AI and You

Transcript

Guest: Tony Gillespie, part 2

Episode 55

First Aired: Monday, July 5, 2021

Welcome to episode 55! We will be concluding the interview with Tony Gillespie today. He is the author of *Systems Engineering for Ethical Autonomous Systems*, which is about the issues of transfer of responsibility for decisions and actions from humans to machines in the context of product design.

That means he's applying systems engineering principles to the design of autonomous systems, like weapons and cars, and the ethical and regulatory frameworks they operate within. What first grabbed me was when Tony said that we can make AI that can obey the Geneva Convention.

He is a Visiting Professor at University College London, a fellow of the Royal Academy of Engineering, and a fellow in avionics and mission systems in the UK's Defence Science and Technology Laboratory. He has applied the techniques in his book to autonomous cars and autonomous weapons and has given technical advice to the UN meetings discussing potential bans on lethal autonomous weapons.

Last week we talked about autonomous systems and in particular military systems and how to engineer them to stay within rules of engagement and other ethical boundaries. We're going to do more looking at those systems and autonomous vehicles through the lens of systems engineering as we get back to the interview with Tony Gillespie.

We've been talking about machines here that are at the pointy end of the conflict chain but what about further up? Are you involved in decision support systems that would be used higher in the command chain, because those are surely evolving to the point of using AI to decide things like targets, strategic targets? And if they're reflecting the trends in AI in enterprises, they are encountering issues of explainability or accountability, of, can I trust this system when it says this and I don't know how it arrived at that decision.

I'll make it clear, I retired from the military side in 2014. So, I don't have any hidden insight that I'm quietly hiding away, just the experience, and what I've been thinking about and looking at unclassified things since. But yes, this question of explainability is fundamental, I think, not just to the military, but also to in cars, and the whole of AI. I think that the problem with explainability is, how much information do they have, and getting them the right information. I think the question they have to keep asking is, is what I am being told a reasonable. We're back to the lawyer versus the engineer. But at that level, you're talking about a military man using his judgment. But that's where international humanitarian law says that, in effect, I mean, a lawyer might shoot me for the way I'm putting it. But basically, he can make his decision in good faith based on the information he had. Now, that means any reasonable person would have come to the same decision, what you cannot do, and somebody from Yugoslavia was actually prosecuted for this, is to withhold information. In Yugoslavia, that particular person withheld information from his forces, that there was a ceasefire, and the snipers carried on shooting people.

But with AI, we could have information being withheld that is not deliberately withheld, but the AI just didn't bring it up, and it would have been material, but no one knows that it's been withheld.

That's where I think there's a whole range of problems that come up there that I don't know the answer to and I hope there are people in classified rooms around the world worrying about this. Again, this brings out the parallels between the military systems and civilian systems, that a lot of the technologies that are being developed in the commercial and civil world for all good reasons, if only insurance companies, on explaining the AI decisions, it's actually beneficial to feed that back into the military. I know this is in complete disagreement with the people who have been saying, we don't want our open-source software to be used by the military, and so on, because they think the military are always going to do evil with it. I think the problem is we've got to make sure the military have access to the best technology, to make sure they do do the best jobs that can be done, and made sure that the systems are such that the international lawyers are on their necks to make sure that the decisions are legal and humanitarian. In fact, there was a very good comment when I was giving my presentation in the question session. I was one of several engineers giving presentations, and several of us had taken the viewpoint of the developers of the autonomous weapons, the high-tech end, and there was a very good comment from one of the other countries, it may have been Cuba, I'm not sure. They said, "We are the people who are being surveilled by these systems and they will be used on us. How is the law going to ensure that we know that the people looking at us are actually following the law, and they're using weapons that meet international law?" That made me think, because if you're talking about transparent processes--and this was fascinating, as an engineer, listening to diplomats discussing these things, and military people, and lawyers, actually listening to some of the problems that people have--and that really struck a chord with me, and makes me think that, as engineers, we've got to engineer the systems in a way that doesn't betray the nation's little secrets or tricks that give them the edge. But they've got to be engineered in a way that an independent person technically clued up, can look at the evidence you're presenting and be sure that he's authorizing something that is legitimate, and that you on the inside are actually telling him enough, and that enough is sufficient, you're not actually hiding something, and that's the trick. Very hard to do, that's why it's taking years for them to work it out.

You bring up something there, which I think is important to remember and easy for us to forget, that most of the people listening to this podcast are going to be in countries where they're not routinely surveilled by another country, in the way that people like you mentioned are. That there are these people in the Middle East and other areas where they're under scrutiny by the Armed Forces of the United States and other places. It's very hard for us to imagine that, even when it occurs to us what that must be like, and that's the point of international humanitarian law, is to address things like that and make sure that we *do* think about them.

Yeah, it's a completely different psychology, and if we don't recognize that, we're actually failing in some of our some of our aims, in our professional guidance, our code of conduct or something like that.

Now, you have mentioned several times here the commercial sphere. So let's look at that crossover because you work equally on automated self-driving car systems and the impact of laws and liability with respect to those. That, of course, is an area where there's a lot more speculation than there is reality, but it is having an incredible amount of money thrown at it. At the moment, I think I identified something like seven possible targets for litigation in the event of a significant accident. I know that lawyers are just waiting for the chance for this so that in the event of an AV having an accident, who's responsible? Is it the car, is that the manufacturer of the car, is it the driver of the car, is it the owner of the car, or the person who configured the car's optional settings? Maybe they were irresponsible in setting it too sporty for a teenager who learned to drive. Or is it the third-party company that made the self-driving software that the manufacturer uses, or the lab that trained that, or all the drivers whose test data they use in going into that? I mean, that's going out on a limb there, but lawyers love to think about these sorts of angels on a pinhead thing. So, to what extent is that fanciful, or to what extent is it just a test case of autonomous car accident and lawyers start suing all of those people that have deep pockets?

Yeah, it will happen. When I wrote my book, it is aimed not just at weapons, but seven chapters out of the thirteen are weapons related and I applied it to a mythical robot, iRobot type person, and I felt that was a bit nebulous. Then I thought autonomous cars might be a good way of doing it. I was using a thing called the 4D-RCS architecture for the systems engineer nerds among us they were recognize it. I was actually triggered by a paper from a guy from a university in Nijmegen in the Netherlands; where he'd apply that same architecture to cars. So, I then developed that and put out the paper you mentioned earlier with Steve Hails from UCL. Lo and behold it actually worked in the sense that you could see how to divide things with the decisions and actions and put limits into that. So that really got my interest going in cars, or developing that. So, I've actually taken that thinking a bit further forward, and you've gone through the owner and so on. and the driver, and you can actually write down the chain of command from the manufacturer and what he's saying, right through to who owns it, who maintains it and updates the software. When do you stop it learning? How do you get the options? How do you stop the car learning one driver and adjusting itself to them? Then the sporty teenager kid gets in and gets frustrated, or it's learned the sporty teenager kid, and a 75-year-old person gets in and can't react as fast and so on. So, there's that sort of problem. But you've got all these factors coming in and yes, it will reach a legal problem. The two-parted point is that the tests of more autonomous cars and delivery pods and so on are being done in the states in special areas which are well controlled. In the UK, we've got them as well. But what the insurance companies did in Britain was they got together with the government and there are a couple of extra clauses in one of the road traffic acts in 2018, where basically, the insurance companies said, we know this problem is coming. So, let's do it in a manageable way. But you've got to do the experiments and get things on the road. So, they said, okay, if there's an accident in a test, where it's under controlled conditions, and all the protocols of how to control it on a let's call it an open road, but on a public road under controlled conditions, if an autonomous car has an accident, we the insurance companies will pick up the bill in the way we would do with any road accident. Then they look to see if somebody has been negligent on the test side, but the public know they will get their bill

paid, and no doubt grumble because the insurance company never pays as much as you want them to pay, but it is definitely a positive step. So that, again, you're bringing the law and technology forward together.

Right and it's in the insurance company's best interest because if AV technology works out, then they'll be paying much less in claims.

Also, what they have warmed to as well is that it's identifiable who you make the claim against. The thinking bringing the military side through, you get to the point of, where does the manufacture stop? You probably familiar with the concept of a design authority who signed something off to say, yes, that's a correctly working product, then with a current car, that car isn't safe, and he goes out and is used in the maintainers follow the maintainers procedures and so on if you get it serviced regularly, all the other good things you do, then, you know, the fault is on the driver. Then you get this interesting point, which is I think analogous of people like Volkswagen, where they had the emission test problem, I think it was Volkswagen, and Mercedes and now all these court cases that are about to happen, saying the driver and owner wasn't responsible for this. The manufacturers are, and the maintainers aren't because they did what the manufacturer has told them or not followed this in detail, right? It's opening this door. You understand this phrase, plumbers harvest, which is hard frost is plumber's harvest, all the pipes freeze, the plumbers have lots of work the next day. You can see it's almost a lawyer's harvest, where the door is beginning to open and other people can be brought in for litigation. I think we as engineers have to make sure that--not to keep the lawyers under control, but to make sure that there can be rapid agreement as to, after the event, which organization was at fault, or even better, before the design happens, say, where are the weak points? Where are these things going to happen? We can design to make sure that the people know they're responsible for that particular happening.

Just to save our listeners some googling there, the Volkswagen incident that you're referring to was when they engineered their cars to lie about their emissions when they were being tested so that they would pass tests they might otherwise not pass for smog emissions. So, in that case, the normal principle that the owner is responsible for what their car does is in abeyance because they could not be reasonably expected to know that this had happened, it was concealed and only uncovered at considerable difficulty of investigation. I want to see if I've correctly interpreted some of the principles in your book about system engineering for autonomous systems, because I think there's an interesting point--at least interesting to engineers like myself there—that, in the great amount of technical detail you go into, it looks to me that you were defining broad principles of architectures for autonomous systems, such that you could delineate boundaries between subsystems of these such that you could encapsulate matters of decision making, or autonomy as they related to the law inside specific systems in the hardware and software. Did I get that all right?

Yeah, you did, and there's an assumption, which maybe not all this is one I've picked up because it's fairly early on. It's one of the bees I have in my bonnet, that sometimes people say they have an architecture, when they've simply drawn all the connections between the boxes, and to some

extent, it's a description, and is quite useful. But if you have a real systems engineering architecture, you would like to do a top-down analysis. Very often, you've got half the boxes, or half the functions, and you do it by function, and start at the top, which is where decisions and actions are functions, then you use that structure to drive the requirements, and all the interfaces, and the point of bringing it in or as you say, all the stuff you've got right about the authorization, the legal, and the requirements, it's actually to make sure that the architecture meets the top level purpose, meeting the law, and then the requirements flow down. As any engineer knows, when you've got a flow down of requirements, you can go away and do your job and know that the requirements are right. Okay, you've then got going up to the other side of the V, or--nothing's ever quite done in a V anymore, but in the spiral, or whatever it i--you know that you're doing what you're supposed to be doing, and then in the next iteration round, you can find out whether what you were told to do was right or not.

So, is your book, then, prescriptive that it's a blueprint for people building or automated systems, autonomous systems to say, this is going to tell us how to design the hierarchy of systems here such that it becomes as easy as it can be, to ensure these questions of industrial safety and liability and authority?

Yes, it's guidance. I wouldn't say it's prescriptive. It's not a standard or a regulation. But it's saying, here is a method that has worked, and the car paper says it's worked in a different field. So, interpret the book, and apply it, and use it to answer questions. A lot of the introductory stuff about what systems engineering is, and so on, is to really understand what you're doing. You need to have a top-level view what I always call the PowerPoint presentation level plus one level of questions below it, understanding of the whole process, so you know where your bit fits in. Then you can ask the questions of the next person up the chain, or somebody outside the chain, what are we doing? Is this actually covered in your bit because it's not in mind? So, it's guidance.

Right. Because at the end of the day, these things have to be built by humans and our brains are only so big, and so, we've got to have this simplified to the point where I's possible for us to build them.

One of the other points about the book is I wrote it in a way that a lawyer should be able to pick it up and look at it and one or two have, and said yes, they like it. So, the lawyer has an appreciation of the engineering problems, as well as the engineer having an appreciation of the lawyer's problems.

What do you think they will do or should do with that appreciation?

I think have a good dialogue with the engineers, and establish a level of mutual understanding and identify where the problems are in their understanding, because usually it's a problem of the language and the concepts.

So, can you tie it together? We've been talking about weapon systems, we've been talking about cars but at the levels of abstraction that you're dealing with, they have a great deal in common. From the high altitude, what are those things that they have in common?

They can cause harm to humans. You've got, in the weapon case, to make sure it's the humans you want to harm, but you don't want to harm anybody else. In fact, the idea of war is not to kill people; the idea of war is to make the enemy change their mind. That's again, an underlying principle, and you don't have to kill them to make them change their mind. But the idea is, you've got something that can cause harm and with machine learning and AI, it's going to be able to cause harm in ways that we don't yet understand. So we need to ensure that we keep it under control no matter what. I thought of something the other day. Back when I was in the sixth form, which is many decades ago, I went on a tour of a computer factory. I don't think they had valves, I think they really were transistors and the like, and they said, "The only name for a computer is TOM, totally obedient moron." I think now we're in the 21st century, we've lost the O, they're now totally moronic and our problem is making sure that we do keep them obedient.

Well, let's then ask the question, how can you limit the damage an AI system can do when it is not under immediate human control, since the development of weapon systems is clearly heading in that direction?

Yeah, the idea I put forward in the book, and this is my take on what it is, is a thing called authorized power, which in each of these subsystems, it's a definition of what it can and can't do, and it doesn't matter if you do it the way I propose it. But basically, I reckon in every function rather than a subsystem, every function, which is doing something. It may have many subsystems or parts and subsystems, that that function, when it's made all its decisions, you have another little module or function, which says, here's the range of decisions, here's how it writes them: number 1, 2, 3, and 4 in choice. I've actually got a rule book, which says, this is what you can and can't do. It's different. For those who are interested, it is very different from the ethical governors that have been around, this is bottom up, because it says at the lowest level, where it's a mechanical system, you've got to stop. So, it won't go beyond the mechanical bit, won't go beyond that stop, you can't turn the steering wheel too far, because you physically stopped and you work up the system. So, the behavior of the whole system is actually the sum of all those stops. Now, what I would love to do, but I know I have not gotten the mathematical ability at the moment to do it, is to derive that mathematically. Because what you want to know is that if you sum it up, you've got a complete solution and it's not at that stage yet. But I think that could be an interesting development. I think that's the way we've got to approach it. I like cars and weapons because they're practical, they're with us, and you've got poor engineers, marketing people and everybody else actually struggling with it when they know the law isn't clear and they're concerned, what's the impact? If you can give them some assurance of the risks, then they can do it and sleep at night.

I've long said that ethics is defensive programming writ large, that it's putting in these kinds of exception handlers for things that you don't want to have happen even though the system shouldn't do that in as far as you think you've engineered it. The kind of modules that you're talking about there are that I think at a higher level and make me think, well, if only HAL 9000 had those, maybe it wouldn't have killed off the crew, just because he got confused. I'm also reminded of--I read the SAE standard on autonomous levels a couple of days ago. I remember that it says, for instance, at level three, a vehicle should be able to tell if, say, a tie rod has come

off and the steering is no longer reliable, and bring the car to a stop in a safe place. I think that's an example of the kind of thoughtfulness and the kind of defensive programming that you were talking about that is not often associated with those levels of autonomy but that's looking at them from a safety engineering standpoint.

Yeah, it may be a little bit off at a tangent but I think it was the first fatal accident where a car went under the back of a lorry.

Yes.

And what frightened me about that wasn't that the poor sod was killed- I mean all sympathies to the poor chap who died, but the car didn't stop. It was stopped by a pole or some obstruction, it didn't know it was going along with a dead driver, and badly damaged. I think that lesson has been learned. But it's that sort of thing you've got to take care of.

So, let's look at some final questions here. What do you think the biggest changes from the development of artificial intelligence in the way it's proceeding are going to be to our concepts and practice of law and liability with respect to autonomous vehicles and weapon systems? Looking out ahead a little bit?

I think that the lawyers and the engineers are going to get together. I know they're doing in the UK, there's a thing called ALKS -- Automatic Lane Keeping System which the UN put out regulation 157, or whatever it is. And all the countries, or most of the countries, are looking at how you can implement it. In fact, I know this is probably a very parochial view--but, in the UK, the Law Society, which guides government on its legislation, has actually put out a consultation document which the professional engineering community has responded to, saying, not as well as the nitty gritty of, is it defined as autonomy. but is this just on the automated side of this whole problem of AI and machine learning? Well the engineers are saying, yes, it is. The lawyers are beginning to suck their teeth and say, we didn't want that answer, but we did expect it. So, how are we going to go forward? So, I think, like with many things, if you can get the law and the technology together, they evolve together, and from an ethical point of view, I think that if you get the lawyers in early, and guide the technology in its applications, then you can get to an acceptable use of the technology, the worst thing of all, is the technologies roar on ahead, and then have to get hold up, because you've developed all the wrong technologies. It's back to that original point about manned and unmanned aircraft, where we realized as a group, we could solve any of the technical problems rather than say, here's an interesting one, from a techie point of view, let's actually ask a military lawyer what we can do. That really guided an awful lot of thinking for about 10 years. I think if we do the same on cars, it'll guide car development and then I think that'll bleed across into areas we've not talked about, like medicine, and robotic surgery, and things like that, because they're all looking at it. I know the medical professional are looking at it, but they're even more conservative than many other people, they are definitely keeping the surgeon in the loop, if only for their professional pride.

Right. I like the conclusion and hope that it materializes, that there's more of this kind of dialogue going as opposed to the traditional engineering development cycle of, let's build

something, toss it over the wall to the outside world, ignore what they do with it, and go on to build the next thing. I think that kind of dialogue is very important and necessary. We've had other lawyers on the show already talking about this kind of thing, so I can see that this is a movement that's happening. For our listeners who want to find out more about you, of course we have mentioned your books and have links to them in the notes. What do you want to tell them about how to find the things that you have done or will be doing?

I think if you look at the book and look at the car paper, which you put a link to, and the Russo paper, Good Practice for the Development of Autonomous Weapons, they looked at the problems of risks, and how you ought to regulate what the machine does after you've let it go. Also, I gave a presentation, that's a AAAI conference workshop a few weeks ago, and I'm hoping to write that up. So, I think I'll just push it out. I'm on LinkedIn, but I'm not on most of the social media networks, so I'm a little bit conservative like that, because that's why I'm very grateful that you have invited me to do this podcast. Thank you very much.

Just if there's someone listening, to go back to the question we opened with, someone listening, who's thinking, well, I am inspired, I want to do that and they're at the beginning of their career and maybe still got some education choices ahead of them. What would you suggest that they learn, study or do for experience?

I would say engineering school. Many years ago, I was customer technical advisor for a particular program, and we had a very clever young lady doing the engineering. I understand fully why at one meeting it was announced that she was going, and she was going to be a patent lawyer. I was rather surprised, and she said she was but the lawyers who interviewed her said, we can always train an engineer to become a lawyer. We can never train a lawyer to become an engineer.

I resonate with that. I had a good friend of mine, who was an engineer at Jet Propulsion Lab who left to become a successful lawyer; not aware of any transitions in the opposite direction. Tony Gillespie, thank you for coming on AI and U.

Thank you.

That's the end of the interview. What do you think about how we looked at the legal and ethical issues of autonomous systems through the eyes of a professional engineer there? And do you think a lawyer *could* be trained to be an engineer? Do you know one?

The show transcript contains a couple of links from Tony for anyone interested in a similar career as we were talking about, and a link to his book:

- https://www.theiet.org/career/routes-to-engineering/
- https://www.raeng.org.uk/education "Useful but aimed more at providers. The Engineering Careers page may be the most useful"

I've also put there links to the two papers he mentioned, the one on legal responsibilities for decisions by autonomous cars and the one on Good Practice for the Development of Autonomous Weapons:

- https://ieeexplore.ieee.org/abstract/document/9159671
- https://www.tandfonline.com/eprint/RTCSN8YA5QDRMAEM35XQ/full

In today's news ripped from the headlines about AI, LG, the company that makes fridges and phones, has an AI research arm that has announced plans to spend more than $100 million over the next three years to create a general-purpose artificial neural network, which it calls "mega-scale" AI.

Research chief Bae (BYE) Kyung-hoon said, "Despite the improvement of AI technology, autonomous vehicles are still far less responsive to outside stimulus than Formula 1 drivers, and the communication skills of chatbots are not comparable to those of psychiatrists. To realize the potential of mega-scale AI, we will funnel all our resources in developing an AI which will beat top human experts. We will not make many AIs for specific areas. It will be a general-purpose AI, which will be better than human experts across the board."

So that's a pretty ambitious plan, and there are plenty of people who will say that artificial general intelligence isn't close to being realized even if you spend $100 million on it. However, I'll make two observations: (1) They didn't use the words "artificial general intelligence," which are well-known as meaning AI equivalent to humans at general tasks; they said "general-purpose" AI. That could mean that it is good ay many things but still not human-scale. And (2), if you spend $100 million on AI, you're going to come up with *something* useful. Can't wait to see what it is.

Next week I will be talking with Przemek Chojecki, all the way from Poland, where is in in the Forbes 30 under 30 list. He is the CEO of the technological group ulam.ai and of Contentyze, a new platform that provides an AI-powered text editor that can help write your content for you, and I swear I didn't use it to write this bit.

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