AI and You

Transcript

Guest: Tony Gillespie, part 1

Episode 54

First Aired: Monday, June 28, 2021

Welcome to episode 54! From last week's completely speculative look at AI in fiction we're going to the nitty-gritty of engineering today's AI. I will be talking with Tony Gillespie, who came to my attention through his book *Systems Engineering for Ethical Autonomous Systems*, which is catnip for engineers like myself but is going to be a bit dense for anyone else. The blurb for that book is perfect, so I'll quote it here: "The transfer of responsibility for decisions and actions from humans to machines presents difficult problems for all those concerned with new concepts, their development and use. This book gives practical help by discussing the issues in the context of product design, and gives a methodology to solve them." How that translates to our show's theme is in this fascinating application of systems engineering principles to the design of autonomous systems and the ethical and regulatory frameworks they operate within. Which comes to a real practical application when Tony makes the arresting claim 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. So let's get to the interview.

Tony, welcome to the show.

Hello, thank you for inviting me.

You're welcome. Now you are working in the fields of artificial intelligence ethics and as it's applied, especially to lethal autonomous weapons and self-driving cars. I could hardly think of a hotter field to be working in AI and governance right now and I just want to know how that happened for you. If we've got people who are listening, who are thinking, wow, that is hot, how does that happen to someone? Can you tell us what you were doing beforehand, and what spark or moment or coincidence led to your current line of work?

Okay, you've already said in my bio, the time got various bits of background, academic industry, and government. The reason I got really interested in this was when I was working for DSTL, which is the government MOD science, and technology arm, and I was actually involved with avionics, and aircraft and weapons. I led a group, NATO group, who was looking at the problems of unmanned aircraft and flying them with manned aircraft. This is back in the early 2000s. So, we're not talking about intelligent unmanned aircraft, but ones that would be slaved to the manned aircraft. And the problem was, How do you get them to work together, What technology do you need, and How are they going to operate? And initially, we thought it was problems of different speeds of the aircraft and so on. But it became very obvious that we could solve any of the technical problems if you gave it enough money or time, which of course, you don't always have that. But it became very clear that what was really dominating everything was the law, what was then called the law of armed conflict. Now, it's called international humanitarian law, or IHL. And that dictated everything that happens, because if the pilot release is fine with an unmanned aircraft, then under the law, the pilot or whoever presses the button on the bomb is legally responsible for the effect and that then drives the technology. So, the light-bulb-coming-on-moment came after a lot of thinking and scratching my head and so on. And literally, I was on holiday. A wet afternoon in a tent somewhere in Cornwall, I won't bore you with where; suddenly I realized that he was actually quite simple to put the Geneva Conventions into engineering requirements. So, a lot of scribbling on paper, then going off and failing, the rain cleared, and came back and put it into the first paper I put out with a good colleague of mine, Robin West. That really opened the door to all sorts of things. Having written engineering requirements, then I could go away and say, this is what we've got to solve, these are the technologies that we need to solve that particular problem and go on from there. I actually managed to turn them into essentially procurement requirements, which then if you impose those on the contractor, there's a pretty good chance you'll meet international humanitarian law. It sounds quite simple. It's a long, hard chain.

To me, as an engineer it sounds incredibly difficult. There's the sort of thing that sounds superficially easy, perhaps, but you're taking legal requirements and approaching them as an engineer and turning them into formal engineering requirements. Those things don't usually work; courts make their money on resolving ambiguities in the law; not that courts necessarily make money. But that's what keeps them going. You are taking the Geneva Convention, which I am associated with things like you can't kill someone after they are surrendered, or you have to treat prisoners of war ethically, and many other things that I'm not thinking about at the moment. Are you engineering those to the extent that you can verify whether a machine is following them or not?

You can engineer them to say there is a requirement that the machine must be able to identify that someone is hostile. Now, there are a lot of subjective decisions to be made. But if you've actually brought the question down to something quite specific, then you can say what tools do they need to decide that? Or the most important thing that actually comes down is to say, you don't know; don't do it. That's one of the biggest things with particularly now the systems have moved on to being -- they call it autonomous, I actually have a personal distaste for the word because it's badly defined and it means all things to all men. But I'll use 'autonomous' because it's easy as a shorthand for something that goes away and does some thinking, and then makes a decision. The other thing that became actually a key for this is, military people to have a command-and-control chain that relies on authority being passed down the chain, and responsibility. I actually had to merge that with the systems that I was proposing for using unmanned aircraft and manned aircraft and that's a different addition to standard engineering processes, but you can bring it in; and that led to the concept of separating decisions and decision-making from action. Now normally, people say they've decided to do something and assume that the action then takes place. The thing that I brought in is to say you've made a decision; who reviews it, and who then authorizes the action. Now that can be a machine; a quite low level, you don't think about an ordinary control system, you just know it's deterministic, and it will do it. If it's non-deterministic, you can begin to query it and say, do we allow it to do this, and the other critical point, which is often forgotten in a lot of these things, is stopping it doing what it's not supposed to do. Again, putting engineering requirements on that helps with the ethics and the safety. I say helps with the law, rather than the ethics. I don't want to get into the trolley problem argument.

Right! That's fine. We've visited that plenty of times already. Now, you don't like the word autonomous? What's a better word?

Highly automated.

What's the distinction?

Highly automated, basically still implies human control. Autonomy implies no human control.

Isn't that a continuum, though? Is this the kind of process that's being executed in drone strikes right now when they're controlled remotely from someplace in Nevada, and they're operating over places in the Middle East? There's a lot that's written about this, about the processes that those pilots go through? Is what you've been talking about central to their challenges and the way they operate at the moment? Or is it applicable more to some kind of technology that we aren't operating yet, as far as I know?

Both is the answer to that. As I explained at the beginning, it grew out of unmanned aircraft and air control, and that's still very much what they're doing at the moment. In fact, a nice break between autonomy and highly automated might be that if you think about a drone going off to do whatever it's going to do, the actual control of the sensors and the weapon is under the control of the man in the Nevada desert, or wherever he is. The actual flight path the drone takes--he may be aware of it, but the drone actually probably plots its own flight path, flies it, and the pilot doesn't have to worry about it, unless there's some emergency. So, I call that autonomous flight; its control is highly automated, because you'll have sensors, you'll have intelligence coming in. But you're absolutely right. It is a continuum of fully autonomous at one end. I used to have a joke slide where you have the man arriving in the empty hangar, scratching his head and saying, I wonder what those autonomous aircraft are up to today. That's one end. The other end is pulling a lever and widening it with a bit of electronic string between the two and you've got the whole concept of autonomy levels.

Right.

Everybody knows the car autonomy levels or is aware of them. But there is something like 11 different types of autonomy level that I found when I was digging around, and they're a very specific particular application area.

Are those specific to weapon systems, or are you talking about general ones like the Society of Automotive Engineers standard for autonomous vehicles? I learned the number for that reference from reading your paper; it's standard J3016 that talks about those five levels. So, you talked about different levels of autonomy.

They are general, the space industry has one for NASA and ESA have one for the control of lunar landers and Mars landers where they've got a time difference of several minutes. So, they have a different set of definitions.

Which probably aren't impacted by legal decisions or liability.

Yeah. The Red Cross came up with a very smart idea back in 2012, or something. What they said is if you take a system of systems, and I think this is applicable to cars and every type of autonomous system with AI in it. You actually say we've got a whole system of systems, which is made up of systems and then subsystems. What you look at is the level of autonomy in each of the systems and subsystems. Then the chain of systems or command that is critical to safety, or release of weapons or with cars, with the safety of the car and all inside, you then say their critical functions, and you then have a very close look at the autonomy of them. You don't put a number on its autonomy level, but you can use that as a vehicle to say, what's it going to do on its own? What does it do if it doesn't know what to do or is being asked to do something stupid? So, that was quite a very good step forward to the Red Cross, which is interesting that that's a charitable organization, and it came out for a meeting of lawyers with some technical input. So, again, this is back to the point that if lawyers and engineers can get talking, this is one of the lessons after my light bulb moment, I then talked to the second and third and then on to the seemed-like-infinite time to some very smart lawyers who were also worrying about this. By working quite hard together, we could begin to agree on the classic problem I always put up, which is a lawyer says that's reasonable. The engineer says that's 95% certain.

Yes, that reminds me of my own experience into that kind of terminology when I was foreman on a jury, but we won't get into that. Now, I want to describe some scenarios and you have to tell me at what point the principles of what I'll just call AI governance at the moment come in at different levels. I could go to artillery pieces, long range weapons that are firing inert shells called a shot.

A ballistic shell.

Ballistic shell, right. So, it leaves the barrel, there is no control over it. But there is clearly an expectation that the people firing it knew where it was going to and so they can be held responsible for that. Now, in the case of an intercontinental ballistic missile, I don't believe those have abort functions; I could be wrong on that. But once they're fired, I think they're going to come down where they're supposed to. I think they actually left abort functions out of them because they didn't want them to be subverted. So, same kind of principle; when you launch it, you know where it's going to come down. If you're pushing the button, you're liable for that. But now there's an added element that perhaps the person turning the key doesn't know where it's going to come down, that decision may have been made by someone else. Now we get to things like cruise missiles, and the path they take to get to their destination is up to them to decide. But that's not part of a combat or an offensive role unless it collides with something it wasn't supposed to along the way, but still less liability. But now we get to the loitering weapons, the ones that are just told to go and hang around some area until they see something and then take action. So, what lines have we crossed along the way there?

I would say I put lines in to make sure you haven't crossed them. If we take the loitering munition as a good example, as you put it, it just hangs around until it sees something and hits it. Now, when you fire it, the question you've got to ask is--the person who's responsible for firing it, is he confident that there is likely to be a hostile target there that is legitimate? If he doesn't have good knowledge that there is a target that he may be hidden, but it's there, or the idea of the loitering munition might be to keep the target hidden, so, it doesn't come out of its hiding place. But unless he's got that knowledge, he is in breach of international law if he fires it. So, you can say it hangs around, then there are questions about how sure are you that it will identify a target correctly? Is he looking for a tank, or if it's looking for a SCUD, which is a big missile on the back of a lorry, how do you tell the difference between that and the petrol tanker which the skirt is hiding behind? That's where they have to decide the initiation of the strike left to the munitions or to somebody with control or that be an issue. The other thing you have to have is you have to have an abort mode that if it runs out of fuel, or the parachute is letting it go down slowly, it has to have a place which it can safely explode on or make itself non-lethal by disarming its fuse or something. So it has to have a way of coming down and doing nothing. Now the preference is to land in a safe area and explode because probably the last thing you want to happen is for your enemies to get hold of your latest gizmo and to be able to take it home and have a good look at everything that goes on in it. But there are balances there. But the principle is, you've got to know there's something there there's hostile (a good chance of it), and that you have got a safe place, if something goes wrong, you get into arguments about what's the last moment in which he releases it or makes a decision, which is where you get into real autonomous systems. The paper I put out in the Russo Journal actually says that you need to define a time when the commander says "Over to you." That may not be a weapon release time, you're got to define that, then you can define the information he needs and you also need to know what it's going to do after that time, and you don't let it learn anymore.

This is really interesting, I believe, and maybe you can correct me if you're familiar with this, that there was a drone that was captured in Afghanistan, and that they did that by hacking its idea of where its home base was so that it landed there instead of back where it was supposed to?

I think that was the Iranians who captured a new American one.

Yes.

I don't have any inside information, I've heard all sorts of rumors about what it did and so, I won't comment on how it happened. But yes, that was a good example and I think, dare I say, one up for the Iranian?

Yes, well, not easy, and should have been impossible. Now, just to visit a question of the command chain here, what you're telling me seems to imply that the operator, the person who's pushing the button, has to know the effect of the weapon. That can't be concealed from him? Is that a matter of law?

Yeah, it has to be under control and the commander has to know exactly what's going to happen. That's back to why I put this question in about the actual handover time, that he's got to be confident at handover time that he knows what it's going to--I'm saying "he," it's often "she"--knows what's going to happen and that's law. What I think people also don't always realize is that in the command chain, there is what they call the targeting cycle, which is not really a cycle, it's a chain that goes; on the closing of the loop is the report back saying this is what happened, you have to trace this all the way down from the high command, right the way down with people making decisions and imposing what they call the rules of engagement, which are the restrictions for that particular attack. And quite often people don't strike because the rule of engagement isn't met. And the opposition understands this. So, they also play tricks, which make sure that the rule of engagement isn't met for the attacker,

In your process of engineering these international humanitarian law, Geneva Convention; did you discover bugs in them? I mean, I can find bugs in the driving code, for instance, there are speed zones which have the sign at the beginning, but not at the end and my car doesn't know when it ends, you kind of assume that. Now, let me boil this down to a question. How should that type of law, that type of international law be adjusted to make it compatible with highly automated systems?

There's no straight answer. The United Nations have got a group of governmental experts looking at the problem of lethal autonomous weapon systems--LAWS, which can give confusion. That's under the convention that bans chemical weapons, biological weapons, landmines, and so on. I actually went twice in 2018, to act as a technical expert in the big forum in Geneva, where—we were careful to say it wasn't negotiations, it was the stage before negotiations; you had something like 85 nations. It was fascinating to see the problems that were coming out: there was mutual recognition that the law may not cover autonomous systems or autonomous weapons: what are we going to do about it? There was an agreement they weren't going to have definitions at that stage. Because once you get the definition, and nail it down, everybody looks for ways around it. But it came to a question of human control, and what you mean by meaningful human control. And so that went round and out ends up, I said, an individual person who releases the weapon has responsibility. Well as an engineer, I could see the problems they were having, and also, they began to move to say, well, if it's a machine learning system, you don't know what it's going to produce, necessarily, but under international law, you've always got to have somebody responsible. So they're actually looking to move the responsibility back into the supply chain. Because if you have the poor soldier sitting there, and he doesn't know what the thing is going to do, because it's going to go off and learn, then how can you hold him responsible?

It must surely be a primary goal of creating autonomous weapons systems to enable them to make better decisions than some grunt on the ground who's had two years of training. You could embody much more sophistication in a weapon. Surely it is a goal to be able to make better decisions in the machines. Is that right?

Yeah, and this is--when I said about separating decisions and action, you can have lots of intelligence in making the decision. I always divide it into decision making. You can have relatively slow time. very sophisticated processes, doing image processing, intelligence sifting, and so on, to come up and say, that's what you expect to see or what you don't expect to see. Then the person with one or two years training can actually be told, yes, this is as we expected to see, or they have quite good electronics in the battle pack now. So, he can see what he's expecting to see on his screen and compare that with what he's seeing or even have a comparison thing that says this is the difference. Again, one of the things--I don't know if it's been taken up or not--is that I always feel you need as well as you have object identification. Military people talk about target identification; I think it's just as critical to have non-target identification. In other words, something that comes out and says this is not a target. This is an ambulance.

That's a very good point.

It brings you to maybe one of the patterns you ought to recognize is the Red Cross on the white background because this drives technology, and an autonomous system should be primed to look for things it can't do, as well as things it can do.

I think that's a very important key insight there.

That's the end of part 1. We're breaking the interview up so we can keep the download under an hour for file size and attention span. That was a really interesting point about how autonomous systems should also decide what is not a target, which is the sort of insight that when you hear it from an engineer we think, "Well, of course," and yet – well, I don't know about you, but *I* hadn't thought of it before. The whole discussion about the role of AI within command-and-control decision loops is fascinating to me and we haven't touched on it before in this show.

Tony's book can be found here.

If you're enjoying these podcasts, think about how many other people might also enjoy them but don't know about them, and then think about how much fun we could have if we grew our community of listeners to the point of having live shows with audience interaction, maybe on Clubhouse, or did live classes, or had subtopic forums, higher-profile guests, more panels of multiple guests, summits or symposia… you see where this is going. I'm nothing if not ambitious. But does that sound like something you would enjoy?

And the only way we get there is by getting more listeners. It doesn't happen from the media fairy visiting in the middle of the night and waving her wand. And the only way we get more listeners is from you sharing about the show, posting links to episodes, giving us five stars and positive reviews. It just doesn't happen through taking out an ad in the *New York Times* Classifieds. So put the word out, please. I believe in the value of this show. In fact we've scored some scoops already. For instance, in 2020 there were many interviews in international media with Audrey Tang, the information minister of Taiwan, who had made newsworthy use of AI in combatting the coronavirus and also disinformation; yet very few of those interviews predated when we talked with her about all those topics in June 2020. And as long as

I'm blowing our trumpet here talking about scoops, there's been a lot of coverage recently of the release of the Moxie robot for children from the company Embodied. But those of you who've been listening to us since last September would have heard about Moxie then, when I interviewed Paulo Pirjanian, the CEO of Embodied. Anyway, that's enough of the self-promotion for now. Share and like.

In today's news ripped from the headlines about AI, a team of scientists at Aston University is going to be using human brain stem cells on microchips in an effort to "push the boundaries of artificial intelligence". It sounds a lot like a subplot in an obscure sci-fi paperback I once read called StarFire, about lab-grown brains, only without the dystopian paranormal conflict.

The Neu-ChiP project been awarded €3.5m (£3.06m) to show how neurons – the brain's information processors – can be harnessed to supercharge computers' ability to learn while dramatically cutting energy use. The human brain, after all, accomplishes more than we can do with a supercomputer the size of a building and all with a 20W power supply. It does not need a nearby river to cool it. That's rather more efficient. So the Neu-ChiP research team is embarking on a three-year study to demonstrate how human brain stem cells grown on a microchip can be taught to solve problems from data, laying the foundations for a "paradigm shift" in machine learning technology. They will layer networks of stem cells resembling the human cortex onto microchips, and then stimulate the cells bsy firing changing patterns of light beams at them. Sophisticated 3D computer modelling will allow them to observe any changes the cells undergo, to see how adaptable they are. This imitates the 'plasticity' of the human brain, which can rapidly adapt to new information. Professor of Mathematics David Saad said, "Our aim is to harness the unrivalled computing power of the human brain to dramatically increase the ability of computers to help us solve complex problems. We believe this project has the potential to break through current limitations of processing power and energy consumption to bring about a paradigm shift in machine learning technology."

Next week we'll conclude the interview with Tony Gillespie, when we'll talk more about issues with AI responsibility in autonomous military systems and also in autonomous vehicles, from a systems engineering perspective, which is really interesting in contrast to the way we've previously looked at them from the perspective of lawyers, activists, and philosophers on the show. 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