

Kriti Khullar:

Hello and welcome to the Actuaries Institute Podcast. My name is Kriti and I'm an Associate of the Institute. I'm currently a lead at IAG, and today I'll be discussing the digital dilemma of navigating ethics into generative AI. I'm joined by Victor Bajanov, Product Analytics Executive at Quantum and Tiberio Caetano, Chief Scientist at Gradient Institute and Adjunct Professor at UNSW.

Welcome.

Today's podcast will have a two-part approach. The first part will be focusing on introduction to ethics and generative AI. The second part will be focusing on regulation and governance. I'll direct my first question to Victor. What aspects of generative AI ethics require consideration?

Victor Bajanov:

Look, I think that question has an answer with many facets, right?

So, if you think about what's been live in the press at the moment, there are a lot of different concerns. And if we kind of rank them from the biggest to the smallest, just the ease of organisation. I mean, a lot of people who are concerned that we're at a breakthrough stage and we're closer to coming to creating an artificial general intelligence.

And then the problem that it may not be aligned with our interests becomes very serious. So colloquially, we might say that it'll go like "*Agent Smith*" and it will decide that actually it prefers to destroy humanity rather than to coexist. So, that would be an existential threat.

The next biggest one, I think is one I would categorise as bad people doing very bad things, right? So, the power of generative AI has democratised access to certain tools that maybe were not as easy to do before, not as possible for maybe people who are socially maladjusted.

So, for example, anyone can use it to create biological weapons and things like that. And actually, you can buy equipment from eBay for quite a small amount of money that will enable you to synthesise viruses and so on, that you've designed with generative AI tools.

And so that could also be an existential risk, but it's not going to be coming from a generative artificial intelligence. It will be just bad humans doing bad things with more powerful tools. And that's a risk that as a society we're probably familiar with. But it's just another vector.

Working down list, I'd say bad people doing mildly annoying things. So, things like deepfakes and fake news and so on. I think it's definitely true that you cannot generate fake images and fake videos much more easily using real people. So especially if they're in the public domain, you have a lot of training material. Equally, I'd say that the ability to write things that are not true and publish them has been around for a very long time.

Kriti Khullar:

Thanks, Victor. Tiberio, was there anything you would like to add on to that?

Tiberio Caetano:

Actuaries Institute

Level 2, 50 Carrington Street, Sydney NSW 2000

P +61 (0) 2 9239 6100 | actuaries.asn.au

That was a great summary there. And I just would add that, as Victor was pointing out, there are many different categories of risk and potential harms, and there's significant range of different opinions about which particular types of harms are more or less important than others, right?

So, it's really important in this conversation about ethics of AI that we recognise that the outset, that even though there are some objective truths about what is and what isn't dangerous, what isn't harmful when the technology does imply that there will be some types of risks. There's also a lot of uncertainty in this space, and there is reasonable people and experts significantly. Well-established experts, will reasonably disagree on many topics.

And the second point I'd like to add is that Victor mentioned a really important risk, which is the risk associated with the perhaps unwise excessive democratisation of power. If power is democratised without appropriate checks and balances, then you can have bad actors doing very, very bad things. That's one risk and it's a legitimate risk.

But there is another risk that's kind of contraposes with that a little bit, which is also very, very significant, which is the risk that if we don't democratise AI, if we don't give this power to people, then the incumbents will further increase the influence and cement the existing advantage as well.

So, I just wanted to flag that there's this power balance in the AI dynamics. We need to make sure that not too many people are too powerful because some of those will be crazy people who may design biological weapons and kill everyone, around the one end of the spectrum.

But on the other end of the spectrum, we also need to make sure that there isn't too much of an imbalance between the power of the people and the power of the most powerful incumbent organisations that are developing this technology.

Victor Bajanov:

I think that's an excellent point that I missed. And I think a lot of the press and the discussion around this topic is concerned with things like impact on the labour market, and that's also a risk of harm because every time there's a market correction, some people are harmed. Yes, Tiberio correctly said, but there is a real risk that will have a market correction as these tools become more embedded because they will change the nature of work for a lot of industries and a lot of professions.

And do we have a moral responsibility to help the people who are not on the benefiting side? But the receiving... On the negative side of the scale is a very interesting, I think, ethical debate.

Suddenly in the US in the 1910s, the answer to that was no, we let the free market do the talking and whatever happens, happens. But you also had a lot of very bad outcomes for individuals. So that that's one we could discuss.

I do think that the wrong approach would be to try and stifle the innovation and maybe restrict access to the hands of a few. I think there are a lot of other incentives that they quite transparently have, such as creating a moat that would prevent competition from taking over their market lead. And so looking at this debate, you do have to have in the back of your mind a lot of... It's really kind of a bit of a minefield.

Actuaries Institute

Level 2, 50 Carrington Street, Sydney NSW 2000

P +61 (0) 2 9239 6100 | actuaries.asn.au

Kriti Khullar:

Thanks, Victor. Tiberio, for the next question. When people are talking about AI ethics, are they talking about AI safety?

Tiberio Caetano:

Well, I guess a simple way to think about it is that there are many different legitimate concerns about what may go wrong with AI, okay?

And from bias to misinformation, to copyright violation, to labour market disruptions, to bad actors creating deadly pathogens, to the skilling of the population, to manipulation of users. There's a wide range of concerns, all right?

That's the starting point as Victor and I were just discussing now, and that's all legitimate. That's the first point.

Now the second point now, which connects to your question is where does the word ethics comes into the entire set of concerns? Where do the words like safety or accountability or fairness or transparency, all these words that we use to try to refer to actually how we want to deal with these concerns, which types of values and attitudes do we want to develop in order to deal with those concerns?

Safety is a word that's usually associated with situations in which very bad things can happen. If you go into an airplane and the airplane principle can fall, right? So that's very bad. If you drive a car, the car can crash and that can kill you. That's very bad. So often the word safety comes in when the worst-case scenario involves some personal injury or death or something quite significant like that.

And I think the whole debate around AI safety, that word has been used largely to refer to these situations of risk where the worst-case scenarios could actually be something quite significant in terms of death or permanent personal injury or permanent psychological injury and so on and so forth.

Now, ethics is different because ethics is pretty much everything that matters. That's kind of the definition of ethics, right? So, safety, will interior include ethics and just like transparency and accountability and so on and so forth.

Victor Bajanov:

Look, I agree a hundred percent, and I will add though, just because something is not an immediate danger to the lives and livelihoods of millions of people doesn't mean that it isn't important.

And so, I think there are real risks related to the safety of AI solutions where they would lead to real serious harms like death or injury. And it is important that we think about how we can address those and mitigate those risks. But it is also important not to get distracted by scenarios that may be possible, but probably on the scale of probability, more in the realms of the distance future or current science fictional kind of scenarios. And for that distraction to divert... I'm going to use in the broad sense investment, but I include under that an umbrella word, the mental effort of intelligent

people who could otherwise solve other relevant problems to divert that from very real ethical issues.

And because the problems of ethics in AI, even without generative AI, the problems of how to do ethics correctly in AI, they're not a hundred percent solved. And I think that will impact the way that society can extract the benefits of AI. I think that will impede the adoption and it will create negative value from bad actors and from systems and safeguards not being in place to ensure ethical behaviour and give the industry a bigger license to operate. I think that's a real economic risk, and that is probably less important than someone dying as a result of an AI solution not being well done. So I think in that sense, that should get priority.

But again, the scenario where the AI evades human control and becomes super intelligent superpower agents, that then becomes an existential threat to humans. I think in that scenario that would be very bad. But if it's a hypothetical scenario, right? But we're still, I think a very, very long way away from that. And I don't think ChatGPT has really brought us materially closer to that scenario. I don't think we have to worry about that today. I think devoting a huge amount of resources to that today would be a distraction from other issues that are arguably more important over the time horizon that matters.

Tiberio Caetano:

I do agree with this period of what Victor just said.

I think I do have a slightly different perspective on one element here, which is the question of what constitutes a hypothetical, right? It's really hard sometimes to know what is true, to find out what is true. And then if experts and intelligent people have a wide range of opinions about very, very high stakes issues, I think we need to be careful not to take the same actions that we would take in calling it hypothetical or in calling it speculative or calling it or labelling some risks.

Let's say for example, evasion of human control as hypothetical, we should be careful not to take the same actions as we would take if we would label it as non-existent. Because what really matters is what we do, right? So often I see us framing the situation, and I'm among the people who will say, first thing, big priority here.

We have finite resources. We need to use those finite resources wisely. Our attention, our intelligence, our money, we can't spend everywhere. There are lots of risks. We can't give equal attention to all of those. We need to prioritize. I fully agree with that. But the question of how to prioritize isn't an easy one, and we shouldn't, I believe, be thinking that by labelling something as hypothetical that sort of gives a license not to actually take some type of action in that space.

So to me, the real challenge is what do we do about each of those risks, no matter how hypothetical, no matter how concrete, but the question is what do we do about them, right? Because if someone comes to me and say, *"Well, look, I'll do the same thing, for a hypothetical risk, then for a risk that I don't recognise even, and it's hypothetical,"* then you are basically saying the hypothetical risk is not hypothetical. It simply doesn't exist.

So, I think it's important to really focus on the pragmatic actions as opposed to labelling as hypothetical or not hypothetical and so on and so forth.

Victor Bajanov:

I agree, and I don't think... the alignment problem is very interesting, and suddenly there are people who have dedicated a lot of research time to that, and we shouldn't tell those people, "Look, your work is not valuable because it's addressing a risk that's never going to happen." I don't think that's the case.

I do think though, some of these questions, the alignment problem is one of them actually, is what should we do about it? And the answer is... We actually haven't got a good answer to that yet. We need to go away and think about it a bit more. And then it's like, okay, well, how much time should we spend thinking about it and how many people, because as you said, the number of resources is finite and we have a fixed number of people with the right qualifications to really make an impact on any of these problems. It's not a small number, but it's a finite number.

And how many do we dedicate to each of the problems is what we need to be cognizant of. And in doing that prioritisation, I think it would be reasonable to say, what's a one-year horizon problem? What's a five-year horizon problem? What's a 50-year problem? What's a 500-year problem? And humans have got those wrong very many times in history.

Tiberio Caetano:

It's very interesting because a few years ago, the buyers of the top AI minds in the world was exactly in the opposite direction. Whether they're exactly saying, *"Well, look, we won't have AGI for the next 20 or 30 years."*

The guesses were much further in the future, but almost no one has updated the estimates of AGI for a longer period of time in the past two years, or everyone has shrunk the estimates to the near future, right?

So, the average of that distribution has just approached the present dramatically. If you just look at prediction markets or even metacules estimates, you will see, I think, estimates for AGI in the 2030s or somewhere.

But I agree. I mean, people can be very, very wrong in their estimates both ways.

Victor Bajanov:

If you look at historical investment in AI and the opinions of top experts, the '60s and '70s, they had some very thick gains in computer science algorithms and were making these quite aggressive predictions and then failed to deliver through the '80s. And then in the '90s, it was almost like you tried to avoid having AI in your grant proposal because it would automatically mean you didn't get funding. Because it was like a toxic label. And then during that period, the prediction of denial, that's maybe that's not even possible. And if it is possible, it will be hundreds of years away.

And so, I think it does have a cyclical nature, and at the moment we're obviously in a more positive side of the cycle, and it could be that there's some truth to that, but I think it's more likely that the cycle will just swing back the other way as it has in the past.

Kriti Khullar:

I'll move on to the final question of the first part of our podcast. So in which ways does GenAI offer different ethical considerations compared to traditional technologies? Victor, I might direct this question to you.

Victor Bajanov:

Sure.

So, look, I think it's worth saying first that the ethical considerations in traditional technologies, it's not a solved problem. It's a problem that is still being worked on, right? And even though we have some very good tools and techniques that have advanced that field significantly in the 20 years or since I started my career, there's still quite significant barriers to getting to the position where you could say, "*Yeah, that's a tick.*" And we have general acceptance from social stakeholders that we are kind of nailing that.

But there is one specific thing where generative AI... Specific way, where generative AI adds an extra challenge that is very difficult to overcome, which is transparency of the actual algorithms. So as a bit of technical background, all the generative AI techniques today are basically special cases of neural network models.

So, there's specific neural network architectures and neural networks inherently just don't lend themselves to transparency measures or techniques the way that tree-based models or boosted models or linear models do.

Tiberio Caetano:

I think this point that Victor is making all the interpretability and transparency of models is really a fundamental difference between generative AI and traditional types of... more traditional types of AI.

I would add two additional important differences. One is that as Victor explained, these generative AI systems, they are built with neural networks and neural networks are becoming too powerful. It's becoming easy to make them more and more intelligent because just more compute is sufficient to make them more intelligent.

So, it's much less about innovation and research than it is about resource scaling. If you've got a hundred thousand GPUs trained during a month, on data that's available online, you can build a version of GPT-3.5, probably.

Kriti Khullar:

Tiberio, should there be regulations or standards for the ethical developments and deployment of generative AI systems like ChatGPT?

Tiberio Caetano:

Actuaries Institute

Level 2, 50 Carrington Street, Sydney NSW 2000

P +61 (0) 2 9239 6100 | actuaries.asn.au

I think it's probably fair to say that the global community has already come to the almost consensus that regulation of AI is an imperative regulation that is not self-regulation.

So, some type of government enforced rules and types of law that make sure that we have appropriate checks and balances for AI. Now, what that should look like is where the debate is.

Some people think that that type of regulation should focus on a risk-based approach, which I think in principle makes a lot of sense because as per all our conversation so far, there are different risks and some risks are more significant than others. And the controls that are going to be effective to address different risks are going to be different. So, it's important that we take a risk lens to addressing the regulatory question.

And there are efforts happening in the European Union, in the UK, the United States. There have been several senate hearings recently on AI, addressing several different types of concerns. And there is, by and large, a great agreement that AI needs to be regulated. The Australian government has taken recent action in this area by writing us a paper, asking for submissions for opinions of experts on how AI should be regulated in Australia. So Gradient Institute alongside many other parties who have made submissions and-

Victor Bajanov:

And the Actuaries Institute as well. So, for our listeners, they need not worry that we're not up to the cutting edge here.

Tiberio Caetano:

Indeed, indeed. And in Gradient Institute submission, we actually refer to the Actuaries Institute submission because we collaborated very closely with the people who wrote that submission as well.

Victor Bajanov:

I'll add to that. I think it's pretty clear that regulation is necessary, and that's because it is a set of tools and systems that can impose external costs so easily. Like in economics, when you have an external cost, as in the person doing the action is not the one who suffers the damages, then generally you need some kind of intervention to prevent that because the incentives are not there for the person doing the action to check themselves and make sure they're finding the right balance. And examples of really successful regulation has been where the regulation is focused on the outcome and it drives technical innovation.

So, for example, emission standards regulation for cars in the EU have been very successful at driving automakers to innovate how they can make engines that produce less emission per kilometre driven for the same amount of power or torque output.

And so, where we are now with efficient petrol engines and efficient diesel engines compared to 20 years ago, is like night and day difference as a result of that incentive that is artificially created for them through regulation, it can't lead the technology.

So I would say for AI, it's no different. We can have a successful regulation that treats the outcomes, right? So, we don't want self-driving cars that don't meet the safety standards that we would expect of human drivers, right? We don't want Chatbots that tell people to harm themselves and things like that. And those are outcomes. And then it's not about how you implement the Chatbot or which version of GPT you're using or whatever. It's kind of your responsibility to find the good way to do that, and that will drive innovation and competition, and that's how we'll get better at making quality guardrails for LLMs, which by the way is going to be the thing that distinguishes the successful companies from the failures in the LLM landscape.

Tiberio Caetano:

Yeah. No, that's a good point. I think that the question of... The question is whether AI favours defence or offense, right?

Fundamentally, because if the answer is... well, if favours defence, then let it rip, right? Because people will eventually figure out how to protect -

Victor Bajanov:

Assuming you can withhold it from both sides equally, right? But if you can only withhold it from the defenders, then even if it favours offense, you're still better off on the defensive side than without it.

Tiberio Caetano:

Yeah, yeah, that's right.

So, there is a line of argument that would be put roughly like this, right? So there is scope, there's so much potential for AI to actually cause good for the world. It's truly unimaginable. What we could do with... Intelligence is everything, right?

It's what gave us pretty much everything that's different about humans and other animals. It's all about intelligence. It's all about creation of knowledge. It's all about improving wellbeing and science and everything, right? So now that's unquestionable.

Now the question that I think to me is really important one, is how do we don't screw up? Yeah? How do we make sure that we really protect ourselves against really bad things happening? That won't even give us a chance of realising that upside. And I don't think we can gamble too much. I think we need to leverage the good and old precautionary principle.

We need to compare the worst-case scenarios of different courses of action and do appropriate risk management. It's hard. It's difficult. Doing good risk management involves-

Victor Bajanov:

It's a lot of work.

Tiberio Caetano:

It's a lot of work, involves a lot of domain expertise. There will be a lot of disagreement, legitimate disagreement, but it's not like we have a lot of options. I think that what is really important is that we think in terms of prevention, because by the time... It's just like health, right? If you start paying

attention, whether you have high blood pressure in your thirties, not in your fifties, right? Or whether you have high cholesterol in your thirties, not in your fifties, because those things are going to kill you in the long term. The last decade of your life will be very different if you started taking statins or monitoring or controlling your blood pressure in your forties as opposed to in your sixties.

So, we kind of need to do the same with AI, I think. It needs to be preventative, it needs to be proactive. We need to design the systems responsibly. We need to put the right guardrails in place, and we need to be inclusive of different views, but we need to be ready to make tough decisions.

Kriti Khullar:

So, any final thoughts from yourself, Victor, as we wrap up today's podcast?

Victor Bajanov:

Look, I think it's been said a few times, but the potential of AI to better society is just enormous. And I heard of a study that said... Basically, correlated individual intelligence with basically every positive outcome you can be better for a person. So better health, better financial security, better outcome for your children, basically all are correlated with intelligence. So, if you can have an assistant that just adds 20 IQ points to you, whatever you've got, like if you're IQ of a hundred, you get 120. If you've got 200, you get 220.

Just the power of that to unlock good outcomes for everyone is mind-boggling. So I'm personally very excited for that, and it's an amazing landmark moment in history.

Kriti Khullar:

Thanks, Victor. Tiberio, how about yourself? Any final thoughts?

Tiberio Caetano:

I've been working in AI for over 25 years, and what brought me to this field was the conviction that a good future for humanity would hinge on us building machines that were more intelligent than us and could be our partners in helping us towards the future of flourishing. So it was the future of more wellbeing for everyone.

Today, I am still hopeful and believe that that can be achieved, but at the same time, I do recognize that it's not going to be achieved by default, by just making these machines more capable.

We actually need to deliberately be wise, as opposed to just smart. Intelligence and wisdom are not the same thing. Intelligence enables us to be wiser, but doesn't guarantee that we effectively become wiser. Wisdom requires other qualities like prudence, caution, humility, generosity.

If we broaden the scope of our thinking when we talk about AI to include a wider range of values to come along with the value of intelligence, then intelligence, I believe, will deliver on its promise, at least on the promise that I had for it 25 years ago.

Kriti Khullar:

Thank you, Victor and Tiberio for today. Listeners, we hope you enjoyed this discussion with Victor and Tiberio.

If you're hungry for more conversations and insights, we've got you covered. There is a wealth of Actuaries Institute podcasts waiting for you at our podcast hub, which you can access via the podcast section of Actuaries Digital, or via all mainstream podcast apps.

If you'd like to continue the conversation with us social media, join us on Facebook and LinkedIn at Actuaries Institute, or you can reach us via Instagram @ActuariesInst. Please write in with your questions and comments on the show. We'd love to hear from you.

I'm Kriti Khullar. Bye for now.