Edmon Begoli
Amir Sadovnik
Oak Ridge National Laboratory
Center for AI Security Research

Defense Writers Group
Project for Media and National Security
George Washington School of Media and Public Affairs

9 April 2024

Moderator: It's great to have you here.

Normally I ask the first question. Because their Center for AI Security Research is so new, I'd like for them to kind of sketch a little bit about what it is and what they're doing, and then three people emailed in advance to get on the questioner's list. We have an hour. There will be time for everybody so no need to throw shoes or anything.

So gentlemen, tell us about what the heck you're doing.

Mr. Begoli: Thank you. My name is Edmon Begoli, and I'm a founding director of the Center for AI Security Research.

We really started forming the Center the beginning of last year. The whole motion was triggered by multiple motivators. One was, from my point of view, very personal as being a researcher in this field, and really I got myself attracted to AI in 1989 when I was in high school and studied logic and psychology and math.

The reason why we formed the Center was multilateral. One, I was a principal investigator at a project for Department of Veterans Affairs and we were building large machine learning models at the time. One of the ways to build a machine learning model, or the primary way, is just to develop it in a large [GARBLED], we have the largest veterans bio-bank at Oak Ridge National Lab, and the question was like okay, once you develop this model, for instance, how safe it is to export and let others use it? And in those days nobody could answer the question properly. But in reality, one can take a machine learning model and can invert [trial data] out of it and compromise privacy. So we couldn't do it. But in those days there was nobody to answer those questions.

Another thing was, just to give you a little phenomenon. If you are training data, for instance, for cancer research, and we were doing this for cancer research but also for suicide prevention, the way how data is structured can influence how this model operates or what's today called data poisoning. One can poison the data in such a way that the model can misperform and cause all kinds of catastrophic, have all kinds of catastrophic effects later.

But again, this was 2017. Very few people, maybe two or three, were thinking about it at the time. By two years later, this became a much more pressing issue. But after development of AI we observed that AI is being so massively deployed in all kinds of political systems.

An important thing to say about the lab, I mean this really is you can say idealistic. I always say I'm an idealist. This whole legacy of Manhattan Project and all that -- people come to work at the lab typically stay for their whole life. We are working on serious problems that impact humanity and impact national security and our lab nurtures that kind of thinking. So we decided to form the Center to address the problems that I would say at this point nobody is truly systematically looking at them. Nobody is looking into the problem of what kind of harm, threat or risk AI can pose. At the same time, what are the ways to protect AI systems from exploitation? Because we are putting them left and right as humankind and national security system and all that, we are putting AI systems everywhere. We are possibly leaving a gaping hole for what kind of exploitations are possible against those systems.

So as enthusiasts, researchers, we had, like I said, this intellectual means, or we had a lot of support to do that we start the Center and we can tell you everything you want to know in terms of research we're conducting.

I'll just finish with this because my colleague Amir, we work very closely. I'd like for him to say a few words about things.

But things, if you're looking at things from a very practical spectrum and that is protecting AI systems and understanding what are the threats to AI in almost like a cyber-like manner that one can exploit to make it misfunction and miss malware or do some things like deep fakes and all that, all the way to the existential threat to humanity. Because today there are all kinds of articles you'll see that AI poses existential threat. AI can end humanity. And to be honest with you, I don't think nobody's very seriously, systematic looking. Is it true? What does it mean? What are the possible impacts? Is it hype or is it the real thing that AI can pose existential threat? And we are engaging in this discussion that really involves all, starting from hard sciences to philosophy and psychology and neuroscience, like what would that mean? Let's quantify the existential threat.

And this goes back to the Manhattan complex because you know the history, and [GARBLED] still are seeing people bumping into each other thinking, and if you saw Heisenberg's lecture, thinking what if Germans are working on a nuclear weapon? Because it looks like nuclear weapon can be created out of this. You're kind of in this moment trying to -- we're talking about AI as a thing that can end humanity and I don't think that anybody is taking it any step further to really trying to understand what does it mean that AI can end humanity? I mean we can all thing about all kinds of scenarios or something, but our mission is to look into those problems but also look into problems of today . That is if you are deploying cyber defense system with an AI model inside, can this thing be exploited in such a way that it can be penetrated [GARBLED]. The Center was created to work on those problems.

Moderator: Will AI kill us all? That's a pretty picture. [Laughter].

Quickly before Amir speaks, my favorite moment in the film Oppenheimer, which is great, is when somebody asks him, there’s some science that says if you set this thing off the chain reaction will burn the entire atmosphere. What do you think about that? He says, I don’t know, 50/50.

But that’s what you guys are doing. Like if we push this button, does life as we know it end?

On that note, Amir.

Mr. Sadovnik: I wanted to say a few words before we dive into kind of national security and talk about it. I just wanted to set the stage a little bit about AI to kind of get us on the same page.

People say AI, what do they mean? Well, AI these days, I’ll be a little “scienty” here for a second. But when people say AI these days really what they mean is what’s called deep learning, which is like an AI algorithm. And really all of kind of modern AI -- when you think of large language model, self-driving cars, image generators -- they’re all actually based on the same thing which are deep neural networks that operate in a very similar way to achieve different goals. Language generating, identifying faces, identifying pedestrians, or generating images.

So the question we ask is what is special about the deep learning algorithm that it produces kind of new threats both to the systems they operate in and to us as a society?

So we think of kind of four main dimensions. I’ll just mention them.

One dimension is the fact that they’re reliant on data. They’re reliant on a lot of data. What that means is that, Edmon kind of alluded to it. That introduces certain threats. For example, you need so much data sometimes that you don’t even own the data. The fact that you don’t own the data means that other people can mess with your data in a way that you won’t know, right?

There was a paper recently that showed that by buying -- for example, a lot of language models don’t own, they don’t train on their own data, they train on web sites. If you buy expired web sites -- a paper just showed that if you buy expired web sites and basically put whatever you want up there, you can poison the language models to say things they’re not supposed to say. Even though you’re buying like 0.1 percent or 0.01 percent of the actual amount of data you’re training on, that’s enough to introduce certain problems in the issue.

So that’s one thing. They’re very [maligned] data and that causes one issue.

The second thing is that these deep neural networks have billions and billions of parameters and we don’t exactly understand what they’re doing. So we can build them, and we know that the work, but we don’t know how they’re doing what they’re doing. We can explain each single element but they all work together in kind of a magical way that they produce results and we don’t know.

Why is that a problem? Well, besides the fact that we don’t understand why they’re doing what they’re doing, it can introduce other things. So bias is obviously a common one. They can be biased in a way that we don’t anticipate, they’re also reliant on the data. They can be fooled in different ways. So if you think of like a face recognition system. Maybe if you were building it you’d say okay, look at the distance between the eyes, look at something. But we don’t know how they’re deciding that a person is themselves. So there are ways we can change the face I very weird ways where for humans they will look exactly the same face, for the computer it will look like a completely different face. Why? It’s coming up with its own feature set.

So there’s kind of this black boxish element to AI that we have to deal with. So that's kind of the second dimension.

The third dimension is I think maybe popular right now, is the generative issue. It can generate things that look human-like. That's a problem as far as like misinformation, deep fakes. I'm sure you all know about that. But it's also a problem for now, for example, for the cyber defense. Like it can generate code. It can generate code pretty well. It's only going to get better. So think of generating malware or generating metric intrusion. So all those kind of things are going to be -- you don't need a human to sit and do this. Algorithms can do them by themselves.

And then the fourth dimension which Edmon kind of alluded to is the intelligence in artificial intelligence. Are we building something intelligent? They're controlling systems, they're making decisions, how do we make sure they're aligned with what we want them to do? So the alignment issue. How do we make sure they're doing what they're supposed to do.

As Edmon kind of alluded to, we don't want to dismiss it completely, the fact that we are building something intelligent and just kind of decide later on that, you know, it doesn't like us or something like that. We're not there yet. We're not saying it's going to happen. But we also don't want to like throw that completely away. We want to be aware of it.

Mr. Begoli: And we don't want to over-react to that too. Because we can talk about this, what we see --

Mr. Sadovnik: There's been a lot of hype about that. But I want to mention it kind of both in the sense that we're not ignoring it, but that's not our main focus and we understand that that's not part of the main threat right now from AI.

Moderator: Great.

For those who came in late, a reminder. This is on the record. Please record it for your accuracy and quotes but there's no rebroadcast.

I'll ask the first question. Three emailed early to get on the list, but we'll have time to go around the table for sure.

My question follows on what you said and focuses very much on the national security space, and we talked about this before.

Talk a little bit about AI and counter-AI in network security. I mean is network security coming to an end and we won't be able to protect any networks ever again because AI is so fast and so smart? And the same thing with encryption. There's all these great science fiction stories about keys that decrypt everything. Is AI going to be a key where there is no privacy, where there's no more secrecy?

Mr. Begoli: We can probably introduce a [protocol] because we both have opinions and work in this area. We do study this at ORNL, so we have this initiative right now focused on [coming to] cyber and that is dynamic adapt to cyber given that we are moving into a space that has two [GARBLED] phenomenon. One is the AI is driving certain capabilities that have not existed before at the same level of scale and sophistication. So there's new types of threats that are being generated. Either there are more of them or they're more sophisticated, so to put it in the context of your counter-AI, to give you a little background on that. What counter, I mean adversarial AI.

We can generate inputs to these models that are trying to make decisions that [default]. That's what Amir was alluding. And it can be done for faces and we demonstrated this at the lab. Left and right. Amir has some T-shirts with a really cool looking logo that look very reliable and they can confuse facial recognition systems.

Well this, we can also build a software or malware that has those characteristics. It's still a malware but it's confusing AI to think that it's not. And actually we have demonstrated that it can be done.

On the other hand, the cyber threat in general is growing so broadly. Think about the internet of things and edge devices -- 70 billion to be deployed by ’27. Don’t quote me on exact numbers, you can find them in the literature.

So A, we need AI to defend our systems today because the scale is such that we don’t have a capable means. And on the other hand, AI is introducing vulnerabilities through its own nature that creates a new hole in the cyber defense systems. There’s a number of literature, there’s a number of experiments we have done to demonstrate that.

Speaking to the question of privacy and security, that’s a really broad area in terms of, you know, we were creating deep fakes of my face, actually my colleague speaking, and live cameras looking at his face and generating my way of speaking and speaks with my accent which you will obviously notice, pronounced incorrectly. We didn’t know where to begin with this, the capabilities are so large.

So I think there are very few areas -- and I’ll stop with that and turn to Amir -- is that again, understanding that the threat exists, understanding the vulnerabilities exist, and doing things today that we missed doing 20 years ago in cyber because 20 year ago we were all rushing to build, connect everything, to deploy state of the art internet systems, get e-commerce [GARBLED], and nobody’s worrying a lot hmm, what happens if somebody puts ransomware, invents ransomware?

Well we are kind of with AI in a similar space right now. There is definitely a market base, in defense sector race as well. Put AI in everything. And we do need to look into vulnerabilities that are existing that are obvious and try to build in safeguards and controls.

Mr. Sadovnik: To add to that. You asked if cyber, cyber defense or whatever, is doomed. No, it’s always been kind of a cat and mouse game. So yes, now our adversaries or the people that want to cause harm are going to be using AI, but us at the lab, what we’re doing is we’re actually figuring out ways to use AI for defense. How do we build AI that can anticipate the different moves? How can we build AI that can detect things early?

So AI is really good at generalizing over unseen conditions. That’s really kind of its specialty. Where most traditional cybersecurity was more like looking for things it’s already seen. AI is better at predicting things that it hasn’t seen yet. So there’s kind of hope, I guess, to answer your question of like we use AI for defense and I think that’s the way we get out of this idea of like oh, no. AI’s coming to get us. Well, no. AI’s also on our side, protecting us.

Moderator: Thanks.

Georgina DiNardo of Inside Defense.

DWG: This is very interesting.

You mentioned earlier [inaudible]. What would you say is the biggest issue you're seeing today, and how are you [inaudible]?

Mr. Begoli: Excellent question.

One, is the I think proliferation of deceptive content that AI can generate. It can just do it at scale and volume and fidelity that is real hard to deal with. We do not have a really reliable way.

There’s the whole idea about creating water marks and we put them in place but they’re not good enough.

Then two is these existing vulnerabilities that are inherent to the AI-based models that are there and I’m not sure that we have a really broad and systematic way to [get them]. That’s why the Center exists. I mean really, it’s like the entity -- I’m not saying we are the only ones who want to focus on that. It is very inherent vulnerabilities, threats, that are still present in the systems that are being deployed.

Mr. Sadovnik: I guess the one thing I'll add to that, I'll say it and then I'll kind of caveat it. I think that large language models have obviously caused a lot of hype and there's a lot of threats they may pose. So I think they may allow adversaries to do things. You know, if they've got to do things they're not supposed to know how to do. But I don't think there's yet a scientific answer to if they actually pose a threat. I think that the companies that are releasing them are worried about it to a degree. So it's hard to say if they're worried about it because it's really worrisome or it makes them look more exciting. But I like to believe they really worry about it.

So I think understanding that threat in a way, because it's such a new technology, because it's evolved so quickly, I think that being able to understand if there's a real threat there, that's kind of one of the first things we have to do. It's out there, people are using it, and we have to make sure that we know what damage it can cause.

Moderator: Nuray Taylor of Signal Media.

DWG: Hi, thanks for doing this.

You mentioned earlier cancer research [inaudible]. Can you speak more about that?

Mr. Begoli: That's a project that is being run now since 2016, 2017. It's called MVP Champion. It's for the Veterans Affairs. It really doesn't have to do necessarily anything with AI security but it's, right now it's the host of the largest I think in the world, bio-bank. I used to [GARBLED] this program before I moved to do this. And we can connect you with more resources.

But the big picture is that the Department of Veterans Affairs is hosting their Million Veterans Program Bio-Bank at [GARBLED] and there’s a set of electronic health records, I believe it’s for all 24 million veterans since 1999. It’s a fantastic resoruce for research. But at the same time we need to guard it exceptionally closely because it’s a super-sensitive -- and that’s our service to the veteran community, that we need to protect it. But at the same time, veterans have contributed this data so it can enable many veterans to say this is my second opportunity to serve. Giving my data so it can be used for problems such as veteran’s suicide prevention. This is [where I specifically] worked on. Understanding different cancers. And the fact that veterans predominantly prostate, liver, lung cancer. And then third was just the cardiovascular disease.

So this is now ongoing where lots of good things have came out of it. It had also some connection to Covid studies in the early days.

My link to that was that given that we were guarding this set so closely, it’s housed in the same place where our Kaiser Labs are housed right now to do this high security research on AI threats. So it’s a very, very well guarded data set. But the best practice in medicine is that you download the AI model, machine learning model from, I don’t know, Harvard Medical School that is really good at detecting Condition X, and then you apply it on this veterans’ dataset to make some logical decisions. For instance we were looking to [GARBLED] highest risk for Disease X or Condition X.

This very question came up a lot, how much can I trust this model? Because it’s kind of hard to think about it for us now, but in 2016, ’17, you just download the model and you use it and this thing will help to answer questions like this veteran is more likely to have a higher mortality for prostate cancer. It should be recommended for surgery. And the question was like okay, these are a lot of critical decisions. How much I can actually trust this machine learning model? It’s not just the quality.

Just to let you know this. Today's state of the art is to test machine learning model how well it recognizes Edmon is Edmon. It's still not state of the art to figure out how good is this model from making sure that it's not purposefully recognizing me as Amir? And this is what we are examining and studying right now. For if this model has been poisoned from the birth to make wrong decisions at the right time for the adversaries. And that really is a problem. This is not exaggeration. Data poisoning techniques can influence the models to make the wrong decisions at the right time.

For instance, if you want to slip the malware into network you can poison the training dataset to, it's a complicated story, to put certain markers into benign datasets so when you encounter malware that has the same marker the model will think oh, this is a benign there, let it go through, and it [GARBLED] can go in and affect that [GARBLED]. And on and on and on. But this is really the original link to how we start looking to that is like how safe is AI to these critical decisions? The answer then was no. We did not actually, because our primary goal was to protect veterans' data and veterans' health, not to play with things. But it triggered this whole research that we are now pursuing to have a systematic approach to that.

Moderator: Josh Keating, Vox.

DWG: Thank you so much for doing this.

[Inaudible] something of an inflection point when it comes to the use of AI for battlefield targeting. There was a [inaudible] public awareness of it. There was an Israeli media investigation last week about how the IDF was using it in Gaza. I think there was a story earlier this week in The Economist about it starting up on the battlefield [inaudible] for targeting.

I'm wondering, one, where do you think that sort of capability is going? And two, what are the sort of particular risks around the sort of fast-paced battlefield use of AI for targeting that you worry about or you think people should be aware of.

Mr. Begoli: We don't deal with those kinds of things in the energy lab, targeting and battlefield and all that. But to tell you this, when it comes to AI the technology is so democratized, and it's surrounding things that enable it such as systems, drones. The thing that we as researchers are concerned about, that the barrier of entry for anybody wanting to misuse AI is very low.

On the other hand, AI itself is not hardened, there's no signs to harden AI. So that if somebody is using it for some of those combat scenarios, to what degree is it going to be reliable? And these are things that we worry about. The reliability of AI to be either misused -- sorry. The resilience of AI for misuse, and then two, malfunctioning at times when it should function and can have all kinds of collateral damage. And it goes, like I said, we are not experts in defense systems and target and all that, but these are universal problems that plague other areas as well.

Mr. Sadovnik: I'm not commenting on any specific use of it, but in general, these type of systems would use what's called computer vision. So they have to basically interpret images and recognize what they see and make decisions based on that.

The reason we would see something like this now is because computer vision has also had a really big increase in accuracy and the way they can do things, so it would make sense that that would be more used.

So on the one hand, yeah, computer vision works very well. We see this also like in self-driving cars. That's another technology that you have the radars and all this other stuff, but then in the end you need to differentiate a person from a garbage and from a cat. So just knowing that something's there is not enough. You have to know what it is, what it's doing.

So being able to -- as I said, computer vision's come a long way. It's allowed us to do a lot of things. But as Edmon kind of said, it does introduce, and you kind of asked about these kind of new threats. So for example, as an example, there are kind of new camouflage ways of hiding things from AI that are a lot simpler than having to like actually hide. You can actually, like I said in the lab, we show that by adding certain patches to certain things -- for a human it almost doesn't even seem any different. For now you can make it look like whatever you want. Right? So you can make a tank look like a school bus or you can make a person look like a cat. So those are real threats that need to be thought about. Because of that black box you're seeing, we don't exactly know what features are used, and we know that these features are somewhat brittle in the way we can kind of change them.

DWG: I'm just wondering, are you able to say -- Anne Flaherty with ABC News.

Is it a foregone conclusion that world powers including the US use AI in targeting I warfare? And we're talking about [inaudible], but I think I saw something on your web site about AI can pick out maybe a cat but not the right breed of cat. That there are these limitations to it.

How comfortable would you ube with using AI in targeting when it comes to warfare? What is the limit on that?

Mr. Sadovnik: I wouldn't answer like it was a foregone conclusion. That's kind of a policy question, how are people going to us it or not, and --

DWG: -- they're using it now.

Mr. Begoli: As a scientist, I do not know.

I guess to Edmon's point of the fact that it's going to be democratized, I assume that adversaries will be using it regardless of what governments do, because it's low cost, you can do it with low computation powers, small drones. So it seems like, just like cyber attacks. It seems like something that you can kind of hack together and put together.

DWG: I think you've kind of touched on this just now, but [inaudible] with precision.

Mr. Begoli: The limits? We are not nowhere near the limits yet. This is where the bar is --

Mr. Sadovnik: -- precision, she's saying.

Mr. Begoli: Oh.

DWG: But that's a good question too. [Laughter].

Mr. Sadovnik: Let's move it to the self-driving car because that's a little bit of a safer -- it's an analogous problem. I think that we don't have -- well we have kind of test self-driving cars. We don't have them yet out. I think that at least the people that are building these cars and are worried about like people getting hurt, and maybe the insurance companies, are really making sure that they're not releasing anything before it's completely ready to be released.

And I do think that with the help of what we're doing at the labs and making sure that we identify these novel vulnerabilities, we can get AI to a place that it will be secure enough to be used in those types of applications.

At the end of the day, will it make mistakes? Maybe. But do humans make mistakes? Yes. And I think if we can get it to a place where we can understand why it's making the mistakes, improve them, and kind of keep working on it, I think we could get to a place where we can rely on AI to do a lot of this stuff for us.

I personally am looking forward to like my k ids not having to drive a car. When they drink they have the car drives itself.

So yes, are we there yet? Apparently with cars not yet, but I think we're getting there.

DWG: Dustin Volz, Wall Street Journal. Thank you so much for doing this.

You mentioned earlier the biggest threats right now, one of them being the proliferation of deceptive content. I just was hoping to maybe unpack that a little bit more in terms of any particular forms that you're working about. I've heard, for example, that audio [defects], audio deceptive content is sort of an area where some are more concerned because it seems like that's an easier space to convincingly create separate content and also harder to identify and convincingly disprove, compared to sort of video or imagery and so forth. I just wanted to get your thoughts on that and sort of what forms particularly are most concerning to you.

Mr. Begoli: Honestly, we have a bill to present, work we've done to demonstrate this. So I'll have to give you a little bit broader story because it's interesting stuff.

Beginning of the pandemic, [actual] tele-veterans program, they received some materials and they were asking us how do we use these to confuse the mass, the public about it and create deceptive content? So what we did is we trained, at that time it was [Verse], so this is before ChatGPT was released. We trained it in all the available news sources, a bunch of data, and asked it to generate an article about Covid 19. I still have it. Two slides. A beautiful prompted piece of nonsense that is written in excellent English, reads like it makes total sense, none of it's true. And then we demonstrated you can proliferate it easily through API to [get to] where you can create thousands of instances [GARBLED], and you can, of course now that's known through ChatGPT and all that.

So textual content, piece of cake to demonstrate. I mean it takes no time to do it. The number of projects -- of course ChatGPT kept fixing some of those issues where you can, you know, the jail break, the prompt, and ask it about misinformation [GARBLED] prose, but it’s all fake and it’s very targeted in terms of what it’s doing.

So in text, no problem. Voice, absolutely easy. Verbally, and it’s a relatively restrictive modality because it’s a sound just like a text so it’s kind of harder to detect because you can be more precise I what you do.

But we’ve also recently done experiments with deep fakes, video deep fakes, a colleague of mine took my, something I was talking on YouTube about some science conference, and then he morphed it with his own face. So this person looked like sort of like him and sort of like me, spoke in my voice, and we did it in 2.5 hours and cost us $20.

The state of the art deep fake detectors talk about it. You can detect the veins and you can look at blood flow. I suspect that it would probably cost $100,000 and would take months to detect, where this thing took $20 and 2.5 hours to do it. And if you take this super high fidelity made up video content and start spreading it, imagine what kind of message you can put into the deep fake.

That’s the problem if you’re observing. If it’s easy to make. I think it’s hard to detect because these are not trivial things. I mean it’s a very high fidelity. In the early days you would have like well, there’s a crease between the neck and the face. State of the art deep fakes you don’t have any of that. I mean they really look high fidelity videos and you need a very sophisticated infrastructure to detect some of those minute differences in the face like looking at the blood flow through the veins and other, you know, what kind of equipment you need to do to differentiate that.

You can also generate other content so you can generate signals. This is something Amir was talking about, deep learning. Deep learning has been this kind of magic technology that you can use it in multiple different domains. So it’s not restricted to any form of -- it's not restricted. Any kind of digital content can ultimately be deep fake, and capabilities to do it are just getting better because more data and bigger models are really good. So generative AI, it's a kind of a deep fake generative used for those purposes.

Mr. Sadovnik: I'll just add that when you think about this generative content there's kind of two elements. There's a technical element of like generating it, and then there's like kind of a social side of it of like what is it going to cause? I don't know if Edmon [GARBLED], but I'm definitely not an expert I social science, and actually that's where we need to have more collaboration with these type of people to understand.

Like how much worse is LLM generated content -- this is kind of like just bringing down like fake -- big of a threat is it, right? How much worse is LLM generated content from like hiring 100 people sitting in, you know. Since you can do it more targeted, maybe there is an extra fit there, or maybe -- what would it actually cause? Would it cause people to believe in things that are not real? Will it just cause people to not believe in anything anymore? So there's a lot of kind of social questions that are here that I think are more important almost than the technical ones.

I think part of what the Center's trying to do is kind of bring people from different fields to be able to answer these questions in a more holistic manner than just kind of saying oh, you know, deep fakes are going to, you know, ruin this or ruin that. Well, is it? How bad is it?

Kids are aware of deep fakes. They kind of are suspicious of a lot of things. So anyway I just wanted to kind of put that in as well.

DWG: [Inaudible] from Foreign Policy Magazine.

In terms of a lot of the harms we [inaudible], with deep fakes, with misinformation, a lot of it tends to focus on sort of the

US or Western or English-speaking content, especially with the number of elections happening around the world this year. I'm wondering how much your research has kind of focused on other languages, other cultural contexts, and countries where digital literacy may not be as advanced as it is in the West. So the societal events you just spoke about.

If you could just talk about that a little bit, how you think about and see how these models perform in these different contexts.

Mr. Begoli: First, as you can tell by my accent, I'm coming from western Balkans and I come from a group of languages called [world resource] languages comparing to English, Hindi, Spanish, Mandarin.

It does, so that domain, that's a technical domain. In areas where you have high resources, you can create high fidelity whatever AI [base things]. So there is a little bit of that disparate, you know, that AI is more potent in the areas where you have lots of data resources.

On the other hand, this is something we just kind of tangential explored before, and this goes back to our work with the elderly and so on, also age impact. So we talk about kids. Kids know about deep fakes. But I remember at the beginning of the Russian invasion of Ukraine, 2022, and there was some really badly made deep fake of Zelenskyy telling troops to give up. And you know, we had this discussion about yeah, for teenagers or us it's obviously a badly made deep fake. But for a granny who has a cell phone and looks at it, oh, my village, we are losing, or something like that.

So this is something that we haven't spent as much time at the lab looking into it, but it's these different cultural groups. Different cultures and then different populations, circles based on their sophistication level in terms of understanding technologies. So there are, I think, vulnerable targets.

Sorry, I'm kind of answering from two different angles. One is A, certain groups have more resources to make AI more potent to do good things and to do bad things. And on the other hand, also certain groups within a population are more vulnerable to exploitation. Because again, if you receive a call from some trusted source that speaks exactly like your child or something like that, I mean you're more likely to be a target for exploitation. I know it's a kind of esoteric answer, but --

Mr. Sadovnik: It's a great question, and that's exactly why I brought up the idea of bringing social scientists and other people in. Those are the exact questions I don't think computer scientists are qualified to answer. But people that deal with like the disparity of how digital media works in different parts of the world and how it might affect it, I think that's where that's really important, and I don't think we've done enough in that realm, so that's a really good point.

Mr. Begoli: To add to that, it comes to existence of our Center. To be honest with you, this is not a [GARBLED], we are kind of trying to bring humility into this field because if you go to YouTube and look at lots of our colleagues in computer science, they frequently make statements like they have absorbed themselves all the knowledge of the world when it comes to social sciences, neural science, psychology, AGI can do that, and -- I mean there's lots of [opinions] among computer scientists. What we advocate is that this needs to be interdisciplinary research and interdisciplinary field that brings expertise from social sciences, psychology, neural science, philosophy, law, ethics, and so on because this is a field that's going to have a high impact. Right now it's still kind of pretty much, like I said from this technocratic sources that have limits to their own expertise and knowledge, although I'd like to have more than that.

Moderator: I've never heard a scientific lab smackdown like I just heard. [Laughter]

DWG: Julian Barnes, New York Times.

I’m struck by what you’ve laid out here of the humans poisoning the data for the AI and giving them bad datasets and the AI poisoning our world with bad data out there in the arms race between us -- who can confuse the other faster.

But this idea of sort of bad data corrupting the AI is really interesting. Especially in what a lot of us here are thinking about, the defense space especially, where we talk about doing, the US is probably not going to be the forefront in using AI to target offensively, but we’re obviously using kinds of AI for defensive targeting like in the cyber defense.

In your mind, where do you, what is the balance between restricting the data -- especially in a government context. Restricting the data sources so that your AI has less stuff but you’re more confident, versus give it more data so it is more robust but there is more risk of corruption in there.

What’s the balance right now? And how do you purify that data in the government context?

Mr. Begoli: I’m planning a paper for next year’s conference called [GARBLED] on this specific subject. I’m going to share my beliefs, I’ll let my colleague share his beliefs, because they really are beliefs or opinions. At least speaking in my own name.

I think from a government’s defense purposes, I think we need to have a data certification as a source, like we have a food that we eat. You want to make sure that somebody is making sure that you don’t have something wrong in that food. It’s that important, that’s one thing.

Thinking much more far-fetched, I do believe that within number a number of years we will move past this purely data hungry or data-based AI because it’s kind of first step. The next step in the AI evolution is going to be the models that will advance themselves structurally. So it’s mathematics advances, not just

-- I only know what I’ve seen.

If you think about the human mind, we don’t only know what we read in certain books. We learn, we have a system that helps us evolve and learn through life. AI is not there yet. AI is right now, here’s the data, what you see in the data is what you know.

So that’s a more far-fetched concern that either we need to make sure this AI we’re developing right now is not defective so that we can introduce the defects, but I do believe in a more stringent certification of data sources, and how the AI was [GARBLED] because it’s just way too easy to poison the data. I mean you can just mess with it left and right.

Just to give an example. Sorry, I don’t want to belabor the point, but to really bring it down to specific examples. So you’re training the cyber defense detector that detects PDF files. And PDFs are notorious to have malware carriers because you can embed scripts in it.

Well, the way to point this would be, you have a set of known clean PDFs and known malware carrying PDFs. Then what you do is in the clean PDFs, you insert somewhere a specific let’s say phrase. And then AI has learned that this PDF has all this malware and this clean stuff has all these phrases. And then you as an attacker, first you go into this clean dataset and insert some specific phrases. They might be English phrase or it can be some computer code. So this AI has learned that the benign PDFs are the ones that look clean and also have these phrases. The next time you introduce another PDF as an attacker, and then you insert this phrase it has seen in the benign dataset, so what cyber defense is going to do, it’s going to say I’ve seen this malware PDF ad this kind of looks like it, but ''ve also seen this clean PEF and it has this phrase and this thing has this phrase. Lets it through. Then it’s going to have embedded malware.

So you need to get into supply chain, mess with it, insert things that will later confuse AI, and we don't control it. I mean it can be -- what happens in the AI is this process called fine tuning. You take one model, then you refine it, make it your own. Then you refine it. Then you refine. So you have like six, seven chains down the line of refinements. If this thing has been corrupted up front, these corruptions can propagate all the way. So you might have original developed AI for some civilian purposes, whatever, for some completely benign scenarios, but that model can be used six steps down the road to develop like state of the art cyber defenses, or for image recognition and all that.

That's why I'm saying that I do believe that from a defense, national security point of view, we do have to do much more stringent view and understanding of the [cleanliness] and the sourcing --

DWG: Which is why you can't off-the-shelf buy an AI model and then just keep training it because it might have the corruption?

Mr. Begoli: It might. And it's one of the things. You can sort of try to test it, but it's so difficult to do it later because there's such large models, train on, you know, whole internet or something, you know. So it's kind of hard to find is this model truly poisoned.

DWG: A few needles in that --

Mr. Begoli: It is. And they need -- It requires very sophisticated computational intensive approach to actually detect data poisoning, presence of the data poisoning in the model and stuff. It's very, very difficult to do it later. So that's why it's important to do it up front.

Mr. Sadovnik: I'll just add, that's a great question. Obviously that's something that companies and the government are dealing with. On the one hand if you restrict yourself only to the dataset you have, you're missing out on some data.

Especially in like cybersecurity where you know new thing are coming. You do want to keep learning because if you don't, if you're going to stay with the data you already have, you're already out of the loop.

So this is an active research area of how do we deal with the fact that we might have adversarial intruders that try to introduce these PDFs. So there's different methods of like trying to detect poisoning, trying to detect things that are out of distribution, things that look weird before you train on them to make sure. So it's basically using AI to defend from it.

Or how do you weigh things differently? So maybe if you know certain data has a chance of being less reliable, can you weigh it differently in the training process that if it has an effect, it's a smaller effect, for example, than the trusted dataset.

So this is an active [GARBLED].

DWG: Does that exist now? The way --

Mr. Sadovnik: Yeah. There are different methods of doing that. I mean again, it's an active research area, so it's still not mature, although maybe there are some things out there that are already used.

DWG: Rick Webber, Inside AI.

I want to talk a little bit about process. President Biden issued an executive order in October, and part of that executive order, DOE's tasks were assisting on testing the safety and security of AI. Can you talk a little bit about what role the Center is playing under the order? Or if not directly under the order, how is the order sort of -- What's happening within DOE?

Mr. Begoli: The Center started before the executive order was put in place. I think the process for the executive order started probably in the summer of last year. We started a little bit earlier, so it was fortunate that it all happened at the same time.

I think there are lots of areas that are very important, so from that point of view in terms of scope and in terms of -- I mean it's a 100 page. It took a long time to go through all the executive order. There is lots of important areas that are covered.

DWG: I think it's been identified as one of the longest EO's --

Mr. Begoli: And from what we know, a number of competent people worked on contributing to it. So yes, DOE is named out and DHS and a few other agencies. So we are supporting as a national lab entity, we are supporting other agencies with expertise. And specifically in the areas of this experimental evaluation and particular testing. That's really our primary role is that we -- there's lots of policies and there's lots of papers that exist, but our primary role serving very different agencies are doing these [Red Teaming] experiments, trying to trash AI by developing very sophisticated attacks that can test if the AI is reliable, and at the same time trying to work on advancing the science of AI security. That's a really big thing.

I know it sounds kind of very lofty, but today we have cybersecurity. We have that field and there are degrees and there are colleges and there are centers of excellence and all that. We don't have anything like that in AI. And one of the major function we have is to actually propel the science of AI security because we'll need it. It may sound esoteric right now but we will need AI security. So that's our big contribution is also experimental methods, methodologies, and then science of AI security.

DWG: You mentioned Red Teaming is there separate from [inaudible], there's really not even any -- when it comes to AI There's really on even definition. How do you Red Team AI? Is that something that you're working on?

Mr. Begoli: Yes.

DWG: Do you have any thoughts about how to define Red Teaming AI?

Mr. Begoli: Red Teaming became a popular term because it's established in cyber, so in AI is a little bit more complicated than that. So we are using this term because it's easier to communicate to others who are engaging in Red Teaming, but we are I think taking far more vigorous approach than just this general term, Red Teaming. And really what it is, Amir is really leading this, is that we are understanding from the literature and from our own research what are the true vulnerabilities of AI, and then trying to replicate it.

Because what happens with literature, especially scientific literature which is right now the main source, somebody can do it just to get dissertation and kind of semi-works or in some instances real exploit. So we try to understand these and then create experiments that are used against AI models to demonstrate -- I mean it's reliability research, that's really what we do. AI vulnerability research is much better formal term than Red Teaming. But Red Teaming, sure it's coming up with scenarios, how could one exploit the systems that rely on AI, and then exploiting them and highlighting those.

Mr. Sadovnik: One of the main goals of Kaiser's and [GARBLED] security research is this kind of testing and evaluation. But under adversarial conditions. So that's why we need the adversarial research to understand how can -- what shall we look for? What are [GARBLED]. And then yeah, give us the algorithm and let us test it and see how it works under these kind of real conditions where especially in the national security domain these will be attacked. We know that for a fact. So let's see how it performs under those conditions.

DWG: [Inaudible] from Washington Times. I have a couple of questions.

The first one is about your visibility. You mentioned the computer vision model. How it’s been used in Gaza. Obviously the US [inaudible]. So what I’m wondering is, what’s your ability to know what else is out there? If you’re doing vulnerability research and that sort of thing, you may find a way that facial recognition, you mentioned earlier, that’s one thing that’s different from another. It works for one system in the US but not a different adversary’s system.

So what kind of visibility do you have into other AI systems and [inaudible] applicable across different models?

And my second question, with respect to, you mentioned before about there not really being any AI research [inaudible] in the same way that there’s cybersecurity research [inaudible]. Can you kind of help me understand what you guys do differently from say a company like [Entropik], that is very out there, talking about [inaudible].

Mr. Begoli: You do lots of work in this area, so I’m just going to open with a few things.

This field is developing so fast that it’s not even what kind of visibility we have is that how can we absorb everything that’s happening? To the point where we need to develop probably our own large language model to study this field itself to tell us what -- there are 340,000 papers on adversarial AI. And it’s not like there’s some secret entity somewhere developing this. A lot of it is happening by enthusiasts, by academia, by student, by companies are doing their own stuff and that’s going to lead to [Entropik].

So our biggest challenge is to keep up. It’s not like having some secret knowledge. It’s absorbing everything that’s happening. Because lots of things are being published because people are excited about it. And they are coming up with some real exploits. I mean first exploits were shown in 2014, famous [GARBLED] and all that kind of stuff. Since then there’s just been cyber and all these biometrics, these are all very right there. So we are really skimming every resource we can get our hands on. The most important thing is to replicate and validate.

Two, like Entropik, companies -- our expectation is companies are going to move into that type of thing. There's already [GARBLED] teaming effort and Google.

The biggest difference that we play in this space is that we sit between academia, government and companies. Companies in the trenches. They want to make sure when they give you [API] that the [API] is working for your banking application, for instance. You know? Academia is not so much focused on applications in the basic research. We kind of -- then we also understand the interests of the government and agencies we support. So this is where we are. And advancing the state of science. Because Entropik, which are great, and we met with them -- they really are good, and boy, do they have resources and all those kinds of things. But they are very narrow focused on things that will ultimately hurt their profit, rightfully so, and their products are great and all those things.

We are looking a little bit further than that including, like I said, this far-fetched issue of existential threat. I will say this opinionatedly. The scary stuff about AI also makes AI sell better. Oh, let me try and see, you know, if it's really that smart. Then there's also this hype. And we are trying to understand, are we moving in direction that can truly hurt United States and can hurt humanity? That is the primary question. It's not like well, it could hurt my stock options. No. It's like will this thing become dangerous to the point that it's going to cause our control systems to fail, that's going to cause -- just to give you this little example. I'm sorry, I cannot contain myself. But like when I talk about this existential threat, it may not be the SkyNet, it may be simple AI that is so well designed to perform its function that [GARBLED]. It's not going to align with our own interests because it's going to be so good and efficient that it's going to drill the holes through everything, fine, to solve the problem that it's optimized to do and it's going to do it at the expense of us. And it may be so good at doing it, it's going to be also hard to stop. And if you embed it into every single aspect of society and if it's interconnected, it's just so omnipresent that you cannot go back and delete it from everything. And so again, it's not like some big [mine] trying to kill humans, it's just a thing that is so good at doing what it does, it can hurt us because it's misaligned with our role.

Mr. Sadovnik: I'll just add, I think Edmon kind of alluded to it, but AI is not like somebody sitting somewhere like in another country like developing its own AI. We know what AI is. It's pretty much the same everywhere. Yeah, they have different training sets, they're using maybe a little bit of different architectures, but at this point we kind of know what it is and those vulnerabilities do carry through. And we actually see that here. We can see it here, but if you take a model developed one place and use that effect to develop models -- companies are almost as private as a country. These things transfer very well.

So we are looking really at the kind of more fundamental threats to these AIs that are really, I don't think we have to worry too much about like what other people are doing. They're basically using kind of the same thing.

Moderator: We're at the hour mark. I think we have time for one more question and they said they could stay a few minutes late, so please.

DWG: [Elias Cole], [Cyber] [Inaudible].

I wonder if we can go back to the Red Team issue. Can you speak to the maturity level of the AI Red Teaming discipline at the moment?

Mr. Sadovnik: Personally, I think it's not very mature. I think it's still being developed, and I think that -- There's a scientific way of doing it I think that we're still figuring out how to do. Large language models might be, again, the easiest kind of thing and most in the news and people are Red Teaming it. A lot of what it is is basically hiring people to try to break it, right? And that's one way to do it.

But I think -- and the reason is because of the Black Box-ish nature of it. Like we can't really tell what it's going to do so we need to kind of just test it.

But I think there are more fundamental ways -- we built this thing, right? We can look inside of it. We can see what it's doing. And I think there are more fundamental ways we can look at it, like what does it actually know? Can we push it to the extreme? Basically I'm going to use AI to Red Team AI basically, right? Can I use algorithms to find ways to stretch these other AIs to the limit to make sure they stay within the bounds? So this is a very active area of research. I assume that even if companies were here -- I don't know if they would say the same thing, that we're all kind of figuring this out together. We've been to a bunch of different workshops with these companies, trying to figure out what are the best practices and how do we do this? So I don't think we're there yet. It's a relatively new thing and I think we're still working on figuring that out. But that's one of the things we're doing at the Center for sure.

Moderator: One last question.

DWG: [Inaudible] with [Inaudible] Magazine.

I was curious, as you mentioned this is not happening in isolation and this is the first time I think in [technological] development of this magnitude that is not really being developed in federally funded research labs. It's being funded by private companies [inaudible] by everybody. And we're really seeing this, I just saw it in Ukraine, you see these big companies just really, again, kind of like putting this into tools that people can easily use and then selling them. Right now they're saying go to the side of the West, the side of the United States [inaudible] national security goals. But in the context of national security, it's so easy to turn around and sell it to an adversary.

I'm curious as researchers, kind of how you're thinking about that balance between, again, how do companies, often in the hands of like a handful of people, having so much power, being able to not only develop but sell this very powerful technology to others, to adversaries. How are you thinking of that from your perspective, from your side. What do you say about this intersection?

Mr. Sadovnik: I would say the Center really is focused on AI security. So what we're trying to do is really, we kind of said this over and over, but maybe I'll say it one more time. Understand the threats to these AI systems and how do we make sure that they stay within the bounds that they're supposed to.

So for me as a researcher, I guess, as looking at that, feel that I actually don't -- I feel like it's good for everybody. I don't know who's going to be using it but I know that whoever's going to be using it, we want to make sure that it stays within the bounds. I think that's an important field regardless of who takes it.

Mr. Begoli: I don't have much more to add. The reason [GARBLED], we really are in a very new, weird state.

One thing I want to add to the previous question too is we are developing these Red Teamings, and everybody talks about LLMs, but as an AI scientist we know that we are probably within a year or two of a new architecture because this is the trend. Every two or three years there's a major breakthrough in AI technology. So the strategy you develop for today, we cannot over-focus on like Red Teaming LLMs because a year from now we're going to need Red Teaming for architecture, we don't even know what is it.

That goes back to the question you asked, and that's about companies are doing, but there's also very strong open source community. There are some companies like Meta who are very much interested in open sourcing, other ones are not. It's an international phenomenon, so it's kind of democratized.

So just to give you an example, so Hugging Face is kind of a clearinghouse for, it originally used to be -- I worked with those guys for a long time, and when I teach I use their stuff. I really used to support, I still do, what they do, but they were [GARBLED] processing open source company. And then they created this clearinghouse for models for [GARBLED]. They started with like 55,000 models. Right now, since the LLMs, they have 500 and some thousand models and if we track the trend they'll probably have 600,000 models by July. This is never seen before, anything like this. But anybody can contribute a model and use them. So how is it proliferating? Who's going to use it? Companies versus government entities. I don't know. I wish I could give you a better answer, but it's a very dynamic and kind of wild field, and moves extremely fast.

Moderator: Thanks. Those of you who know the Defense Writers Group know that we usually have three-star, four-star commanders, Under Secretaries. We've never done a lab before or Oak Ridge. I think this is your first time that DOE has let you out in public -- [Laughter]. And probably the last.

Anyway, again, having visited, your work is so important that I'm really honored that you came to meet with us here today. And for follow-ups, Eric Swanson is a crackerjack public affairs officer and can follow up with people.

Mr. Swanson: Is that a compliment?

Moderator: Yes, sir. Because it tastes good and there's a prize inside. [Laughter].

I'm dating myself. Crackerjack is an ultimate compliment.

I don't know if you have any final comments or wrap-up? Then we'll just say thank you for your time and thanks to all of you in attendance.

# # # #