's fair. I'm obviously not an expert either. I guess if we had the author of this sitting here, what I would ask or what in my mind makes a difference comparing, let's say, humans and these models, is that let's suppose the model was experiencing some harms when you ask it to do something very evil or something nasty, but you can always sort of roll back to the same state where the model was before you asked this question. So what's the entity of whose welfare you care about? Because with humans, you cannot roll back. To me that makes a difference.

Katarina:
I think that's a really good point that relates not only to model welfare, the question of what is one unit model. It's much clearer for humans. And that, I think, even plays into some of these loss-of-control risk scenarios. Because suppose that a model develops a drive for self-preservation and it wants to avoid being shut down. What is the unit model? On whose behalf is it going to take these actions? And I know there are people doing research as well on that. That's really fascinating.

Arno:
Let me ask you about consciousness. So are these big models already conscious or do you think they will be one day?

Katarina:
I'm not a consciousness researcher. There are people who are. I'm happy to chat about it as a layperson, if you want to have that conversation with me.

Eric:
Yes, please.

Katarina:
Anecdotally, because we were telling stories, actually before I joined OpenAI, at one point, I had a consciousness coffee with Alex Krizhevsky, the first author of the ImageNet paper that kicked off the deep learning revolution. So back in those days, I gave these topics a little bit of thought. I think a bad argument for the models are conscious is that they feel conscious. And I think that's a very tempting trap to walk into,

as the models are made to be more and more anthropomorphic, right? They can speak like humans with human prosody. They'll let you interrupt them and they'll have emotional intonation. And some people form relationships of various kinds with their AIs. So it must feel viscerally to many people, I imagine, like they are conscious the same way that I assume, viscerally, that you two are conscious. Right? But that, importantly, is an illusion. There was a paper by Meredith Ringel Morris that made this very stark for me. It wasn't actually the purpose of the paper. It was an AI-human interaction paper. They were talking about how you can control image generation diffusion models. The thing is that if as an artist, you want to steer an image generation model, you can't just pass it halfway through generation. Because that output doesn't look at all the way that a human artists' painting would look like. So what they proposed is that let's train a surrogate model that makes the intermediate output so that the human can give feedback on it. To me, that makes it really stark that the way that a model paints is not at all the way that you paint. It's just that the final output is made to look as if it's made by Van Gogh or somebody else. So they are not conscious because they appear to be like you. That's sort of a last stage illusion. I've also heard the argument that they cannot be conscious. And I think, Eric, you mentioned this to me before that some people make the argument that they're just matrix operations. And actually Alex Krizhevsky made a similar argument. He said, "I can write out a neural network on a whiteboard". "Are you telling me this whiteboard is conscious?" I think that's a stronger argument. And, you know, I wish we had Alex here to argue with us. But my thought is they're not just matrix operations. Right? They're matrix operations and nonlinearities to begin with. And those ones were inspired by computations that are made in the human brain. And they seem to be pretty good abstractions, don't you think? They can classify objects. They can read, speak and do lots of things that brains can do. So in some sense, in principle, why wouldn't they be able to also do the thing that brains do? At least as a neuroscientist, I think consciousness is implemented in the brain. So why shouldn't they also be able to do the thing that brains do, which is to become conscious, if you will? I'm curious for your thoughts on this.

Arno:
Yeah. Basically, the more I think about it, and now that I listen to you, the more I start thinking that maybe we aren't conscious.

Eric:
That's one possible conclusion. I mean, I'm thinking back to the example, that you can rewind these back to a previous state. Okay, I don't know how that goes together with the definitions of consciousness, but that makes this model fundamentally very different from the human brain. Maybe there is some connection to consciousness there as well.

Katarina:
Yeah, I agree with that, that it certainly makes it counterintuitive to picture it being conscious the same way that we are conscious.

Eric:
It's not like continuously processing some information the same way human brain is.

Katarina:
Yeah.

Eric:
Anyway.

Arno:
I think it boils down to definitions. So how to define consciousness.

Katarina:
For sure. There was a recent paper from Anthropic on something they called the J-space that got blown up in the media saying that now it looks like Anthropic's models are conscious. But actually, the authors were very clear even in the intro of the paper that what they're talking about is a specific subtype of consciousness, to your point about definitions, that's called access consciousness. That's a little bit similar to an internal monologue, if you will. And they weren't making claims at all about phenomenal consciousness or qualia, that what it is like to be a human, what it is like to experience red. They didn't make claims about that at all. And yet because they sort of gestured towards the term consciousness in their paper, now in layman's terms, everybody is debating if maybe the model is conscious. And then, maybe at worst, people draw links from that to model welfare considerations, which I feel is definitely premature based on that data. What do you think?

Eric:
No, I mean, you're more of an expert here with your neuroscience background.

Katarina:
But I'm definitely not a consciousness researcher. And I recommend folks read some books by actual consciousness researchers as well.

Eric:
That's fair.

Arno:
We just know about linear algebra.

Arno:
We have been asking this from the other guests as well. But maybe this is a very relevant question to ask you. So, what is your P(doom)?

Katarina:
Oh, can you define doom for me? Is doom that all humans die?

Arno:
Yeah. Unfortunately.

Katarina:
So I'm not a forecaster. This is one of these cases...

Eric:
You have to give us a number.

Katarina:
I would prefer to defer to my colleagues. So in that case I'm going to give you a number based on authority. Because I talked to a senior safety researcher at GDM who has thought about this a lot more than me. And he said 10% or less. Actually, I think my own P(doom) is probably even substantially less than that. Especially if we say that all humans die. I think that's really, really unlikely. But I think actually there are two things to notice about it. If I refer back to this sort of authority's take. Suppose you say it's about 10%. One thing to notice about that is that 90% probability is that it doesn't happen. And then we do have to concern ourselves with all the other AI related problems that do with the highest possible probability happen. But also, none of us would go on an airplane with 10% crash probability. So given that the outcome is absolutely disastrous, even if the probability is small, it's still too high. What's your P(doom)?

Arno:
You're not supposed to ask questions from us.

Eric:
Let's leave that for a future episode.

Arno:
I still have to think about it. Yeah.

Eric:
Yeah, we have covered a broad range of topics. Anything else you'd like to cover before we wrap up?

Katarina:
I'm interested in how you two see Finland's role in the AI revolution and in AI safety.

Eric:
I mean, I think there's a broader discussion on the Europe's role. There's a nice essay called Europe 2031 that lays out one scenario. I don't know, I think it's on a lot of people's minds given that there's more and more sort of geopolitical tension around this. Like they become so useful that actually you can't, or maybe you can, but people are not willing to share anymore these with everyone. I think part of this podcast's mission is to show that Finland has long roots in what has developed into the current models that we have. So there's a long history that people don't necessarily realize. But I think it's probably both of our mission to make sure that it's not just in the history, but we also play an important role in where this technology is going in the future. So I don't know. It's a race, as you said, but I think we have an important role to play there. Yeah, I

think, of course Finland has long traditions and history in AI. But people ask what is the relevance of people doing these models in the 1970s and 80s. Of course, it's important for us to know what has led to the current AI models, because the roots go quite deep. But then at the same time also we need to be relevant today and in the future. I think Europe has a lot to give. Finland has a lot to give. And people are forgetting this quite often, I feel.

Eric:
And maybe about AI safety specifically, one observation we discussed before this podcast is I think there's a lot of interest among students. I've observed that a lot of people are really passionate about that. So I think also as a university we should answer to that demand and be doing more in terms of teaching these topics. So hopefully this podcast can also serve as a part of that purpose.

Arno:
Related to that, I've been visiting schools, quite young children, and high schools' older children, and typically the children are quite well versed in AI these days. Often actually I get better questions from the children in the schools than from adults or students on courses. I think they have this curiosity and open mind to ask relevant questions. And many of their questions are about risks and concerns. Some students in schools are asking about electricity consumption and natural resources that the data centers use. Some are asking about that truthfulness of the answers. The teachers of course ask about de-skilling and other things. But what would you tell the youngsters?

Katarina:
What would I tell the youngsters?

Arno:
Related to risks?

Katarina:
It sounds like they're asking all the right questions. I love it. I'm really impressed with these kids that you met. Yeah, I would encourage them to keep asking questions and to keep educating themselves about these topics. For English speakers, actually specifically, there's a great newsletter that's called The Transformer if you sign up for that. But it sounds like they already have great information sources. And I think I agree with you that Finland has really great potential both in AI broadly and in AI safety. And we should really use that at this moment in time.

Eric:
Okay. Thanks a lot. Great pleasure having you.

Katarina:
Such a pleasure to be here. Thank you so much for having me.