**EP5 Transcript**

**CRAIG:** [00:00:08] Hi this is Craig Smith with a new podcast about artificial intelligence.

**CRAIG:** [00:00:14] I'm a former New York Times correspondent and I've been talking to people who are making a difference in AI. I'm bringing the most interesting of those conversations to you. This week I speak to Miles Brundage, AI policy research fellow at the Future of Humanity Institute, which was founded by the futurist Nick Bostrom. Miles is focused on governance of AI, developing policy responses to AI advancements. We talked about the challenges of regulating AI as the technology progresses. I hope you find the conversation as interesting as I did.

**CRAIG:** One of the things that I wanted to ask you about is the whole security issue that's moving front and center as companies are commercializing A.I. implementations and putting them into the public domain. Is that something that you work on.

**MILES:** [00:01:32] Yes. So, I don't work on it from a technical perspective but I look at the sort of broader context and what it means for decision making. So, generally, even without someone trying to mess with your A.I. system, AI systems are already pretty fragile today - and the problem gets even worse if someone's trying to mess with the inputs to the system and confuse it or poison the data that's being used to train your system. So, in addition to being an interesting and important technical question there's also the question – and we had a debate at this conference earlier about this topic - of what does that mean for whether we should apply AI systems in the real world particularly in safety critical applications when there is this potential for the systems to be subverted.

**CRAIG:** [00:02:16] I haven't heard much public debate about the adversarial attacks on vision systems that are pretty important for self-driving cars. There is a famous one that we were talking about earlier of stickers on a stop sign that causes the vision system to read the sign as a 45 mile an hour zone sign as opposed to a stop sign. That's obviously pretty frightening. How do they contemplate putting self-driving cars into the public with those kinds of issues?

**MILES:** [00:02:49] Yeah, I mean, there's a lot of research in this area and that's a moving target. Both the attacks and the defenses are getting stronger and it's hard to sort of characterize what the situation is. But, generally, there isn't a total solution to the problem and you know there's some ways in which you might be reassured, in particular applications, that there's a solution, like stop signs. I think it's pretty reasonably likely that one could come up with a bespoke solution to that particular problem or you might benefit from the fact that you have multiple sensors so it's not just, you know, a still image that's confusing the sort of a moving camera and you can look at the same thing from multiple perspectives and also cross-reference with LIDAR or another sensor sources. So, I think you know on the one hand you can think of solutions these particular problems but the more general phenomenon of AI systems not being very robust is very much the case.

**CRAIG:** [00:03:41] Yeah, the designers of the systems can never anticipate these attacks. That's the point. It's the attackers that are always going to find a place to exploit a system. So is there talk about standards or regulation or testing of A.I. systems before they're deployed, some sort of regulation at the governmental level.

**MILES:** [00:04:03] So first of all in terms of anticipating all the possible failure modes, I think it's not totally dissimilar from digital systems generally. So, there are all sorts of ways in which systems can be hacked or backdoors put in software, and there've been various efforts to find solutions to those problems and it's sort of an ongoing battle for computer security generally, even without A.I. In the case of A.I., there is some discussion of how can we have confidence in these systems and, if we have good solutions to these problems, or at least partial solutions to these problems, whether it's through testing or whether it's through formally verifying the properties of the systems, then is there a good way to require that. And, so far, it's sort of a piecemeal, or you know, sort of a patchy regulatory situation with different states within the U.S. and different countries having different approaches.

**MILES:** [00:04:55] But you know, both in the autonomous car case as well as medical A.I. systems, so far, it's sort of multiple conversations going on. But there isn't a general sort of A.I. regulation that you know even at a national scale let alone at a global scale that would address all these issues. That might in fact be premature given that we're still, you know, in a fast-moving space.

**CRAIG:** [00:05:17] Has there been any regulation of AI systems proposed or passed.

**MILES:** [00:05:23] I think you could consider the state level autonomous car policies as AI regulation in a particular domain. Similarly, you could think of the approval process of say medical A.I. systems as, you know, some form of A.I. regulation. But it's fairly domain specific. You could also think of GDPR - the general data protection regime - in Europe as having some A.I. relevant properties of accountability for how certain algorithmic decisions are being made and the ability to get your data deleted or have an explanation of certain decisions that critically affect you.

**MILES:** [00:06:00] I think there are all sorts of different - what a colleague of mine, Joanna Bryson, and I call - de facto AI policies. So, it's not always called a policy per se, but it's sort of different jurisdictions, different domains are developing different policies that apply to A.I. But it's not an explicit, you know, global A.I. policy.

**CRAIG:** [00:06:18] You've also done work on how to prevent unethical or dangerous uses of A.I. How do you prevent it?

**MILES:** [00:06:27] There's no easy solution. So first of all, I'd distinguish between misuse of A.I. or unethical uses of A.I. and then there are a lot of things within that bucket that aren't necessarily intentionally malicious.

**MILES:** [00:06:40] So there are things like biased or unfair or insufficiently transparent or just, you know, otherwise harmful A.I. systems that are designed not explicitly to harm those people that it's affecting, but due to carelessness or lack of oversight or other sorts of reasons. And then there is a separate category, which is what my research has been on recently, which I call malicious use of A.I. So that sort intending to cause harm to a certain individual or group of people and that might be through generating fake media order to cause confusion or it might be an automated spearphishing system. Some type of hacking system that instead of the human doing all the labor of crafting a phishing email and figuring out who the right person to target in a particular organization is, some of that task is automated; or a drone that homes in on particular people's faces and drops a grenade. So, these are all extensions of existing things that we see in society that are already happening, so there already are drones that are dropping grenades on people but they're remotely piloted by a human. And, similarly, there's a lot of hacking going on. There's a lot of fake news going on. So, you know we argued in this recent report that some of these sorts of malicious uses of information technology and other forms of technologies could be scaled up and automated in various ways with A.I.

**MILES:** [00:07:58] How do you prevent that? There is no silver bullet but I think a combination of developing certain technical measures - for example, with hacking. So an automated hacking system in some cases you can fight A.I. with A.I. In other cases, it's beefing up existing forms of defenses against sort of the non-A.I. versions of these problems. So, in a world without A.I. enabled hacking you already have to worry about some form of hacking and you might want to have two-factor authentication for example to make people's emails and other accounts more secure. Similar sorts of things would be applicable to A.I. enabled attack. So, we should probably double down on those things that we should be doing anyway and we just aren't doing.

**MILES:** There are also things around the optimal degree of openness around A.I. So, there's a strong push in the AI community to publish and publish code generally make it easy to reproduce the results of your research. But in some cases that could be harmful or go too far if the particular thing that's being published is especially ripe for abuse. So, this is something that has been negotiated and discussed a lot in the cybersecurity community and there are norms around responsible disclosure where, say, you notify potentially affected party before you just publish something or you only release certain information that doesn't make it too easy to go ahead and exploit that and use it for malicious purposes. So, I think a combination of, you know, developing new technical measures possibly some regulations and responsibility by the AI research community itself are, you know, all part of the solution. But there isn't one silver bullet.

**CRAIG:** [00:09:30] Although a lot of these malicious use cases would primarily be by state actors. So, there would have to be international conventions that everyone would have to sign on to. Are you at all hopeful that that would happen? I mean just the case of fake news as a tool of propaganda or subverting the democratic process, it's pretty clear that Russia has embraced that and I can't imagine that if they were party to an international convention condemning it.

**MILES:** So first of all, I'd say that, you know, there are a variety of things we might worry about both in terms of unintentional misuse of AI and intentional ones and then, you know, in the intentional bucket there are state actors and non-state actors. And, you know in the case of ISIS for example and Hezbollah, they are using drones already to carry out certain attacks. And, you know, there are all sorts of rogue actors in the world who carry out mass shootings and so forth. So, I think we have things to worry about even before we think about state actors.

**MILES:** But you know in the case of state actors I agree that there's going to be difficulty getting an agreement on the full range of these things. In the case of autonomous weapons for example there's a lot of active discussion at the level of the U.N. between Russia and the US and China and various other countries on what would it mean to have meaningful human control of lethal decision making in combat. And is there a way to come to some agreement to ensure that that control stays into the future - that we don't have autonomous killing machines. And it's difficult. I think that's a particularly thorny political problem. But I think it's not all or nothing. It's not just either we have a global, you know, perfectly enforced ban on all possible malicious uses of A.I. or we have no governance at all.

**MILES:** [00:11:15] I think there is also a middle ground, say, agreements between some countries or an imperfectly enforced agreement or some rules of the road and norms that are developed through dialogue and so forth on particular areas like certain ways of ensuring accountability of weapons systems in certain areas, or dealing with or cooperating to prevent domestically oriented terrorist attacks leveraging these technologies.

**MILES:** [00:11:41] I was actually just at a conference recently in Singapore organized by Interpol and the UN where there were people from a bunch of different countries talking about how they're using A.I. in law enforcement and trying to stop you know murders and rapes and other crimes within their countries. So, I think there is common ground in some of these cases. So that was a case where they're talking about using A.I. to prevent malicious things and I don't see any reason why they couldn't also have talked about using A.I. to prevent misuse of A.I. But that's sort of one case where there might be common ground, because the US doesn't mind if Russia stops a murder within Russia or vice versa. But when it gets into technologies that are potentially going to be used in conflict between these countries then it's a difficult situation. So, I, so I don't see it as all or nothing.

**CRAIG:** [00:12:32] I would imagine there will be international conventions over, you know, killing machines and that sort of thing. But I would also imagine that each country will develop the capability before they sign the convention. I mean this has happened with nuclear arms. And then there's always the threat – as you said - of a non-state actor getting hold of the technology. I mean one of the disturbing things about this technology is as complex as it is, it's relatively easy once you understand the math and the algorithms to implement. It's not like spinning up uranium isotopes or something.

**CRAIG:** There has been a lot of talk about national strategies being published in the last year or so and an emerging quote unquote arms race between state actors. In talking to the research community, I don't feel that because there is a lot of transparency in the scientific community, things are published. There's not a sense that anyone's hiding developments.

**CRAIG:** [00:13:35] Are those national strategies significant or is that really a matter of an economic choice by these countries to ensure that they have a cadre of capable scientists so that they're in the game commercially.

**MILES:** [00:13:52] I think a lot of the talk about an arms race is somewhat overblown or at least conflating multiple things. So, there is AI as a driver of economic growth or AI as a tool for ensuring domestic political stability as we can see in some cases in China for example or AI for particular military applications or more generally AI for a range of military applications so I think there's a lot of things happening simultaneously in different countries that have different emphases across these areas and some national strategies are more significant than others. Roughly the way I see it is consistent with your point that there is a lot of openness and this is not analogous at all to what was happening in the early stages of the Cold War where there was very secretive development projects for developing nuclear weapons and you're not publishing anything.

**MILES:** There are hundreds of countries represented at major conferences you know generally being open about what they're working on. But I think it's a complex situation like that academic world of relative openness is happening alongside, you know, specific applications of A.I. for military purposes, sort of the development of, you know, decades long strategic plans by militaries for applying a bunch of different areas. So, it's a complex picture and I think calling it an arms race is totally overblown. On the other hand, I think you know the truth is somewhere in between.

**CRAIG:** [00:15:16] On the question of transparency one of the complaints I've heard here in this is not with regards to state actors but with corporate actors.

**CRAIG:** [00:15:25] There are a lot of papers published where the source code is withheld and so the reviewers have to judge the paper basically on faith and there's kind of a growing complaint in the scientific community.

**MILES:** [00:15:41] I think there are a couple things going on there. An example of two things that make it sometimes hard to replicate results particularly from corporations that have a lot of resources at their disposal including computing resources is that these can be very large-scale experiments that are tried in a million different variations and it's hard to figure out what is the main driver of performance here. Is it just that they're throwing a lot of computing power at it.

**MILES:** [00:16:06] Another problem is that often the research underlying a particular paper is intertwined with the larger infrastructure of the organization. So, it's hard to sort of publish the code for one thing without also revealing some proprietary information.

**MILES:** [00:16:19] So, there are reasons why these things are happening that are you know structural at that level. But I think there is merit to the idea that in order to have a science be something that can be incrementally built upon, it's important to know what exactly is going on here. Is this a new algorithm that's actually useful or is that intentionally or unintentionally deceptive about what the improvement is? So, I think there's a lot going on in the debate and the underlying structural factors but I don't think it's inevitable that there will be efforts to sort of close down research just because corporations are involved, with those limited exceptions that don't, I think, apply to all research.

**MILES:** So like AlphaGo as an example. Not only was it extremely computing intensive but also it built on larger infrastructure. So, it's kind of hard to just totally open source the whole thing. But the same organization, DeepMind and you know in similar cases Facebook and others, there are a bunch of companies that are still publishing code and publishing enough information in their papers to be able to reproduce it. So, I wouldn't say that corporations are you know causing there to be a lot of openness but there are some structural factors that are causing things be more open than they would otherwise be.

**MILES:** So, for example it's hard to hire people these days. So, the AI talent market is pretty scarce. So, it's very much a seller's market in terms of selling labor. So, companies have to compete to get the best talent and they have to sort of allow people to publish in order to get them in the door in the first place. So, I think as long as that's the case and if, you know, tomorrow there were 10 million new AI researchers and then maybe we'd see a closing up of research as companies can just hire anyone they want, they can decide to get them to do research and never publish it. But that's not the situation we're in today. Right now, researchers want to publish.

**CRAIG:** [00:18:07] At the Future of Humanity Institute, what specific things are you working on?

**MILES:** [00:18:10] Yes, the malicious use of A.I. stuff that I mentioned before is a big part of it and also more generally trying to figure out where this international competition vs. cooperation is going to play out of the long run. And what's a more compelling vision than an arms race. And how do we get there.

**CRAIG:** [00:18:37] Just on that last point China's the one so far as pouring the most money into their national strategy from what I can see. They have this goal of being a leader in a by 2030. To they have the thrust to take the lead or do you think that the US, despite an underfunded government plan, because of the company interests, the industry here, that the US will retain its dominance.

**MILES:** [00:18:54] There's a lot of things going on and it's not just the US and China. I mean Canada for example has both been a strong player in AI for a long time and, in particular, has been a recent beneficiary of the weaknesses of the US, particularly on immigration.

**MILES:** [00:19:09] So lots of Iranian and other researchers and people who are not happy with the political situation in the US are going to Canada.

**MILES:** [00:19:15] So you know there are other actors, but generally in terms of the US and China, it's clear which directions they're going. You know China's on the upswing in the US, at least in relative terms, you know, there continues to be a lot of investment in the US, but it doesn't have the same privileged position that it had five or 10 years ago in terms of the overall hierarchy. It's really hard to say how quickly things are changing and will change. It's much easier to invest a lot in particular engineering projects with well-defined goals and clearly China doing very well on those things but they're also starting to move into the more creative science side of things and I think that's, you know, harder but also potentially very important.

**MILES:** I think they're aware that they still have those weaknesses, that a lot of the most interesting and cutting-edge research is all coming out of countries like the US, UK and Canada. But you know China's trying to close that gap and it's an interesting question how quickly that can be done or whether that's something that's actually amendable to just throwing a bunch of money at it.

**CRAIG:** [00:20:14] Are you optimistic about being able to control the development of A.I. in very positive ways. I mean certainly Bostrom’s book painted some pretty terrifying scenarios and presumably that's what the institute is all about, is figuring out ways to avoid those scenarios. Are you optimistic that we'll be able to corral AI into a ...

**MILES:** [00:20:37] I try to focus on what can be done constructively and not say, you know, what, whether I'm optimistic or pessimistic. But I did write a paper recently where I painted the view of conditional optimism. So, if we get the governance side of things right then AI can be extremely useful and have all sorts of benefits for society. So, everything from enabling a true leisure society and having widespread distribution of high living standards, high quality of life, I think is something that A.I. makes much more plausible than not having the ability to automate various tasks. Similarly, you know health research and biomedical research and various forms of scientific progress could be accelerated with the AI. So, I think, you know, if we can get the governance side of things right and solve a lot of these challenges around safety and security and avoid this destructive arms-race narrative and you know potential in the future reality, then I think we can get very good things out of there.

**CRAIG:** [00:21:36] For those of you who want to go into greater depth about the things we talked about today you can find a transcript of the show in the program notes along with a link to our [new newsletters](https://www.eye-on.ai/subscribe-page/).

**CRAIG:** [00:21:47] Let us know whether you find the podcast interesting or useful and whether you have any suggestions about how we can improve. The singularity you may not be near, but A.I. is about to change your world. Pay attention.