**RedZone Podcast Episode #65: AI Safety in Cyber Security | AI Decision Making | Wireheading | AI Chatbot Privacy – with Roman Yampolskiy**

Bill: Roman, I want to welcome you to the show today.

Roman: Thank you so much, Bill.

Bill:  
[00:00:30] You've recently written this book called Artificial Superintelligence. Can you take our audience back to what was the genesis of this book? Where did the seed initially germinate that you were, I have to write this book.

Roman:

[00:01:00] I do a lot of research on AI safety and it's mostly published in scientific venues, conferences, journals, and I realized most people would never read a technical paper. The best idea I had for introducing more people to the concerns I had about artificial intelligence was through publishing a popular book.

Bill:

[00:01:30] So you wanted to reach a larger audience with your message which is great because it's quite easy to talk about the benefits of artificial intelligence but I love that your lens is on safety. Where do you like to start safety discussions, like if I asked you what is artificial intelligence safety, where would you start me?

Roman:

[00:02:00] We can distinguish short term and long term concerns. Short term, we worry about the type of decisions the systems are making today, whether it's credit reviews or employment decisions. We want to make sure they don't have any bias, not racist, not sexist, basically not repeating the mistakes humans are making. Long term, there is really no limit to concerns we have. The systems will control most of our infrastructure, will have impact on every aspect of our lives, so it's very important to make sure we fully understand how they operate and still remain in control.

Bill:

[00:02:30]

[00:03:00] What's really interesting is that I was having a conversation with my CTO yesterday. We're leveraging this cloud service for our customers. We're leveraging this very advanced security analytics organization to do a lot of heavy lifting. However, they're really notifying, they're not taking action, which is fine. Notification is super helpful, but when we've got to get down to the customer level we actually need to go do something about what they've found so we've written some algorithms and tools, scripts that really automate the ability to go and actually effect an action from this super-intelligent service. That action is really an algorithm of sorts. It's really an intelligent way to slice through the data and take action.

[00:03:30] I said to our CTO, I said it's very, very important that we programmatically know what this tool is doing and that it talk to us as a human so that we know what was happening. Maybe that's a superficial way of talking about artificial intelligence but what would happen if we program AI and we don't know what it's doing next, like it's not talking back to us? What happens?

Roman:

[00:04:00] That's actually a very common approach. We have those deep neural networks which are just black boxes. We don't understand how they make their decisions. We just measure accuracy of the final decision but it's very important to also be able to understand how they got there, can they justify this decision. Let's say in medical domain, so the system says, okay, you have cancer, this is the treatment, but how do I know it's correct? What is the evidence for that decision? It remains a very open area of research, trying to get explanations out of those pattern recognizers.

Bill:

[00:04:30]

[00:05:00] One of the stories was interesting in your book. I'm going to read it because then I want to use the technical word around what it's called. Maybe we could talk about this a little bit. This is going to be illustrating one of the problems with AI that you highlight. You tell this story about ... you wrote here: "A married couple, both 60 years old, were celebrating their 35th anniversary. During their party, a fairy appeared to congratulate them and grant them a wish. The couple discussed their options and agreed on a wish. The husband voiced their desire. 'I wish I had a wife 30 years younger than me,' so the fairy picked up her wand and poof, the husband was 90." This is one of the things you talk about, literal wish granting which I guess is called, the more technical word is perverse instantiation. That's one story, but maybe tell us more about this.

Roman:  
[00:05:30]

[00:06:00] Human language is very ambiguous in general. A lot of times, we don't take what is being said literally. We understand much deeper what is meant and what would make our conversation partner happy whereas machines don't have the same common sense, they don't have the same background. That's very easy for them to misinterpret a given situation. For any command, for any desire you'll express, there are multiple ways of getting there. Maybe the final result is the same, but how you got there is different or maybe even the actual final state is very different because there are multiple ways of interpreting what is being said. There are many, many examples of doing that.

Bill: Can you give us another one?

Roman:

[00:06:30] Sure. A common example would be, okay, I created this superintelligent system. I'll ask for happiness. I want to be happy. I want everyone to be happy. As people, we kind of understand what that means. We want to be healthy, wealthy, beautiful, all sorts of desirable properties, but there are other ways of accomplishing that, for example, drugs is a very simple way of getting to happiness quickly.

Bill: Oh, I see, so the conversation of what makes someone happy, the literal interpretation for the computer might be take drugs because they're not going to know the subtle nuance.

Roman:  
[00:07:00] Exactly. As a person, you understood that I didn't mean get me so high I can't really do anything about it but to a machine, it would be a perfectly reasonable approach to get to the same state.

Bill: This is something that people have been tackling before ... people are thinking about this now, right? It looks like you've referenced in 2006 the Open-Source Wish Project. What is that? Obviously, people are thinking about these concepts well before today.

[00:07:30]  
Roman:

[00:08:00]  
Today, it is a very hot area of research in AI safety ... human value, learning, trying to understand how humans value certain things and express them. The citation to the Open Wish Project, that was not really AI research, that was more the philosophical discussion of nuances in really dealing with genies and granting wishes, but the parallels are very good between this powerful genie, mythological creature, and superintelligent system. In both cases, you want to be very careful about what you wish for. There is also parallels with religion. We're always saying, okay, if God wants to punish you, he'll give you exactly what you ask for.

Bill:  
[00:08:30] What you're saying is that people are using language ... they're researching use of language for extreme clarity. Is that one way of saying it?

Roman: Right, but we understand we can never get to that level with human language. Programming language, yes, we can be very precise, but human language will always have this ambiguity, this fuzziness, so we really need machines which can work with that. It doesn't confuse them. They understand, okay, when you said I want a hot girl, you don't mean someone with fever.

[00:09:00]  
Bill:

[00:09:30]  
That's interesting because I know that in picking for immortality, on of the quotes is: "I wish to live in a location of my choice in a physically healthy, uninjured, apparently normal version of my current body containing my current mental state, a body which will heal from all injuries at the rate of three sigmas faster than the average given the medical technology available to me." Is that basically ... that's what brought up my comment about clarity. Are we going to need to really rethink how machines are going to interface with humans? How do we handle this programmatically?

Roman:

[00:10:00] Absolutely, so just communicating. How do you communicate with something which is so much smarter than you? We have hard time communicating with animals, for example. Supposedly, we're smarter than dogs but we're really not doing well communicating both ways. Now think about something with IQ of 1000 equivalent, right. How do you communicate with something like that. What they say or try to communicate may be too complex for us to get and what we say may be too ambiguous, too fuzzy, for them to implement correctly even if they wish to do so.

Bill:  
[00:10:30]

[00:11:00] One of the points that stuck out to me in your book and in the research, this is a high level point that we need to understand. Human beings are not infallible, of course, and I guess someone is trying measure our infallibility rate, but even if a superintelligent machine ... and I'm highly paraphrasing, so I'm looking forward to your comments on this. If a machine is making X number of million decisions a second or every five seconds or every ten seconds and they're not going to be infallible either. In fact, even if they were smarter than humans, mathematically you were saying that even if they're 99% accurate, with the potentially tens of millions of decisions that are being made, that's going to be a lot of errors.

Roman:

[00:11:30] Right, and if the system has a lot of impact, even the tiniest of errors, a very small percentage of errors after they accumulate, would have tremendous impact on society if that machine controls infrastructure, if it controls economy, even if it's wrong once every million decisions, billion decisions, that's hundreds of wrong decisions every day.

Bill:

[00:12:00] When you're talking and educating people that are in decision making capacities within organizations, whether they would be governmental or enterprise, what's a good counsel for them regarding bringing in AIs into the organizations. What is a good question that they should be asking themselves and asking vendors when they're bringing these tools in?

Roman:

[00:12:30] Of course, we need to understand what the technology is and how it works. It shouldn't be this magical black box where they assume it's an oracle with perfect knowledge. Also, I always suggest that really important decisions, things like who lives or dies, should never be left up to the machine. There is always a human in the loop who makes the final decision. Once you surrender that control, there is really no way to get it back. We already see we surrendered control in terms of complicated domains, things like stock markets, nuclear power plants. All that is controlled by software now and there is no way to undo it.

Bill:  
[00:13:00] Okay, so making sure you understand what is in the black box and not surrendering control. What does that mean? Does that mean that the black box has to report back to you, essentially it has to give you a certain set of ... you're strong into the safety side, so what would that black box give you from a safety signal point of view?

Roman:  
[00:13:30] It should be able to explain how it makes its decisions at the level a human would understand. It cannot be oversimplified, like we talk to children ... we'll tell them something not quite correct but simple enough for them to understand. That's not good enough to truly evaluate if the decision is a correct one. It has to be able to explain its decisions and we should be able to verify it. If there is any possibility that it is wrong, we should be able to detect it and override that decision.

Bill:  
[00:14:00] I had a guest on recently who said that he thought that AIs were going to be teaching AIs, that there was going to be an AI that would be the exemplar of values or whatever those may be in that particular domain and it would be teaching the AI or sort of be a governor. Do you believe that that's possible?

Roman:

[00:14:30]

[00:15:00] Agreeing in values is already a very difficult problem. We've been trying to do it for thousands of years in philosophy and we've failed. I don't think any two people would agree 100% of anything. What's worse, once we transfer it to machines, now again you're talking about something capable of enforcing that set of values on all of humanity at least in long term, so that's going to create problems. Having one AI monitor another AI doesn't really solve any problem; it just transfers the problem to a different software piece, different function within the problem. It's not making it easier. In fact, it creates additional levels of communication which might cause additional problems so I don't see it as a solution, this idea for AI governor or AI ethical control system. It doesn't seem like as a separate piece it adds anything.

Bill:  
[00:15:30] Yeah, this seems to be moving faster than we are going to be able to govern. I wonder when you think about this problem from an AI safety, if there's no time to govern and set up controls because the technology is moving into our infrastructure as we speak, prior to laws and governance structures or frameworks or protocols or acceptable use, et cetera, how do you approach that problem? How do you think about something after the horses have been let out of the corral?

[00:16:00]  
Roman:

[00:16:30]  
That's the problem. We have this exponential technology. AI is growing at that rate. Everything in terms of control, whether it's political or legal aspects, at best is linear, maybe slower than that. We don't have solutions. We don't have AI safety mechanisms and no one even knows how to get there. It's very easy to see that creating a safe AI is harder than just creating any AI. We've been trying to do it for 60 years at least, trying to create intelligent machines. So far, we didn't get there, so I'm suspecting it will take at least that long or longer to make an AI safety mechanism, which is not very encouraging given predictions on when we're going to get to human intelligence.

Bill:  
[00:17:00] Interesting. What's interesting about one of your chapters ... it's called Wireheading in Machines. That was a new concept for me. If you were running a lecture in your university for new students, like it's a 101 class to introduction to exponentials and you were covering AI as one of your topics, how would you explain wireheading, addiction, and mental illness in machines? That was the topic that was super interesting.

[00:17:30]  
Roman:

[00:18:00]  
It's exactly what you think it is in humans. You know how mental illness works. You know addiction. We're addicted to drugs, maybe addicted to pornography, all sorts of pleasure-driven stimuli. Machines are subject to the same problem, especially the ones which are based on reinforcement learning. If there is a reward being given by a human or coming from an environment, a lot of times it's easier to go directly for the reward instead of trying to do productive behavior. The system may try to steal the reward, influence human programmer to provide additional reward, basically what we see with drug addicts.

Bill: Okay. Have you seen this happen yet? Have you seen basically a machine gone awry or have you seen an example of this?

Roman:  
[00:18:30]

[00:19:00] Obviously, machines today are not at the human-level performance but then we do evolutionary computeration or those rewards-based algorithms, we do see this sometimes where the system is not really interested in accomplishing the goals of a programmer but it just wants to collect as much reward as possible. Maybe I am training a system to, I don't know, play soccer or something like that and there are reward points given for certain behaviors. Maybe controlling the ball gives you certain number of points. The goal of course is to teach the system to play soccer, to score goals and so on, but all it does is really just grabbing the ball and running around with it because that's where the points are.

Bill:

[00:19:30] It seems to me that almost because of the spiraling of this out of the governance control, it almost seems to be very useful to learn how to harness AIs to solve certain really hard problems that are beyond the human mind to solve. Maybe it would be domain-specific like infrastructure of nuclear plants and the machine could be assigned to actually figure out the vulnerabilities assigned to that specific narrow domain and keep working on it domain by domain and getting smarter and smarter with the intelligence so that it becomes almost, it's assigned to the safety side. Is that possible? Is that just a theoretical burp on my part or how would I approach it?

Roman:  
[00:20:00]

[00:20:30] We're definitely using AI as a tool to help us in many individual domains. Cybersecurity is a great example. We use machine learning to identify attacks, to discover new attacks. That's definitely the case, but there is a fundamental difference between the OAI, something that is one domain only expert and a general AI, something capable of knowledge transfer, capable of succeeding in multiple domains. I really see the danger is coming mostly from this general intelligence. So far, we don't know how to do that luckily. It may be that in 20 years, 30 years we'll get there, but right now we're really only dealing with narrow domain AIs. While they fail, and I have a recent paper describing exactly how they fail and what we predict they are going to fail at, the damage is limited to that one domain. We don't have cross-domain problems yet, so that's a good thing.

[00:21:00]  
Bill:

[00:21:30]  
You said AGI, artificial general intelligence ... what would that look like ... first of all, when would you see that happening and then, what would it look like in the day to day life of a human being that is a westerner that is educated, how would they experience AGI from a first level? I've seen some of those slide presentations where artificial intelligence reaches the intelligence of a mouse, and then of a village idiot and then shortly after reaching the intelligence of a village idiot, it's super beyond that. Maybe you can jump in there and say where would AGI fit in?

Roman:

[00:22:00] Right, so that goes back to this concept of singularity. Once you get to human-level intelligence in domains like engineering and science, the system is very quick to improve itself, upgrade itself, do original research on robotics and AI, and so very quickly it goes beyond just human level. It becomes superintelligent and at that point, any predictions we make are really meaningless. We can't predict behavior of such a system, how it's going to impact us. In short term, before we get to AGI, yeah, we can talk about technological unemployment, we can talk about things related to labor automation and so on but long term with superhuman performance, we just don't fully understand what to expect.

[00:22:30]  
Bill:

[00:23:00]  
As you look at not knowing what to expect, how do you incrementally try to put layers of protection into this problem? Are you in your research trying to focus on the end game or are you trying to ... for example, like an AI chat bot right now that an enterprise is using or experimenting with, would you tear into that and really look at and research that and where it could go and how it could go awry, or are you more just dealing with the general theory 20, 30, 40 years down the road?

Roman:

[00:23:30] Both are very interesting. We want to make sure technology we have today is safe and our safety mechanisms are keeping up with it, so maybe for chatbots, we've seen examples of them becoming verbally abusive, so maybe I'll have filters in place to make sure that doesn't happen, but obviously the damage is limited. Somebody will have hurt feelings, but that's about it. Long term, the consequences are much more significant. I'm not sure there is actually a solution. I can't say that I think it's possible to control a superintelligent system. What I'm looking at is a lot of tools which allow us to have more time to figure it out, ways to contain such projects, to study such systems, but all of it as far as I can tell is just temporary solutions. They are not really guaranteed to work long term.

[00:24:00]  
Bill:  
Give an example or two about some of the containment tools that you're ... because that's super practical and I would love to hear how you're approaching those right now.

Roman:

[00:24:30] It's similar to how we study computer viruses. We have computer systems which are not on a network which are creating a sandbox, a virtual environment from which the process cannot communicate, cannot escape. That's one of the projects I'm working on, developing this limited communication system, developing multiple layers of virtualization to make sure the system cannot escape and we can shut it off before it manages to infiltrate multiple layers. It's still in the early stages. We're still trying to get additional funding for it. We published a few papers and kind of general directions in it, but it's not a finished project.

Bill:

[00:25:00] That would be contained meaning it wouldn't necessarily have a connectivity to the outside world, for example. It would just be something that you could, like you said, a sandbox to contain and test in.

Roman: Right, social engineering attacks prevents it from sharing information with global community.

Bill: In your books ... Maybe you could share with my audience, what is the social engineering attack from an AI point of view?

Roman:

[00:25:30] The easiest way to break into any system is not to actually find a technical flaw but to find someone who is going to let you in, whether it's a secretary, a janitor, just kind of talk them up into revealing a password or connecting a cable. As those systems become better at just talking to people ... you talked about chatbots ... it becomes easier and easier to gather information, to communicate, to maybe convince people to engage in certain behaviors. With all of the information about you available through your social media profile, it becomes very easy to create a very targeted attack to really get someone to do what you want.

[00:26:00]  
Bill:

[00:26:30]  
Yeah, because it's interesting now, the social element or the the human element is the biggest vulnerability point whether that be just clicking on malware or hitting sites or hitting spam that's coming into the network. It's inadvertent so a lot of money is spent on training the human, so it's interesting your point on a chatbot that might have the intelligence level of a ... I don't even know, where is the intelligence level of a chatbot right now? Is it at 18, 19, 20?

Roman:

[00:27:00] It's pretty low. Most of them are just tricks and they don't really understand anything they're doing so it's not so bad unless you're really talking to people who have no background in safety, security, computer science. That's of course the problem. We see it with standard spam and phishing emails. Interestingly, even additional training doesn't seem to be very productive in reducing this type of danger.

Bill: Oh, have you seen some reports where that hasn't been as effective as I made the statement earlier?

Roman:

[00:27:30] Something like spear phishing attacks, they would go to a company with maybe 100 employees and they run a spear phishing email campaign. Maybe 85 people would click on it. They go in, we'll do a workshop, we'll explain to them don't click on things, it's dangerous. They do it again; this time maybe like 60 people click on it. They do it three, four, five times cycles and still you have someone clicking on it. It never stops.

Bill:  
[00:28:00] Where do you see the research going right now in creating safety mechanisms? Are you at the tip of the spear now or do you see any glimmers of hope as far as helping drive the engineering efforts in a certain direction where we have a tide that rises all boats? Is there a certain research angle that's being explored that may be able to elevate the security and the safety concerns from an engineering point of view?

Roman:  
[00:28:30]

[00:29:00] Luckily, there is a lot of interest and a lot of projects taking place, some are well-funded, really good researchers. The problem is I'm not sure any of it is actually doable in practice. Theoretically, we can talk about trying to understand human values, so maybe implementing some morals into machines. In reality, I'm not sure it would actually work and this is true of all similar projects, projects on software verification have limits. There are limits to verifiability. Projects dealing with just explaining decisions by those systems. Again, there are limitations in terms of complexity a human can understand. I don't think today anyone has even an idea for a working AI safety mechanism which would work long term to make a human level or higher intelligence safe.

Bill:

[00:29:30] Very interesting. I've had different perspectives on this. Some people that I've had on, literally guests that I've had on, said no, we're not concerned. We're going to figure this out as we go and we're going to let the AIs teach us. You have some significant concerns about this and are pretty vocal about these AI safety concerns and having machines teach us how to be safe. Maybe you could share one of these with us.

Roman:  
[00:30:00] It's definitely the case that there is a lot of good people, experts, who are what I would call AI risk skeptics or even denialists, and it's interesting because the arguments are usually either, well, machines will be good to us just because or it will take thousands of years to be a problem. We don't have to worry about it. We'll get to it when it's a problem. All those can be argued about.

[00:30:30]

[00:31:00] I published specifically on creating malevolent AI on purpose. Like people today design computer viruses, there is going to be the same trend to take intelligent software and give it malevolent goals. Now there is no debate about it. I can always argue, I'll do it just to prove a point. Anyone who says it's not a problem, we don't have to worry about it, ignores this fact. Most problems will come from deliberate malevolent design and it's the hardest problem to solve. It contains every other problem aspect so poor design, mistakes in implementation, value alignment problems, all of them are part of this malevolent design issue. This also has additional negative payload and anyone denying that this is a problem is wrong. They're ignoring evidence. Also saying, we'll work on the safety mechanism when it's an actual problem. This cannot be done; it's too late at that point. It will take longer to design a safety mechanism than to design the actual system. It's like saying, okay, let's make a car first and we'll worry about brakes when we're on a highway.

[00:31:30]  
Bill:

[00:32:00]  
It's interesting. I had a guest on and we talked about autonomous vehicles and the reliance on GPS, but what happens if our satellite networks get taken out? What happens five, ten years ago when a big chunk of the cars on the road, or a larger chunk than today for sure, are autonomous vehicles? What is the safety mechanism in place to run truly autonomously without having the guidance from the network?

[00:32:30] It's a design challenge that we don't want to wait until that happens to architect around. Microsoft ran into this problem ... you know, they were first to market and they came out with their code and they launched their code with Windows 95 and moving through their technology stack. Then in 2002, I think, Bill Gates got so much flak, he doubled down and put a billion dollars into his security teams and now they have quite a secure stack of software. I should say they have a huge investment in security but it happened after the fact. We sort of have a history of doing safety and security after the problems exist.

Roman:  
[00:33:00]

[00:33:30] Exactly. I talk about it again in the recent paper, the difference between cybersecurity and AI safety. In cybersecurity, your goal is to minimize attacks. You want as few as possible and you want to minimize damage. Somebody is going to lose their social security numbers, maybe a few bank accounts will be taken over, but the damage is limited. With AI safety, if you're talking about human-level intelligence system, there is no limit to damage. We really cannot say, okay, it's not going to cost human lives or destroy economy completely. It's a very different problem. We're not just trying to minimize number of AI failures. We need to get that number to zero and it doesn't seem like it's actually possible.

Bill:

[00:34:00] Do you watch the TV show ... I've watched it so frequently through the years, and it's going to sound kind of weird if you don't remember it either, but it's on right now. Basically it's this great show currently where this AI is this beneficial AI has been programmed by this very smart guy and it's been unleashed into the world and now it's propagating and it watches everything. It watches the cameras and it's always looking for ... Do you know that name of that show?

Roman: I don't watch TV, so no.

[00:34:30]  
Bill:  
Basically, this AI is ... and it's wonderful, it's well done ... but now the defense department wants a part of it. They want to try to get at the guy that wrote it because they want it for their own uses. It's essentially, it's all pervasive. I think when we talk about human level intelligence, you called it AGI. What does that A stand for for AGI?

Roman: Artificial general intelligence.

[00:35:00]  
Bill:  
Artificial general intelligence ... is it all pervasive across our networks or are you envisioning it still being domain specific? For example, is it just in a robot, AGI, or is it ... really it pervades our algorithms, it pervades every technology we touch?

Roman:  
[00:35:30] It has potential to be everywhere and it's really interesting how it changes a lot of our privacy notions so things which right now no one has time or energy to analyze, a system like that can go back ten years, twenty years, and put together all those different data points and discover things about you and me which we didn't really anticipate. We see it with our politicians today. They go back and they find this video from ten years ago of a guy saying something but this is like manual labor and it's done on a scale of one. Think about a system capable of doing it for everyone and very quickly with all the data.

[00:36:00]  
Bill:

[00:36:30]  
Yeah, it's stunning, very, very stunning. I've been reading this book recently on blockchain. Actually, the gentleman is coming on the podcast. Could blockchain help in some regards or do you think the AIs will be so smart at some point that they will actually be able to crack into blockchain? I'm bringing this mainly up for privacy point of view, so recovering our sense of privacy back from the pre-1980s era. Do you see blockchain at all fitting into your models?

Roman:

[00:37:00] It's a tool. It's a cryptographic tool. Like any cryptography, it could be useful to increase security. Again, AI is not magic. If there is a computational barrier to decrypting data, it will be there. How we use it with AI is an interesting question; it depends on what you're trying to do with it.

Bill:

[00:37:30] Well, I think one of the pieces is most interesting, and you and I have talked about this prior, is that the exponential technologies give us lots of benefits and I think it's really incumbent upon really smart people like yourselves and myself who bring people like you onto the show, that the message is as much about governance as it is about how to use this for positive benefit for humanity, businesses, personal, et cetera.

[00:38:00] One of the blockchain pieces that I really find appealing is the ability to recover privacy and a sense of the ubiquity of the internet came with it a whole bunch of positives, but we lost anonymity. Sorry ... everything is anonymous but we lost security and privacy, so recovering that is what I really hope that blockchain potentially can help. It also dismantles the middle man and so it makes it very ubiquitous because you essentially can take out intermediaries.

[00:38:30]

[00:39:00] When we talk about AI and we talk about blockchain, it's more the smarter these data aggregators are being able to pull data together about us, but then if we can take that data and actually obfuscate it again and actually make it private and make it ... If I have a wallet that is specific for Bill Murphy, and a certain part of my wallet I can give to Amazon, another part of that wallet I can give to the IRS, another part of it I can give to my employers, so I don't have to give my entire identity. Possibly there is a way that AIs can help participate in a beneficial good related to that. It's more taking the AI and saying this thing is going to be so darn powerful. Maybe it needs to figure out parts of these problems, but if the bad AI then comes in, it can actually break the blockchain, then we're back to where we started.

Roman:

[00:39:30] There are definitely benefits to this technology. My concern is that cryptocurrencies make it very easy for NAI, which manages to get access to internet, to get access to this financial resource, quickly accumulate it and then be able to use that money to pay off people to do its bidding. We've seen people use cryptocurrency to hire killers, to get access to additional server resources. That's my concern. I'm always looking at the safety and security aspects of every technology.

Bill:

[00:40:00]

[00:40:30] This is really important that this be done because it's almost like there is going to be a second arms race in some respects. As many people are looking to positively exploit the uses of these technologies, there is going to be the dark side of this as well. It's interesting your angles on this. As we wrap up our conversation, Roman, what message would you give our audience? There's going to be entrepreneurs listening here, very technical folks all the way up to the CIOs and sometimes board members that are listening to the show to really see where the future is going. How would you guide them through their decision making for more of a strategy point of view as they look at the world?

Roman:

[00:41:00]

[00:41:30] Short term, especially with young people, with students, I always suggest before you commit to any degree, any major, see if it's going to be around in ten years, fifteen years because a lot of it is automated, replaced. If you're starting a new company, same logic applies. Is the product you're developing something which in ten years will be done by machines automatically? Will you still have a competitive edge in that. In terms of human level AI and beyond, there is really little we can do. An individual doesn't have much say in that issue. If you do get to participate in a political process, make sure your representatives are aware of those problems. We had some success with that. President Obama just last week talked about AI safety, so it's very encouraging to see that level of understanding from the White House. We need to make sure they not just know about it, they are willing to fund this type of research.

Bill: That's excellent. How would people reach out to you regarding AI safety? Are you active on social? Where would you prefer to engage with people?

Roman: Yeah, I'm always happy to engage. Facebook, Twitter always works really well. Just feel free to follow me on both.

[00:42:00]  
Bill:  
That's excellent. We'll definitely put this on the show notes, Roman, so people can reach out to you and listen and follow your work as we go along here. I very much appreciate your hard work in these subjects and bringing these important issues to our awareness.

Roman: Thank you so much for inviting me.

Bill: I appreciate that. Until next time, Roman, thank you very much for coming on the show.

Roman: Thank you, Bill.