

Our Human Future with AI

Episode 3: Regulate and Build: The EU's Past, Present and Future in and with AI

Sue Anne Teo and Fredrik Heintz

in conversation with Thérèse Murphy

Thérèse Murphy (TM): Hello, and welcome. This is a podcast series by the Global Campus of Human Rights. It's all about artificial intelligence, specifically it's about our human future with artificial intelligence. I'm your host. My name is Thérèse Murphy. In this episode, I'm asking: what is the AI Act developed by the European Union? What's good about that Act, and what could be improved.

To talk about these questions, I'm joined today by two expert guests: first Dr. Sue Anne Teo, who's the expert on AI and Human Rights at the Raoul Wallenberg Institute of Human Rights and Humanitarian Law, based in Lund, in Sweden. I'm also joined by another expert who happens to be based in Sweden. It's Dr. Frederick Heintz, who's a professor of computer science at Linköping University. Fredrik has been an AI advisor to both the Swedish government and the EU. Fredrick, Sue Anne, you're both very welcome. Thanks so much for joining us.

Fredrik Heintz (FH): Thank you.

Sue Anne Teo (SAT): Thank you.

TM: Shall we get the conversation started? Sue Anne, my first question is for you. I know you have legal training. The EU's AI Act is a piece of law, a piece of legislation. So, could you give us a very basic introduction to the EU's AI Act, and as part of that, can you help me understand why this act seems to have got so much attention? Over to you, Sue Anne.

SAT: Thank you, Thérèse, for asking this question, because I think it sets the scene in terms of looking at AI regulation in general, but also bringing forth the European approach. So, the EU AI Act is a milestone, in a sense, because it's the first in the world to comprehensively regulate AI. It treats AI as a distinct object of regulation using this thing called a risk-based approach, and which treats AI as a product from a product safety legislation perspective. So, the risk-based approach means that the type of obligations that are set out for those who design and deploy AI depends very much on a risk classification.

So, regulation is based upon how much risk a given AI system presents to society, to health, safety, and fundamental rights. So, if your use case is deemed to be high risk, then there are more obligations demanded from you by the law compared to if a use case involving AI is considered minimal risk. So, if a use case falls within the high-risk category, there are strong obligations around risk management measures requiring, for example, logging, data management, addressing bias measures, information sharing, and in

some cases even a requirement for you to do a human rights or fundamental rights impact assessment. Examples of high-risk AI are in the use of, for example, AI in critical infrastructure, so your electricity grids, and so on, the use of AI by public administration, in education, border and migration, law enforcement, and other areas. There are a total of eight of these high-risk AI use cases. There's also actually a small group of AI use cases that are deemed to be so damaging to health, safety, and fundamental rights, including towards human dignity, that they are actually prohibited under the EU AI Act. These include AI systems that manipulate people as well as those that exploit vulnerabilities of groups because of their disability, age, and socio-economic status, for example. In turn, some other use cases are considered minimal risk, which means that people, companies deploying and providing these AI systems have less obligations, let's say oriented mainly around transparency. General purpose AI, for example, the, you know, Large Language Models that we're familiar with, are also regulated and are required to share information, including on copyright, but if your general purpose model falls under a what is called a systemic risk category, where there is a high risk of severe harms that can propagate across society, then you have more obligations around testing and evaluating your AI model, for example.

So, the EU AI Act, in general, approaches AI as a product, but within this framing, through the differentiated risk classification and the obligations that are required, fundamental rights or human rights considerations certainly do shape and feature in the in the legislation as well. And the Act has also gained a lot of attention because we can trace its distinct way of regulating compared to other jurisdictions, let's say. So, for example, China, which, unlike the EU, does not regulate AI in this very comprehensive manner, but according to specific use cases or domains, one of them being let's say the algorithmic recommendation systems that we're familiar with, for example, in, you know, social media algorithms, or in human-like AI, AI, you know, that is showing sort of human-like qualities, such as emotion, emulation and talking as like a human, for example. So, it can also be contrasted with the approach in the US, which has a more, let's say, market-driven approach, so letting the market develop and to a large extent sort of regulate itself in the understanding that, rightly or wrongly, that this is the best approach that promotes innovation. At the same time, even though there's no, let's say, US law, US federal law regulating AI, states do regulate AI in different ways, so the US states, different states themselves focusing on, for example, companionship AI, such as in New York and California, or on, you know, regulating frontier models, which are these highly advanced AI models. So, the EU AI Act stands out a little bit in this regard, because it's the first mover, and because it's a risk-based approach within this comprehensive frame of regulation, which is for now a distinct enough way to regulate AI compared to other jurisdictions. I guess we will talk about the pros and cons in a bit. So, I will leave this for now.

TM: Absolutely, Sue Anne, the plan is to talk about pros and cons a good deal more during this episode. So that's very interesting. Thanks for the short introduction. You summarized for us that how distinctive the approach of the EU is, comparing it with the approach taken in China and the United States, for example. You made it clear to us that the EU was a first mover, you know, describing this as like a world first form of regulation that it treats AI as a product, and then, in very interestingly, that it takes a risk-based approach to regulating AI.

And Fredrick, if I could now use that to ask you not directly about the AI Act, but about aspects of your own current work, because I know that you're involved in work on trustworthy AI, and so they're like, there seems to be a potentially nice counterpart there, between, on the one hand, you know, talking about risky AI, and then, on the other hand, talking about trustworthy AI. So, could you tell us a little bit about your current work in the area of trustworthy AI? You know, what is it? What work are you doing? Why? What

makes AI trustworthy versus untrustworthy, and presumably risky, then, on that basis? Over to you, Frederick.

FH: Thank you very much. So, so I would say, I mean, basically I was part of this high-level expert group where we developed these ethical guidelines for trustworthy AI back in 2019, and then we basically had three requirements for an AI system to be trustworthy. The first was that they should follow the rules and regulations that apply. The second is that it should satisfy four ethical principles, including respecting human autonomy, fairness, explicability, and also avoiding harm. And then the third requirement was that you need to have a robust and safe proper implementation, and if you have all these three, meet all these three requirements, then you could argue that your system is trustworthy.

So, the project that I currently lead is called “Trust LLM”, and it's about building trustworthy and factual Large Language Models, and the reason that we need to do this is because I think we will never be able to actually properly evaluate these commercial models. They will claim that they satisfy all the regulation and all the rules, but it would be more or less impossible for the enforcement to actually verify these claims. So, therefore, I think we need to build our own alternatives that need to be built on data that is compliant within the EU, because I think this will be the major part. I mean, it's, I think, the biggest challenge when it comes to, I mean, trustworthy compliance, and so on, regarding Large Language Models, is very much about the data collection part, but it's also, of course, that we want to be open and transparent about how do we process this data, how do we train the model, what are the objectives that we're trying to maximize, and so on. So very much being open with how these models are developed and trained, which means that as an end user you should both be able to trust the process because it's open, but also you should be able to verify and check what we claim and whether that's true or not. So that is kind of the core.

So, basically, the challenge from our side is very much this: how do you actually collect large amounts of data, which is needed in order to train these Large Language Models? Because in order to provide high-quality AI models, you really need to scale, and that is quite challenging. So, if you're talking about most of the models that currently exist in Europe [they] are very small models, which means that they are not competitive with the kind of commercial models, and in order to build competitive models, you need very large amounts of data, and that's currently one of the things we are doing. We are also trying to come up with more effective ways of doing this, so that we can minimize the energy usage and the need for compute, which is another major limiting factor for us in Europe, so that in the end we actually can develop models that hopefully are useful. We are also working together with end users, so that they are testing and evaluating and making sure that it's both trustworthy within their application domains, but also that this can be helping them to solve real business cases.

TM: Thanks, Fredrik, for the outline. It sounds like fascinating and very important work, tackling a range of challenges and seeking to build technologies that support guess European values, we could say in some ways. I guess what I'd like to ask you now, Fredrik, is: Sue Anne outlined for us the EU's AI act. You're doing your work in a European context. What's your view on the EU's AI Act? Specifically, you know, are there features of the Act that you really like, and equally, are there features of the Act that you find frustrating?

FH: So, first of all, I do think it is very important to have this regulation. I think it's very good that Europe has taken a clear stance, in the sense that we say that yes, we do want AI, but we don't want any AI. We actually have requirements and expectations on this technology, so I think that's the most important part of this in my mind. Then the challenge really becomes that from a practitioner's point of view it leads to so much legal uncertainty, and when the Act was created, one of the arguments for it was that it was intended to reduce the legal uncertainty and provide a level playing field, so that everyone knows the rules. However,

this is not the case in practice, and it will, as far as I understand, not be the case until it's actually been, I mean, ruled in courts, and that will take 5-10, years, which means that we will now have a very long period of a large uncertainty, where it's very difficult to get any clear answers. So, I think this is one of the challenges.

We can take a concrete example. So, one of the things we discussed in terms of this language model training was you can take the Swedish King – the Swedish King is a public figure – which means that then we are allowed to include information about the Swedish King, but the Swedish King is also a private citizen, and according to the GDPR and so on, if you're a private citizen, you have the right to not be. Then basically, we need to ask permission to include information about this person. And now the question becomes: are you allowed to include information about the Swedish King or not? That's an open question. We also had this question of, I mean, maybe it's not much about the AI Act, it's more about... actually, I think one challenge is also the interplay between the AI Act and other regulation. For example, GDPR has this data minimization requirement that you basically should collect as little data as possible, while the AI Act says that you should have complete and accurate training data, which means that you should have as much information as possible, and basically, if some of that data is left out, you no longer have complete coverage. So, you actually are reducing the quality of the training data in order to live up to the GDPR. And so, to me, this is really the biggest challenge from a European perspective, is that these uncertainties force organizations to take a very conservative approach, and be kind of far away from the limits, because we don't know exactly where the limits are, and therefore we have to be very conservative, and this I think is not positive from a European perspective.

TM: Thank you, Fredrik. So, what I hear you say is, you know, overall you welcome the EU's AI Act, but it's already clear that a number of challenges arise, one of which often arises when law is used as a tool to regulate. One of the challenges that has arisen is that of legal uncertainty, and you know, somewhat paradoxically, the EU's AI Act was introduced, as you pointed out, to try to minimize or ideally, eliminate legal uncertainty, but legal uncertainty remains in a number of ways. And one concern, as you explained to us, is that that legal uncertainty is then creating quite a cautious, or even a conservative approach, which might be harmful for a range of reasons.

If I may, Fredrik, can I jump from that idea then to one further, for now, question for you? So, your background is computer science. My background is like Sue Anne's, in law, and one of the claims I hear made typically by non-lawyers is that regulation, and maybe particularly law-based regulation of technology, interferes with innovation. And, perhaps because my perspective is too law-focused, I find myself, you know, getting quite irritated when I hear this claim “but regulation interferes with innovation, it's holding back the progress of the world”. And so my question to you is, you know, as a computer scientist, as somebody who is working to build trustworthy Large Language Models in Europe, how do you see innovation and regulation relating to one another? Over to you.

FH: So, I don't think necessarily that regulation will, I mean, make innovation necessarily harder, so I mean, my point of view is always, I mean, I'm not against regulation, but I think it's if we only do regulation without innovation, then we end up in trouble, and I think the challenge here is that Europe has invested so much in regulation, but almost nothing in innovation. I mean, the challenges here are that current technology have a hard time actually satisfying the regulation, that it's not clear whether you actually can build the type of applications that we want, according to the regulation. And in order to check this, we need to invest a large amount in innovation, and if we're looking at the general innovation, sorry, investments in Europe compared to the US, it's very clear that most of the investment is in the US, which means that if we want to actually have a chance to provide competitive products that is in accordance with our regulation, we need

to invest at least in the same order as the US, and since this regulation, of course, raises the bar - that's the whole purpose of the regulation - it means that we will need to invest more.

So, I would say it's basically a matter of cost, that if we raise the bar, it would cost more to develop it. We are more than happy to do it. I think we are very many that can contribute to developing better solutions that currently exist that are more aligned with European values, but then there is a need to make comparable investments, and that has not happened, and therefore I think the consequence is that currently there are no truly competitive AI models being produced in Europe, and that we are basically left out to use the commercial American models, which, considering the current geopolitical situation, I think will not really be in our best interest, and that's why I'm a very strong proponent for we need to develop, innovate in Europe in order to both live up to regulation, but also to provide alternatives that hopefully the rest of the world also want to buy, and so on, but for that we need to significantly increase the investment in innovation.

TM: Thank you, Fredrik. That's a very clear, powerful message. So, you've identified how in Europe there's been significant investment in regulation, but what you've just made clear to us is, you know, there's a partner, an essential partner arena in which the EU needs to be stepping in, to be moving, and that's investing in innovation itself. So, yes, great to innovate in relation to regulation, great to invest in regulation, but there also needs to be in Europe, for a range of reasons, investment in innovation, and right now that continues to lag behind.

Sue Anne, can I jump over to you now? Because looking at the EU's AI Act and thinking about the term we've used a couple of times in this episode – European values – one of the things that often gets cited as a European value is respect for human rights. It's not exclusively a European value by any means, but it often gets cited by European leaders as a European value, respect for human rights. So, here's my question for you, Sue Anne: are human rights protected by the EU's AI Act?

SAT: Yes, I would say it, it is, and it features in different ways. So, at a high level, fundamental or human rights inform the entirety of the AI Act, for example, through its risk classification, which I talked about before. So, those AI use cases or domains deemed to pose a high risk to fundamental rights or human rights, are either prohibited or classified as high risk, and with this classification, high risk classification, there are specific measures that are required, and these also relate to human rights, right? So, these include, for example, requiring those making and deploying AI to address issues such as bias, which also addresses discrimination risks, and in terms of, for example, ensuring that enough information is recorded and made available for accountability purposes, so that people eventually have access to remedies, which is also a human right.

A more, let's say, direct way is the requirement to conduct a fundamental rights or human rights impact assessment in relation to high-risk AI systems that are deployed for public purpose or by public actors. It's not one and the same, of course, because public authorities that deploy high-risk AI – for example using AI to, let's say, determine social welfare locations, finding possible fraud, or in relation to border and migration – but private actors can also deploy AI in a public-facing way that can impact upon key interests or life opportunities – we can think of, for example, insurance and banking, where banks determine who qualifies for loans, for example. So those actors also need to do a fundamental rights impact assessment.

Now, of course, impact assessments are not new to the human rights world. It's not a new way to account for and protect human rights. We have, of course, due diligence measures that are familiar to the business and human rights space for some time now, but I would say it's a little bit different when it comes to assessing fundamental rights risk in relation to AI, because assessments done to assess whether or not

there are human rights impacts before operating, let's say, a business in a particular country, an environment in a business and human rights context, is different from like how we assess risk in relation to AI. For AI, we need to do more iterative assessment, compared to, let's say, the more sort of one-off nature of human rights assessments, impact assessments that are more familiar in the business and human rights due diligence measures. For AI, we also need to look maybe a little bit closer at proportionality considerations, looking at the severity and scale of risk, and possibly also the types of rights that we will need to look at will also be a little bit different, arguably wider, because AI is deployed in many different areas that affect human rights, but at the same time it also can be quite tricky, because it can be difficult to anticipate these risks in advance.

There are also some things that are, let's say, missing because the EU AI Act is a legislation that is informed through a product safety lens. We don't really have rights in conjunction with the Act, so this is different from, let's say, the GDPR, where we have a bunch of rights – which Fredrik has also talked about before – such as, you know, you can rectify your data, have rights to explanation, deletion, and so on. One can complain to the authorities that oversee the implementation of the AI Act, but this is not really in the nature of individual complaints, right? The authorities are there to ensure that the Act is complied with, and not per se to ensure that individual human rights are upheld in specific cases.

And I mentioned before also that the high-risk AI systems are regulated more stringently, and this is because it poses a high risk to safety, health, and fundamental rights, but one can determine, you know, self-assess for oneself whether one falls within the high risk category, so this is not determined by an external third party or a licensing body or anything like that, but it's determined by the AI provider itself. And there are several examples in the Act, such as when, for example, you use an AI system that only does a narrow procedural task or does a preparatory task, and you can assess that it doesn't have a high risk to fundamental rights. Then you can take out yourself, you can take yourself out of the high-risk categorization of the AI Act. This has been criticized by civil society as it could allow for providers to self-assess certain systems to escape from the more stringent obligations under the high-risk classification. The backup measure is that you have to register this, so there is some transparency around it as well.

And then another way that we could look at it is in terms of the risks that are deemed to be unacceptable to human rights, because, for example, it greatly impacts human dignity – the prohibitions under the AI Act. So, if AI is used for manipulative purposes, or to exploit the vulnerabilities of certain groups, as I talked about before, this is not allowed under the Act, so it's also a way to fundamentally protect human dignity and human rights, and this ensures that human dignity cannot be balanced away, for example. So, the EU AI Act protects human rights in different ways through several checkpoints, through this risk classification, through prohibitions, and through assessments. However, some other parts also fall a bit short, which, as mentioned before, unlike the GDPR, is not a rights-based instrument, it's oriented around product safety, so the exercise and enforcement of rights is not its one key focus.

TM: Thanks for the comprehensive overview, Sue Anne. So, what I heard is: yes, absolutely, there is human rights protection provided by the EU's AI Act, but also a little bit of a mixed report in that there are some things missing, there are some gaps, but then also I guess on the positive side there's an interesting tool available through human or fundamental rights impact assessments, but there will be a need to make that tool fit the specific circumstances of AI, because a human rights impact assessment is a tool that's used in a range of settings - for some of our listeners it'll be a familiar tool – and what you said is it's going to have to be made more bespoke to fit the AI context.

Okay, we're almost drawing to a close with this episode, and if I may, I've got one final question, and I'd like to get a response from each of you in turn to this final question. As you know, with this podcast series, our particular interest has been in our future, our human future with AI. So, adopting that future-oriented perspective, can I ask first you, Fredrik, and then you, Sue Anne: if you look to the future and you have a list of desirables - let's make it simpler - you have a wish list. Thinking about AI in the world, but more specifically the EU's needs, the EU's role, and more specifically, still the EU's AI Act, what would you recommend to us? Yeah, so looking to the future, what would you recommend to achieve what you see as desirable goals? Fredrik, over to you first.

FH: I mean, my long-term vision – or well, I hope it does not, the long term, because then we will be too late – but my short-term vision is that Europe actually start producing competitive AI solutions, because I think the best way to actually live up to the regulation is to build it yourself. But in order to do that we need to significantly increase the legal certainty, and in order to achieve this legal certainty, I think we have to work more iteratively, so I think one challenge with regulation is that it's basically... it's a law, so it applies while in reality I think this will require a lot of technical development, and probably legal development as well, because we haven't even started talking about enforcement, so I have no clue how this is going to be reinforced in reality, and therefore I think we should have a much more iterative approach where we successively, iteratively improve and come closer and closer to the ideal with which I see the AI Act is much more an ideal that we wish to strive for, and something that we need to make significant technical improvements and develop new solutions in order to live up to this. And therefore I think we need to also work in tandem with the standardization, because the idea was that the standardization should provide the best practices, and that as long as you follow these best practices, then you should be within the legal realm of the regulation, but since these standards are not done yet, we don't know what is required in order to be considered to be compliant, and therefore I think before that is known, we should not introduce the regulation. So, we should strive to always improve, but I think it will be very difficult to do all in one step.

TM: Thanks, Fredrik. So, you know what I hear is some very clear recommendations resonating with what you said earlier. Europe needs to be investing in AI. Also a recommendation that – if I can put it this way, you know – we need to continue to invest in how we think about legal regulation in this area, because we need that legal regulation to be iterative. Yeah, it's not that there's a block of regulation, and then we say we ticked the legal regulation box, we've sorted all of the issues, we need to stay engaged and be able to adapt. And then I guess what I found most striking in what you said was the, know, the overarching vision, which was to say, well, essentially the future is now, that we need to be doing these things now to secure the positive future.

Sue Anne, the future is now. We need to be investing in AI. We need to be more flexible in how we think about the role of law. Do you share any of these views? Do you have different views? Over to you.

SAT: Yes, of course. I think the AI Act is interesting in this sense, because it's being changed even before its different provisions actually become applicable. This goes to show that there's a lot at stake when it comes to regulating AI, as Fredrik rightly observed. And in this effort to change some aspects of the AI Act - through this legislative effort called the AI Omnibus - certain changes introduced, for example, notification apps, you know, AI apps that undress people, we might recall, for example, what happened on Grok some time ago, where millions of sexual images, sexualized images of girls and women were generated and shared, that would be now banned, and I think that is the right move. It targets the harm that is experienced. But at the same time, I think also the AI Act doesn't address adequately some different kinds of harms, right? For example, we've heard of cases of harms of people when they interact long term with AI chatbots, so looking more at AI's impact on people rather than the type of risk that AI model presents, so instead of looking at

the model, looking at the impact on people and there should be sort of ongoing monitoring efforts, including from the research side to gauge the impact of AI on people, including working on, for example, benchmarks. That is partly sort of what my own research is about, as I think there's some lack of clarity on this concern at the moment. I also want to see some best practices emerging, for example, around fundamental rights impact assessments, how to assess human rights impacts in practice that can enable us to take on the lessons learned and that can hopefully help us to strike the right balance between innovation and rights protection, not to trade them off each other, but to ensure that both considerations feature together, so that we can have these, you know, European values or universal values to a large extent protected in and through the AI Act.

TM: Thanks Sue Anne, particularly for the very clear message that we shouldn't go down the route of zero sum thinking, that you know, for innovation we have to give up regulation, for innovation we have to give up rights protection. But, also, very clearly guiding us that you know that the challenges here are ongoing and they're not simple. On that note, I want to close out today's episode by thanking you, Fredrik, and you, Sue Anne, and thanking you in particular for providing us with really clear descriptions of what's been happening, AI tech-wise, law-wise in the European space and also for your generosity in sharing with us your own recommendations about the next steps. There's so much to be learned from what you have shared with us today. Thank you for sharing. Goodbye.