AI and You

Transcript

Guest: Peter Asaro, part 1

Episode 40

First Aired: Monday, March 22, 2021

Welcome to episode 40! It's really great to see our listener base increasing as you all spread the word. We're past 12,000 downloads now, so we must be doing something right.

Today's guest is Peter Asaro. He is a professor at the New School in New York, and is a philosopher of science, technology and media. His work examines artificial intelligence and robotics as a form of digital media, the ethical dimensions of algorithms and data, and the ways in which technology mediates social relations and shapes our experience of the world. He has focused heavily on the use of military robotics and unmanned aerial vehicles - otherwise known as drones – in other words, autonomous weapons. He's the co-founder and co-chair of the International Committee for Robot Arms Control, which means he is instrumental in work on the human rights issues surrounding targeted killing by drones, and arms control issues for autonomous lethal robotics.

You might be wondering just what that work entails and what sort of difference it's making, and we'll be talking about that in the interview. But you might also be wondering what sort of person gets into that kind of work, so first we'll be talking about the path that led Peter into being a champion of our need to examine our values around autonomous weapons. So let's get to the interview with Peter Asaro.

Peter Asaro, welcome to the show.

Thanks for having me. It's great to be here.

Your experience is really focused and intense on the area of autonomous lethal weapons in particular. I have not seen anyone with as much output and research on that area as yourself. And maybe you could tell us a little about how you got into that. What precipitated that? What were you doing before that?

Well, I got to kind of an early start on it. But my background is in philosophy and in computer science, and I was really doing my graduate PhD research on the history of cybernetics and early artificial intelligence and was really interested in the machines that were built that would simulate brains, in particular mechanical kinds or electrical neural networks in the 1940s, '50s, '60s. And from that, I kind of got interested in the embodiment of intelligence in robotics, in particular, interest around-- I did a little documentary on social and emotional robotics, thinking about, not just intelligence as a human trait, but what other kinds of human traits are we trying to simulate artificially. And that kind of got me into it, got me an invitation to a journal that was being put together, an issue on robot ethics. And I wasn't really sure what robot ethics was at the time so I was like, "Well, 'What Should We Want from a Robot Ethic?'" was the title of the paper and really kind of reflecting on what should we want. And then I realized quite quickly that I think one of the crucial questions in robot ethics was whether or when or how we would

ever allow robots to take a human life? And what would the conditions and criteria of that be? And what would make that morally acceptable or morally objectionable? And from that, I sort of went into thinking about it more in terms of just war theory and looking at military applications of robotics, and the ethics and morality of that. And then that kind of snowballed into founding the International Committee for Robot Arms Control, which is really calling for an international moratorium and ban on fully autonomous weapons systems, which then became part of a coalition of NGOs called the Campaign to Stop Killer Robots back in 2012 that has been campaigning at the United Nations for an international treaty to prohibit autonomous weapons systems. And I've been heavily involved in that since.

And we'll be talking a lot about those, I hope. I'd like to know what you think has precipitated the urgency here. I could go back to the '30s and Isaac Asimov and discussions then thinking about the ethics of robots that he did, among others. But it wasn't something that showed up on the radar of international geopolitics, which is where you're staging your efforts now. So what takes this or what has taken this out of the realm of academic theory into something that needs to be addressed by people acting in the moment?

I think there's a convergence of factors, all of which point towards real risks and significant sort of societal and global level risks. So on the one hand, you have the rapid development of military technologies, and the increasing use of autonomy, and what we might call machine learning or artificial intelligence at different levels and in very different kinds of applications across the military. But as we get into looking at weapons and the targeting of weapons, in particular, there's a lot of interest in moving from, say, remote-operated drones, where you have humans who have to do a lot of work to interpret imagery and decide what's a target and fire on it. And military is thinking, "Ha, we're just going to replace this with some software." And most of what you need for, say, a robotic fighter plane or something is sort of already there, technologically, it's just the decision to implement software in the decision-making processes of targeting and things and engaging targets with weapons. So the urgency sort of stems from that and the fact that really we're not very far from that and you can already do this. Arguably, it's being done in certain kinds of weapons systems for many years now. But you can also do it very easily or cheaply or if you don't care a lot about accuracy or precision or reliability or impacts on civilians or these sorts of things, it's very easy to automate these. It's very hard to get those systems to not make mistakes. And so there's a certain set of risks around how do we manage mistakes? How do we engineer these systems or test these systems or evaluate the systems, not just for our own army, or military, but for all of the militaries around the world who might implement these systems? There's a set of risks around that. And then if we look at the kind of recent history of cyber warfare and the way in which it's being deployed, I think you see a number of concerning features. So one is the sort of difficulty of attributing attacks, and the way in which that allows states to launch attacks that they might not otherwise do because they feel that they can get away with it, or there's some plausible deniability and some cover for it. We've seen this to a certain degree with the use of drones for targeted killings outside of traditionally recognized war zones, the idea that there's a lower political risk and a lower military risk for deploying force in those ways. And then you start to see, well, autonomous weapon systems are just going to exacerbate this. And they're going to put that capability into the hands of more and more militaries and

governments and give them the option of this kind of low military risk, low political risk use of force, which leads to more use of force and thus more conflict, so that's concerning. And then you start to look at the conflicts or the rivalries between great powers - the United States, Russia, China - who are all rapidly trying to develop these kinds of technologies. And then you start to realize, "Well, this is really the sort of new front of an arms race in which each of these states thinks they can get a significant military advantage by being the first one to develop the system or to develop the system the best, or at least not worried that they're going to be left behind and thus forced to sort of keep up and invest heavily." And so you have another kind of Cold War-style arms race in which all these countries start investing lots and lots of money into rapidly developing and fielding systems, which may not be perfectly safe, which may not be totally adequate. It's a waste of resources. And really, and especially in the last few years, what we've seen is the sort of breakdown of really significant treaties and arms control measures between the states and that's a sort of growing set of concerns in terms of international stability. And there are all sorts of ways in which basically transferring the decision authority for the use of force to these automated systems and removing humans from those decisions increases your political risks, increases the likelihood of conflicts, makes it much more difficult for the humans who do make decisions or have any control to really know what their adversaries might do, or what their systems are capable of, or even what their own systems are capable of. And all of that uncertainty leads to more risk, more conflict. And so the whole sort of picture of what it means to go down this road is bad in a multitude of ways and really should be controlled and be much more deliberate and controlled, basically - arms control.

And I'd like to dig down with more fine-grained discriminating tools at exactly what the risk is here, whether it's destabilization, whether it's use by non-state entities. There's an argument that technology has decreased death in war, that hundreds and thousands of years ago, people would throw 100,000 troops into a single battle to die there. Now, we send a drone, it takes out a power station and maybe hits a half a dozen bystanders. At least that's a common view that the impact of technology has reduced death, unnecessary death in particular. Now, we could argue that nuclear war increases the risk of a great many innocent bystanders dying, and it's an all-or-nothing kind of thing. Once one of those gets set off then that happens, but if it doesn't, then you don't get any. The use of automated weapons then, are we afraid of it exactly destabilizing what or creating what extra risk that we don't already have?

So there's a couple of different ways to think about that or approach that question. I guess in terms of technology, I mean, all sorts of technologies are deployed in warfare and that doesn't make them necessarily bad. New medical breakthroughs help troops who are injured in battle and things like that. So the question is really what kinds of technologies are we thinking about? And what's the motive for regulating those or banning certain kinds of technologies? And I think in terms of lowering the risks to civilians, there's I think the strongest argument so far has been made with regards to precision-guided munitions and these are laser or GPS targeted munitions. And the argument goes that instead of having to carpet bomb a huge area several times and drop a lot of bombs, we can just drop one bomb that we know is going to hit the target, and we can tell whether it hits the target. And so we reduce the civilian casualties or the collateral damage from

these other different bombs. So in a sort of, one-to-one comparison of you can drop one bomb, or you can carpet bomb, yeah, it seems like that's a technology that reduces civilian casualties. But when you look at the actual practice of targeting, and particularly the US use of precision-guided munitions in the first Gulf War in the '90s, versus the later Gulf War, what you see is what actually happens is you reduce the cost per target of bombing a target. So instead of having to drop 1000 bombs, we drop one expensive bomb. But that's not 1000 times more expensive, it's maybe 80 times more expensive. But that lets us then bomb many more targets. And so you see almost exact correlation in the growth of targeting lists. So the US military just bombs a lot more targets because it's cheaper to bomb each individual target. And then you start to ask, "Well, who's really doing the calculation of how much collateral damage and what were the actual civilian impacts of all of that?" And to actually do that in a controlled way is very difficult, of course, but there's still enormous civilian casualties that we're aware of, even when precision-guided munitions are used. And even when we've seen more around, say, targeted killing, where you're really trying to kill specific individuals, and yet, we know that there are many innocents who are killed and Pentagon does these things where any male over the age of 14 is designated a combatant unless there's evidence otherwise. And of course, usually, there's no evidence presented otherwise so they just make those assumptions. And so it's actually very difficult to establish the truth of that claim that precision-guided munitions, in fact, overall, reduce civilian casualties. And, again, there's a strong intuition that that's the case if you do the kind of one to one comparison, but we really should think about this in a much more comprehensive sense, and really do the kind of accounting and see is that in fact true and how true is that? So the similar argument gets made that autonomous weapons will be better than humans and more precise than humans. But of course, if they wind up creating more conflicts, having more engagements, even if they have lower error rates, if they are used more, if they engage more, if there are more wars, then you're still going to see more errors. And those errors translate into civilian casualties and impacts on civilian infrastructure, and things like that. So I think that's a genuine set of concerns, especially when we think like what this technology is doing is scaling things up, right? So it's an economies of scale that we can drop more bombs, we can attack more targets, that you can build a bigger army because now you can use robots instead of soldiers or what have you. And so that's the nature of automation. So that shouldn't surprise us that there would be more war, more conflict, more attacks per conflict if you're further automating, and increasing the efficiency of the war machine in general. And then I think you have these questions about genuine stability and you can look to nuclear arms as a kind of an odd case, where there's this huge risk of threat to civilization and to civilians, posed by a weapon, which actually kind of makes it useless in a military sense. You destroy the thing that you're-- Generally, the objective of a military operation is to take control of geographic territory, and the population centers and resources of that territory. Nuclear bombs just destroy those things - obliterates everything. So in a sense, you've got the territory, or at least you've eliminated whoever had the territory, but it doesn't-- If it's radioactive, and it's burning, and smoldering wreck, it's not a useful resource anymore. So it's questionable whether it's actually that useful as a weapon. It's useful as a threat, that your civilian populations of your country will be devastated unless you capitulate politically or militarily to some demands of the person who holds the weapon and is threatening you with it. And so it works very effectively as a threat, but it's much less clear whether it's actually

effective as a military weapon. But what's interesting, I think, about autonomous weapons, and there's been some discussion about whether there would be a new class of weapon of mass destruction. So if we think about the traditional weapon of mass destruction as a nuclear bomb, which just destroys everything, and it's highly indiscriminate in its destructive power. You can target it, but it has such a huge area effect, and all these carry-on effects to the environment, and health and things. And we put also chemical weapons and biological weapons, which again, have these kind of mass effects that are difficult to control. But I think if you think about it from more of a political-military perspective, the idea of a weapon of mass destruction is a weapon in which an individual or small group could deploy a weapon or make an attack that has mass casualty effects. And in that regard, it would not be impossible to conceive of a small group of people launching a fleet or a swarm of drones that would go into a village and kill everybody, but they could do it in a much more precise way. So you could actually decide, "Oh, just kill all the members of this political party or this ethnic group within the city," and program some parameters into it to try to achieve that, and then deploy that. And then, that kind of threat that that poses is severe. And because it's even more discriminate than nuclear bombs, I mean, that doesn't mean that it's morally better. It's better, of course, not to indiscriminately kill, but that alone, the fact that you're not indiscriminate doesn't mean it's moral. So it becomes now a recipe for genocide and things like that, or for, as you said before, non-state actors could acquire these; could use these for terrorist attacks, and very specific ones. And that's, I think, a concern, as well as the potential for say, assassinations or things like that. So then, I guess the stability question is really then, to what extent does the existence of these kinds of weapons and their proliferation to numerous states and to non-state actors make it easier for people to go to war? Because it's really kind of lowering the cost of raising armies and deploying lethal force and that's very concerning.

And so could we use the technology in ways to improve - I hate to say - the way we make war? Because no one really likes to do that. But if we accept that there are cases where that's going to happen anyway, and in, for instance, remotely guided drone attacks, there are instances of people being caught up in the fog of war and everyone involved in a decision says, "Yes, it's a legitimate target," and then it gets hit, then discover it's the wrong one. And they go back over it, and everyone realizes, "Well, we should have seen this all along. It was obvious except that we all conned ourselves into thinking the same thing." It's a recognized phenomenon in the stress of war, which AI could be immune to. Do you think that it's worth pursuing the use of AI in those situations to reduce collateral damage, assuming that there are entities, that there are actors that would want to use it for that purpose?

Yeah. I mean, I think I've always been an advocate for computer-assisted or the human-machine hybrid doing the decision making and doing the analysis, and that automating certain aspects of the process can enhance the capability of humans to make those decisions even in conflicts, even in stressful, high-pressure situations, and things like that. And I mean, there's a vast literature, of course, in human factors and human-computer interface design about how to design cockpits for planes or all sorts of things to enhance situational awareness of the human operators and studying those cases where it fails. Airline disasters and the USS Vincennes shooting down the Iranian airliner in the 1980s - there's been a lot of studies of that. And as a sociologist of science

and technology, it's always very interesting to see how the narratives are assessed in terms of blaming technology versus blaming humans. "Is this human error or is this a machine error?" And then some sense, if the machines doing what it's supposed to do but the human doesn't expect it to do that or understand that it's doing that, then we just sort of always chalk that up to human error, when in fact, I think in many cases, it's a design error, that the system wasn't designed to inform the human properly in those situations. They didn't anticipate that the human would have other kinds of expectations that the system didn't clarify. And I think there's a lot of work to be done to improve those sorts of systems and to utilize data, information, intelligence analysis, machine learning, and AI to provide human decision-makers with better information, more up to date, and in fact, warn them when they're going to do something that would be dangerous or harmful. I think one aspect that makes this a little tricky is that if we think about a safety mechanism, like a gun that wouldn't fire if it recognizes that it's being pointed at a civilian, I think that would be a great technology. But you have to have the confidence in that case that the human isn't allowed to override it for some other reason. That you're really going to trust the machine's judgment ultimately for that. And then people would say the reverse that if there's a missile coming at you, do you want to shoot it down and the machine doesn't let you, then the human should be able to sort of override that using their judgment. And so who do we ultimately give the authority to, I think, is a significant question. And I don't know that you would just want to base it on performance of a particular system unless you've really thoroughly evaluated that system. And in general, the bigger problem is, whenever you take humans out of that role in terms of their authority and their ultimate decision making, you also eliminate their responsibility from a moral and legal perspective. It becomes the system's fault. And from the operator perspective, they blame the system, they don't feel that they are in control. And so they let the system make decisions. And whenever you're in that sort of situation, you're losing human responsibility, both from the sort of proactive kind of psychology of those operators of doing their full due diligence and fulfilling their duty to the best of their ability, given the uncertainty and stress and constraints, of course. But that's also retrospectively and legally that we can't really hold these people accountable for, say, war crimes or for these mistakes. And in fact, maybe we shouldn't if they're not in control of an autonomous system that's off doing something, that they're not really aware of how it operates, how it's going to interact in a given environment. They released it, so we sort of want to hold them accountable, but they don't specifically intend any of the things that that system does once it's released. And that's a very dangerous place to be. Now if we fully trust these systems, maybe it's okay in some metaphysical sense but we don't know. And so you're basically saying you're not responsible for releasing something even though you don't know what it's going to do. And that seems reckless, right? And in war, we can't really hold people accountable for that kind of recklessness, generally speaking. In terms of civilian applications, it's different. In the civilian case, of course, we have liability laws. So if somebody manufactures a car or device that harms a lot of people, then there are lawsuits, and they have to pay those people. But that doesn't really happen in international armed conflict. Somebody drops a bomb on your house and kills your family, you don't really have any legal recourse to the army, the military that did that, or the state that did that, much less to the manufacturer or the engineers of those systems, even if you were somehow able to show that those systems were responsible for mistakenly dropping a bomb on your house.

And I think you've hit on something that's novel, for me, at least, and has me thinking, which is around this idea that the controller, the operator may not feel as much agency when the machine they're operating has this degree of autonomy. Now, as a parallel with say, self-driving cars, it's clear at the moment that despite what people might like to argue about on talk shows, that the driver has the responsibility. If my car speeds through a school zone, if it hits something else, my claiming that it was on Autodrive at the time and [so] it wasn't my fault will make no difference whatsoever. It won't hold up the policeman, it won't hold up the court for a second. My, as you say, only recourse might be to launch a suit against the manufacturer, but good luck with that one, they've got more lawyers. And in the case of an operator of an autonomous weapon, the chain of command and responsibility might be completely clear but I think it's the psychological aspect that is the more important one because far from decreasing the fog of war, it could increase it in the sense that the operator temporarily feels like the way a driver of a Level 3 autonomy vehicle might feel, that "This thing is doing just fine, I can relax," except when it gets to the point where it doesn't know what to do and it says, "Help!" and suddenly, you've got a split second to recover situational awareness. The same kind of thing could happen to someone operating a semi-autonomous drone. Did I capture your concern there?

Yeah. It's really a set of concerns, and I think it all focuses on the complexity of the interaction between the human and the machine. And if we look at the self-driving cars case, in some sense of if all the cars were self-driving, and we could just regulate them from some central command sort of system, it would be a much easier engineering problem. If people weren't allowed to wander onto the street and children's balls didn't roll out into the street and things like that, it would be much easier. It's hard to do self-driving cars because there are all those unexpected things. And other drivers are very unpredictable in their behavior. And then there's vision things which we've seen with the Tesla autopilot and things like that. The hard part then becomes the sort of trade-off. So you have a car that can drive itself some of the time, and they have a steering wheel, and you have a driver - a human - sitting in front of it. And now it's this question of the handoff. So when does the car decide that it can no longer control the vehicle and wants the human to take over? And how much time does the human get to regain situational awareness and things like that? Versus when should the car automatically brake or swerve because the human driver is failing to do that? The car's systems recognize that a collision is imminent unless those actions are taken. So we kind of want those systems in place, and now they're pretty widespread for breaking. Like if you just are approaching something too quickly, the brakes will just engage on some of these fancy cars today. And that sort of assist seems helpful and well constrained. When you get to the point where you take the steering wheel out of the car, then it's clear to the passengers in the car that they're not in control anymore at all, and they have no responsibility. And you have a very clear situation where this is a fully autonomous system and where you're just going to trust it to take you where you've told it to take you and you don't really have much say in that. Maybe you can tell it to take the scenic route, or not to drive too fast or to stop by your friend's house or something along the way. But you're not, in some sense, controlling the low-level nature of it.

That's a very good point.

That's the end of part one of the interview, and you probably know by now that we'll be airing the rest of it next week. I think it's really easy to think that campaigning against lethal autonomous weapons is futile, that our leaders are not going to limit our abilities in that area, and wouldn't that mean that Peter is wasting his time? So I was really curious to see what motivated Peter to engage this challenge with such commitment, and we'll certainly be continuing to flesh out that question next week. As you've heard already, there are a lot of layers to this question that aren't obvious at first glance. Perhaps we'll end up with some kind of strategic arms limitation treaty like we have with nuclear, chemical and biological weapons. I really admire Peter and the others who diligently pursue this issue, because it shows a deep commitment to his values.

In our news ripped from the headlines about AI, a recent study published in the Proceedings of the National Academy of Sciences showed how AI can learn to identify vulnerabilities in human habits and behaviors and use them to influence human decision-making. In one sense that may not seem like news, and in another it may sound horrifying. After all, you may have already heard about the use of AI in influencing what people see on Facebook, and driving bots that plant ideas to change people's minds, something we used to call propaganda and now call fake news. So maybe it seems as though this study isn't telling us anything we didn't already know. But there's a difference between suspecting something and proving it, measuring it, and so these researchers set out to do that. They used a neural network in a controlled experiment where people were clicking on red or blue colored boxes to win a fake currency, with the AI learning the participant's choice patterns and guiding them towards a specific choice. The AI was successful about 70 percent of the time. You may wonder how the AI guided the choices, and it did that by showing a smiley face or a sad face to the participant.

In the second experiment, the subjects watched a screen and pressed a button when they were shown a particular symbol (such as a blue circle) but not when they were shown a different one (say a red square). The AI learned to arrange the sequence of symbols so the participants made more mistakes, and achieved an increase of almost 25 percent.

The third experiment consisted of several rounds in which a subject was in the role of an investor giving money to the AI. The AI would then return an amount of money to the participant, who would then decide how much to invest in the next round. This game was played in two different modes: in one the AI was out to maximize how much money it ended up with, and in the other the AI aimed for a fair distribution of money between itself and the human investor. The AI was highly successful each time.

As the researchers said in their paper, "Understanding the vulnerabilities of human choice processes allows us to detect and potentially avoid adversarial attacks." So this kind of formal experiment can help us get more knowledge about how humans are vulnerable to being manipulated by AI and maybe prepare ourselves better.

Spread the word about this podcast! The more listeners we get, the more we can do, such as building a community where we can come together for live events, for instance, and have debates, panels, and you can interact with other people who also are interested in what we talk about here.

Next week we'll conclude the interview with Peter Asaro.

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