AI and You

Transcript

Guest: Risto Uuk, part 2

Episode 140

First Aired: Monday, February 20, 2023

Hello, and welcome to episode 140! This week, we'll conclude the interview with Risto Uuk, a Policy Researcher at the Future of Life Institute working from Brussels in Europe, where he is focused primarily on researching policy-making on AI to maximize the societal benefits of increasingly powerful AI systems. He is an expert on the EU's AI Act and previously worked for the World Economic Forum on a project about positive AI economic futures, did research for the European Commission on trustworthy AI, and provided research support at Berkeley Existential Risk Initiative on European AI policy.

In part 1 Risto told us how he got into this line of work, and helped us understand the basic form of the act, what sort of things it regulates, its definitions of risks, and so on. Today we'll get more detail about how it might affect companies, their products, and how that could show up. Let's get to the interview with Risto Uuk.

I wanted to say that I think for many people, their most recent experience of EU regulation around computers and communication would be the GDPR, which for most of us shows up as an annoying box that pops up when we visit a website that we make go away as quickly as possible by giving them what they wanted in the first place. And so then let's look at the application of this act. What would we see if it's enforced, implemented in its current form? How would it show up? More boxes to click on? What would we see?

Yeah. Definitely, for a lot of people, the privacy regulations seem like just ticking a box. And for many, the AI Act might also run the risk of you just having to tick some boxes and it actually doesn't make AI systems safer. So we want to avoid that through these high-level principles or high-level framework in this AI act. But also through the AI standards parties actually trying to think about guidance for companies to comply with regulation and what they would actually do in practice to comply with robustness or accuracy or cybersecurity requirements and these kinds of things. So as I said earlier, I think I mentioned in the AI Act, there's the chapter two, which lists all the requirements for high-risk AI use cases from technical documentation to risk management, all these kinds of things. So AI companies would have to comply with those if they want to provide high-risk AI systems essentially. Some of those might seem very much like ticking the box. Maybe one might wonder when there's technical documentation, what does this do? We just document information around AI systems, but there's already these practices in the field. This is not something that is completely made up by regulators. For example, many companies are currently using model cards, and they are going to start using these more. And this is being popularized by industry actors. Model cards, system cards, data sheets, all kinds of things to document what is the intended purpose of an AI system, what kind of things it's not meant to be used for, and all kinds of details around it, which could be useful for companies themselves to understand why they're developing this AI system, what some of the risks might be. But also for the users of these systems, they would know that these systems shouldn't be used for this or that. Like in the case of ChatGPT that you mentioned, maybe if there's a good model card for users highlighting that these systems oftentimes hallucinate, they provide misinformation, they make up stuff, and then highlighting or emphasizing that if you are working in specific fields and using the AI system for specific purposes, it's actually not meant for, I don't know, constructing legal text or whatever, or it's not meant to write an entire op-ed without human oversight or some kinds of things. This could be helpful for users. And then if the users decide to ignore this then that would be to some extent their fault if that information was reasonably available for them, I guess.

So it sounds like the impact on a company would be something like ISO 9000 compliance effort where you have to document various chains of your process to a certain standard. Is that a good analogy?

Yeah, this might be a good analogy. Maybe just one additional thing to mention to this is kind of, if indeed there are good enforcement mechanisms and institutions set up and good enforcement ability, which seems to be the case with, for example, GDPR, that now over time, privacy enforcement or enforcement of this regulation tends to get better as understanding improves and institutions and capacities increased, more and more penalties are given out for violation of GDPR. And it even turns out that maybe some businesses are kind of entirely not set up from the perspective of privacy regulation like Meta, for example, seems to be in all kinds of trouble recently with that. And in the case of the AI Act, similar things could be the case that in the original proposal, there are some penalties put forward for non-compliance. One is the prohibited practices that I mentioned. The second is obligations of high-risk systems. And third is also the infringement of the duty to cooperate with competent national authorities. And then you get various kinds of fees or penalties based on what you violate and depending on the size of the company, like how big your total worldwide annual turnover is and things like that. So yeah, that's an essential aspect you mentioned here with regard to what companies might face in case they don't comply with this regulation.

Right. Now I want to talk about the risk, because this is obviously pivotal to the act, and two types of risk occur to me. One is the risk that the product won't do what it's supposed to. And the other is that there's an unintended side effect of it doing exactly what it's supposed to. So for instance an AI that is vetting resumes, CVs of people could have inherent bias and rejects female applicants, which happened at Amazon, and so it's not working the way it's supposed to. Another would be social media where things like the TikTok algorithm were doing exactly what they were supposed to but the unintended side effect was that they made people angry and radicalized them and created factions of conspiracy theorists. Does the act distinguish these types of risk? Does it focus on one more than the other? Does it focus on either or both?

So that, at least from the perspective of the commission when they released this, was kind of intending to achieve several types of goals with this. One is to reduce safety risks, the very concrete risks that we talked about, but also protect against fundamental rights but also have all kinds of benefits like reduce legal uncertainty for companies, reduce mistrust in the use of AI, reduce fragmentation of the economic market and things like that. But when it comes to kind of seeing whether the AI Act makes distinctions between side effects and things like that, I don't think it really does. Like when I mentioned, for example, the risk of societal harm, then a lot of people's reaction or policy-maker's reaction, or at least some of policy-maker's reaction to that was, it's very hard to monitor or measure what's societal harm. It's very hard to identify these harms and these risks. And because of that, all of the focus, oftentimes these are more clear risks. But in the case of the CV scanning tool that you mentioned, in that case, people are not thinking about safety or health-related risks, they are thinking more about kind of discrimination-related risks and these kinds of things. So in that case, they are making this distinction, and maybe in this case, it would be useful to highlight what is under Annex III. And maybe in that case, we would basically understand what kind of risks the European Commission had in mind when they released this. So, for example, there's one high-risk category, it's biometric identification and categorization of natural persons. So these types of AI systems, they don't appear to have health-related, like very direct health harm related risks, they seem to be more around some type of other fundamental rights. Then another category is management and operation of critical infrastructure. This has very clear health-related, like physical health and safety-related implications. Then you can go to a third category, which is education and vocational training. Again, in these cases, for example, AI systems intended to be used for the purpose of determining access or assigning natural persons to different educational positions or institutions. This doesn't seem to have any specific safety-related issues in mind, but some other fundamental rights. And this list has eight categories, and we can go through them, and you can kind of understand what types of harms EU institutions have in mind when they discuss the AI Act. But there's lots of civil society members who are throwing other types of things here, "Hey, you really haven't considered this risk." For example, you don't find on this list environmental harm. It's nowhere to be found. Although in various negotiations, environmental harm has been brought up but I'm talking about the initial proposal by the European Commission.

Perhaps you could illustrate for us what sort of effects this would have on some examples of different companies. If you could think of some specific companies and how they might be affected ranging from large well-known ones like Meta to some possibly hypothetical, but smaller businesses. And in particular, which companies might find this no big deal, which might be inconvenienced, which might find it posing a risk to their livelihood?

Yeah. This is a very difficult question because it's hard to think about exactly what kind of companies and what kind of businesses they have. I mean, we can go into that, but it's easier to kind of just generally think that those companies will be affected by this, who provide AI systems for those purposes that are on the high-risk list in Annex III that I mentioned, which are these eight categories. I went through three ones, but I can also mention the other ones are employment, workers management and self-employment. That's the kind of labor-related category.

Or maybe we can look at some specific examples. For instance, there are lots of negative stories emerging about, you mentioned labor management AI systems in say Amazon and Starbucks and Uber for instance, have had complaints that their labor management AI creates unfair and punitive working conditions. Is that one of the things that's a risk in that category?

I would hear mention regarding Uber and these kinds of examples. There's another legislative file in the European Union, I think it's a platform economy directive or something like that, that's actually meant to specialize on this particular topic. So that will be probably more relevant to this issue. But in terms of employment, currently, or at least in the initial proposal, there are two kinds of AI systems that were tried to be captured by this. One is AI systems that are used for recruitment like advertising vacancies, screening people, filtering applications, these kinds of things. Like AI systems or AI companies that are providing these kinds of AI products will be captured by this regulation. And the second one was AI systems that are intended to be used for making decisions on whether to promote somebody, whether to terminate their work, how to allocate tasks, and evaluate their performance. All of these seem somewhat relevant to, I don't know, Amazon's examples or these kinds of companies. But yeah, I mean, those are the two ones that are especially mentioned regarding employment-related AI systems.

These cover such broad areas that there's such a breadth of applications here, it seems impossible for one act to get down to the nitty-gritty of all of them. Like if we're talking about the hiring, there's a company in the US called HireVue, which conducts AI-mediated interviews, actually AI-directed job interviews, and has been criticized for having emotion assessment as part of that. And then there are so many other areas that we can get into here. Is the act going to drill down as far as necessary in those eight areas that you mentioned, or is it leaving some of those decisions to other regulation or bodies?

Yeah, I think there's a tricky balance to be had here. On the one hand, if you just have a very general category, then companies will of course be very upset with that because it doesn't provide enough legal certainty to them. Like, somebody could say, "I don't know if my AI system exactly goes into this category." They might kind of try to find ways how it could be in another category. I guess in the case where you are extremely precise and say very concretely what will be captured by this, then maybe some companies would really try to find ways to modify the system in some way so that they wouldn't be captured by this, like slightly differently, a slightly different version of that same thing, and it couldn't be captured. So there's a very tricky and very challenging issue here, how to find exactly how precisely or how generally to word it. And there's probably going to be all kinds of issues where this is going to be decided on a case-by-case basis. Some incidents are kind of edge cases and not entirely sure whether it is captured or reasonably. Can you interpret reasonably this particular paragraph in this article? And this is the kind of law context just generally, that's why there are lawyers, there are courts, all these kinds of things need to be done.

And I've been doing all along here, the engineers' kind of thing, of finding the edge cases, looking for the ways in which this could break or not work. And I want to look at the opposite now, and imagine that it is successful according to the intentions of the people working on it and see what difference does that make to the world. Maybe it doesn't result in hordes of people writing thank you letters, but maybe there are things that don't happen to them as a result. So if you imagine these two alternate futures, one where the act is as successful you can hope and the other where that didn't happen. How do you contrast those two futures? What is the positive difference the act has made?

I think this is to a large extent speculative because we're trying to come up with--

Oh, it's completely speculative. That's what I want you to do.

Yes. Trying to come up with requirements and then trying to imagine kind of with those requirements, how the world looks like versus without those requirements. And I think what would be useful is to think about and look at recent AI-related incidents that have occurred that are especially visible to people and are in their mind and these kinds of AI incidents that people care about. And maybe there are many ones to choose from, but maybe one that comes to mind especially is the recent Dutch scandal where in the Netherlands between 2005 and 2019, authorities wrongly accused an estimated 26,000 parents of making fraudulent benefit claims. So between those years, basically some payments were made that were asked to be paid back by these parents in their entirety. And basically, it meant sometimes tens of thousands of euros had to be paid back to the government. And some people committed suicide, some people were led to poverty because of that, more than a thousand children were taken away. And I don't know exactly what kind of AI system was used by the government to make this decision but you can imagine that the EU AI Act would reduce the chance of this type of incident because one of the high-risk use cases that we talked about is using AI for determining access to essential private services and public services. And yeah, hopefully, people hope that having these requirements, like for example, if you have documentation requirements and risk management requirements, these kinds of things, then people would monitor AI systems more closely, and they would ideally find out about AI systems being used to make these kinds of decisions. And how sometimes these decisions are made quite randomly, and there's not a lot of really good justification behind it. And then you ask some kind of welfare payments back from people that actually was completely incorrect to do. So that could be one type of incident. Maybe in an ideal world if the act is well enforced, if the high-level requirements are implemented in practice precisely enough and clearly enough, then hopefully that would help.

I'm reminded of something on a lesser scale, but it reminded me of a case in the United Kingdom where many people, I think hundreds of people, who ran post offices were accused of skimming cash and had never done that, but they were essentially convicted without evidence other than a system which was wrong. If I look now at the pressures on this act, in the US, anything like this would be subject to immense lobbying, and I assume that that's the case there as well. What are the external pressures on this right now to change it?

So, I agree. There's a lot of lobbying happening in the European Union as well as in the US. I don't remember exactly where I saw this data, but there was some recent data highlighting that some of the Big Tech players in the US have increased their lobbying efforts in the EU recently, because the EU has started focusing on regulating AI and digital services and products. For example, the AI Act is not the only regulation that the EU is working on, there's a Digital Services Act, Digital Markets Act, Data Governance Act, Data Act. There's all kinds of things that are very relevant to these businesses. So apparently, yeah, many companies have doubled their lobbying efforts from several million to even more. So there’s definitely a lot of influence coming from Big Tech, from other industry actors, but also from civil society and everywhere.

Does the implementation of this, for instance, require companies to open up for inspection parts of their systems that they consider proprietary and would not want third parties looking at that data or their algorithms?

Yes. I haven’t paid very close attention to exactly what these debates are, but I’ve definitely heard about these types of debate around intellectual property. I’ve seen some policy-makers draft requirements where they put a qualifier without violating intellectual property rights and these kinds of qualifiers to avoid some issues or please some actors that are very worried about the requirements. But these kinds of debates are constantly being had, where on the one hand, it seems that there’s some need to have access to data sets, to have access to algorithms and whatnot, to understand better what kind of issues the AI system might have caused. But on the other hand, some of these are in the business’ interest not to share. So yeah, this is definitely something that is being discussed. I don’t have very specific examples in mind immediately.

If we’re looking now to get into our final questions and you were speaking at this point to people about to enter or graduate higher education who are looking at career choices, what kind of people does this effort need the most, and what sort of difference can they make, and where can they go to do that?

Yeah, that’s a very tough question because my own view is that this field can benefit from a wide range of academic backgrounds and interests. It’s very easy for people to say that, “Oh, when we’re talking about AI, that means we need computer science and machine learning background,” which is fair enough. We definitely need that. But we also need lawyers, we need economists, we need philosophers, we need various kind of social scientists, people with engineering backgrounds. I think all of those can contribute in some way. Lawyers can interpret and craft text legally, in a very precise way perhaps, but on the other hand, they maybe don’t focus on really big-picture issues in society. Maybe people who have some type of sociology background to study what the public opinion is, maybe they can contribute in some other way because they can see some other perspective that lawyers wouldn’t be able to. And of course, computer scientists, for example, when I think about the AI Act, some key questions that are there, for example, there’s one article called Article 15, which is about cybersecurity robustness and accuracy requirements. That is a very technical article. In that case, computer scientists and machine learning researchers can definitely contribute a ton there. But when it comes to, say, the enforcement of this regulation and what kind of institutions to set up, this is not really a very big machine learning question. So in that case, you would benefit from getting input from political scientists and some other fields. So there’s really a wide range of topics that people can contribute to in this particular regulation. But of course, this regulation will be passed perhaps by the end of this year, or if it’s delayed, maybe later. But I don’t think that this regulation will be the final say of AI regulation because so far, for example, in 2018 we had the GDPR privacy regulation, and now we have seen more and more regulatory initiatives come out in the European Union, but also elsewhere. And I think as we are adopting AI systems in the market more and more, and we are seeing that, hey, we actually missed some things with regulation, and also maybe through negotiations, this regulation gets watered down in some way, or it finds some kind of compromise between policy-makers that is suboptimal from the perspective of the next AI systems we have on the market. So I think there are going to be new ones and new risks that we might see that the current one doesn't address, or isn't very easily adaptable to those risks. And then you will have new regulatory initiatives, and then you will have, again, lots more people who need to contribute.

Well, it is exciting that one act has such need of so many different disciplines. Do you get to have those kinds of interdisciplinary interactions? Do you work with lawyers, economists, philosophers, computer scientists as part of all this? How does that play into your career goals and how you feel about your work?

Absolutely. So I have a formal background in public policy, economics, philosophy. Those are kind of the areas that I've learned the most about in my education and now self-taught basically in kind of AI-related machine learning-related issues. And I read other people's papers and books and whatnot, but I don't have a strong formal background in that kind of field. So that very easily leads me to developing a network of technical research that I can always reach out to and ask for advice. For example, we recently published a paper trying to define general-purpose systems for regulatory purposes. Like what are these systems exactly and what kind of systems would be in scope and wouldn't be based on some definition? And for that, it has many technical questions. For that, I reached out to people in the US, I have collaborators at Berkeley, at Stanford, in other places, in the UK at Oxford University. Very technical people that I ask for advice from. We also collaborate with people who are very strong, or I personally collaborate with lots of people in, say, civil society who have a very strong background in social sciences and humanities and all kinds of different areas. And we just generally have quite a wide range of researchers and expertise in our Future of Life Institute network. We have in the past, for example, created surveys publicly where we have asked researchers and other specialists to fill out the survey telling us what they think about various kinds of AI-related risks and issues and challenges and whatnot. So we have scattered a lot of people through those efforts and we oftentimes reach out to them depending on a specific project that we are working on and what our specific need is. Right now, for example, I'm working on another paper where we're trying to think about how to offer more guidance to companies and regulators in terms of defining some AI systems. And there, we are directly collaborating with machine learning researchers as co-authors and significant contributors.

Well, thank you. And I don't think I've ever heard regulation sound so exciting, thanks to you. We've covered a lot of territory here. Is there any large rock we should have turned over, but haven't? Anything that you want to tell our audience about the EU AI Act that they should be aware of, look out for, or that we otherwise haven't done justice to?

Yeah, I'll just mention that we talked about various kinds of very technical and very difficult issues, and yeah, in this kind of format, I'm only able to share very brief thoughts and might miss very important details. So if people are actually very interested in more closely understanding this, then we have set up a website, artificialintelligenceact.eu That website specializes in providing information about the AI Act, all kinds of legal documents that people can read if they so want and prefer, pages as well. And it currently has more than 80,000 unique visitors, so it's been followed by a lot of people. And then we also have a biweekly newsletter, the EU AI Act newsletter, which is on Substack, and that one has nearly 5,000 subscribers. So a lot of people pay attention to that as well. It comes out roughly every two weeks covering what is happening in the legislative process, what are EU regulators doing around the AI Act, but also some analysis from civil society, from industry, from academics analyzing various kinds of proposals and issues and challenges with the AI Act. So I think these two resources in case people want to learn much more about it or follow it much closer, then I think these resources can help them. So that's one thing because we were only able to capture so much about this.

Of course, it almost is insulting to try and boil this down to an hour but we've got to start somewhere. Do you have any personal page or contact information for people who want to follow what you are doing?

Yeah, you can pay attention to my stuff on Twitter. You can find it with my name. Risto Uuk is my name, and you can just find me on Twitter or LinkedIn at the same name or check out the Future of Life website as well, futureoflife.org. I think these are the best ways to reach us.

Oh, fascinating. Well, thank you very much, Risto for coming on *AI and You*. It's been a fascinating discussion. Really enjoyed it.

Thank you so much for the great questions. Thanks a lot.

That's the end of the interview. Of course, learning about how AI is now called out explicitly for this kind of regulatory attention only illustrates how much momentum it is gathering. I don't think we can call this one a fad.

In today's news ripped from the headlines about AI, actually in June, a cargo ship completed a voyage across the Pacific Ocean using autonomous navigation. The liquid natural gas carrier *Prism Courage*, went from the Gulf of Mexico, through the Panama Canal, and across the ocean to South Korea using a system developed by the Hyundai subsidiary Avikus. The journey took 33 days and Avikus claimed it improved fuel efficiency by 7% and reduced emissions by 5%. No, it wasn't unmanned, there were people on board and they did take control for about half the trip, but still, the ship used AI to assess weather, wave heights and nearby shipping to figure out the best route while adjusting the ships steering in real-time, and recognize nearby boats, avoiding a collision around 100 times by itself.

Next week, we're going to do a special solo episode, no guests, just me helping you make sense of ChatGPT and the disruption it's wreaking upon our world. Yes, in one sense this is bandwagon jumping, but in another, we were on this bandwagon long before ChatGPT came along. And the publicity about ChatGPT is certainly not dying down. So the angle I'm going to take will be the broader context of disruptive effects of AI, illustrated through the kinds of changes we're seeing from ChatGPT. Which I can't help but point out, I've been saying on this show and in my books and elsewhere for several years now were imminent. So that's next week, on *AI and You.*

Until then, remember: no matter how much computers learn how to do, it’s how we come together as *humans* that matters.

http://aiandyou.net