#185 - Ethics and Artificial Intelligence (AI)

[00:00:00]

[00:00:00] **Speaker:** Hello and welcome to another episode of CISO Tradecraft, the podcast that provides you with the information, knowledge, and wisdom to be a more effective cybersecurity leader. I'm your host, G Mark Hardy, and today we're going to be talking about ethics and artificial intelligence. If you're watching us on YouTube, great.

Don't forget to subscribe. And if you're following us on your favorite podcast channel, make sure you share it with other professionals that you know, so we can help them with their cybersecurity journey. Now, when we talk about ethics and artificial intelligence, at first, it sounds like two disjoint sets, drawing it on a Venn diagram they would not necessarily meet, but when you think about it, is that AI does have a lot of ethical implementations, problems, and characteristics that we want to look into.

And [00:01:00] that's what we're going to be talking about today. So, let's think about where we are in AI. There has been some suggestion that there are multiple stages in AI. I'll put the reference links in the show notes, but I found a YouTube episode on the evolution and the seven stages of AI, where the first one is rule based AI systems.

And we've had those for years. If then else, if then else, if then else. The second stage, of Context Awareness and Retention Systems, which is really where we are with a ChatGPT, suggests that it's able to go ahead, take information from where you are right now, incorporate other data, and come up with some sort of a predictive outcome.

The third one then will become Domain Specific Mastery Systems. We're working that way, but be able to take absolutely all the knowledge about a particular domain, And be able to produce outcome such that we have a true expert system. Imagine, for example, all the medical knowledge that goes in there and being able to give correct results and not some things [00:02:00] such as how to eat glass.

And some of the other goofy things that we might've seen come out from ChatGPT lately. We then get to stage four about thinking and reasoning AI

systems. This gets a little bit scary when we find out that. What we've talked about today for generative AI, it's a long way from thinking. It's really just predictive models, and we'll get into a little bit of the math in a bit, but really something that could essentially think for itself.

And this is getting into the realm of science fiction, but we keep going. And it's stage five where things get really frightening, perhaps for humans anyway, is what's called artificial general intelligence, or if you will, the birth of a new mind, something that becomes self aware decides that it can go ahead and make decisions sort of like an iRobot movie and the like.

We go beyond that to artificial super intelligence where we have Machines that can outthink humans. They can do better than we can. In fact, they don't really need us anymore. And then you start to worry about things such as, are we getting into the realm of looking [00:03:00] at the Terminator? And stage seven, the AI singularity, where essentially it has eclipsed mankind, and then we become sort of a footnote.

Well, we're, of those seven stages, we're only at layer two. And that means then that if we're going to get to the point where these machines are going to be. As smart as us, and then smarter than us, and then potentially become self aware, and the like. We better get it right this time. There's not an opportunity to go back after this has progressed beyond a certain point and redo it.

Now, there are ways that humans can interact with AI. We should be deciding. What we can ask and how, and we see, for example, a lot of the guardrails that are put on generative AI. If you say, show me how to build a Molotov cocktail, it'll probably refuse that. But if you tell them how to build a cocktail for the bar, you'll get answers for that.

And yet we write AI escapes and we find out that with a little bit of clever wording, you can get around these guardrails [00:04:00] and be able to produce results that the designers didn't intend. That, of course, is the essence of hacking. But other things we are just starting to think about is legal compliance.

The regulations around AI are not really there yet. I'm going to talk about that in a bit to find out what should governments be doing. And then AI companies, are they're posturing themselves in a way to say, hey, let's go ahead and do these things, which sounds good. But the reality is they might be cross purposes.

And then data security, something that we as CISOs and security professionals really need to concern ourselves with. Because as all this information goes into

these models, Does it stay? Does it get trained on? Does our proprietary data end up being part of a generative AI training model? At which point it may come out in somebody else's query.

And we've heard some stories about that. And the concern there is that AI could potentially be one of the biggest security and privacy [00:05:00] problems of our time. If we don't do it right. And then ultimately it's important that we understand that these models Currently is Instantiated has some weaknesses.

We're not there yet. We've made great progress. We looked, I mean, especially if you've been owning things like Nvidia Stock, you've done really, really well. Congratulations on the $3 billion Club guys. Um, full disclosure, I don't have any shame on me. Alright, but what about business models? Do we think about business models around artificial intelligence?

And from, so what do we protect in there? The input, the data that becomes something we have breached notification laws and they extend to a lot of computer systems, network systems, et cetera. But from a perspective of an expert system where it's all in there in a training data, it's a one off, it got ingested from something. We don't know necessarily when we have a model that has billions of tokens or trillions of tokens where that [00:06:00] actually came from and being able to identify to say, Hey, we're gonna do some discovery on this. Where did it originate? How do we trace it back to its source? What is the data provenance of what goes into AI?

In a way, there's an ethical implication on that because if that gets corrupted, if we train it on bad or incorrect or false or misleading information, our models will be bad or incorrect or false and give us disinformation. And that's a real danger that we face. The access controls to being able to not just control access to the AI, because today it's just a matter of go to ChatGPT 3.

5 or cough up your 20 bill and get ChatGPT 4. But what happens when these artificial intelligence systems, are agent based, these are essentially agentless type tools in that if I ask ChatGPT a question, I get an answer, but I can't get it to go ahead and unlock a door or raise a garage [00:07:00] door. or change a chemical formula in the process online.

But we're getting there. We're getting to the point where we could go ahead and interact with the outside world and the access to those systems are going to be critical. As we look at the models that we build, we have to go ahead and harden these things so that they can't be easily corrupted. One of the opportunities we

have with open source artificial intelligence Tools and Software and Models is that we can train them, well, any way we like.

Which solves the privacy problem that we talked about, in that I don't have to go to a general artificial intelligence model to get my answer, and worry about coughing up the data. I noticed recently that at least ChatGPT 4. 0, because I do pay for the subscription, allows me to do the equivalent of an incognito window, where it says, we will not train on your data.

Promise. Scout's Honor. Okay, hopefully they're true on that one. But then the concern is, is that it doesn't get smarter. So perhaps in exchange for a 20 bill, they're willing [00:08:00] not to get smarter at your expense. But, I don't know. I have no way of checking that. And so what we have then are ways that we want to go ahead and put guardrails on our AI systems.

Not, Ex post facto, after they've been deployed and put out. It's the way that security has been done for a long, long time. You have a business requirement, you're in a hurry, there's a rush to market to get it out, and we'll get security right later, if at all. As compared to doing it correctly, is getting security baked in at the very beginning.

And that's a real challenge as we're building AI, is that often there's a rush to market. And we often think of security as friction, something that will slow us down. And one of the things we want to think about that from a business perspective, as we look with AI, is how do we go ahead and improve our likelihood that we're going to get it right?

That involves putting in your controls up front. [00:09:00] There's a great lecture that I found also on YouTube on the most important trends 2024. And it's kind of a reality check that gives us on what it is that we can and can't do. We're going to be taking a look this year to say, Is AI going to be able to deliver the things that we expect?

There's an awful lot of expectation and hype. We've seen stock valuations go through the stratosphere with that expectation that we're going to get there someday. My experience over my years is that much like the Gartner hype cycle, we're right about here. And then what comes next is the troth of disillusionment, which is not to say that I'm suggesting you short anybody's stock, but I'm a bit concerned that the momentum that we've seen in the past might not continue and not due to any fault of the businesses, but just to a realistic resetting of our expectation.

Being able to multi modal AI, being able to get involved in different areas, smaller models. Being able to go ahead and get [00:10:00] it down to a device. Imagine if you had a self contained AI engine with the right number of tokens on your cell phone. Think of the power that you could have there at your fingertips.

You already have more horsepower on your cell phone than the entire Apollo program had getting a person to the moon and back in 1969.

Imagine where we can go in a few more years. One of the big costs with regard to being able to do AI correctly is the cost of the GPUs, the graphical processing units. And GPUs, as you know, came from doing video. Why? Because you've got all those pixels. If you've got a Screen, my monitor here is like 3, 000 pixels across by, I don't know, 1980, so there's five, six million pixels.

Every single one of them has to be told. Are you going to be blue, uh, or yellow, or are you going to be green? And then we have, I'm sorry, blue, red, and green. Get that right. And then mix those into the right colors. [00:11:00] GPUs turn out to work really well for parallel processing. It's why they've been very successful early on when we're doing things such as cryptocurrency mining, because you do the same transaction around and around, and also for AI.

So the GPUs that are coming out from companies like NVIDIA, big special purpose things, a lot of parallel work, but expensive. Doing it in the cloud, expensive. There may not be enough electricity available in the world when you compete against crypto mining to go ahead and build our AI models, because we also have to keep our houses, Lights on and the air conditioning running, et cetera.

So there's some energy challenges in here as well. Model optimization is important. Being able to go ahead and trim out the things that are not relevant to your business. Output that you desire. So you don't train on things you don't care about. Again, a way to streamline our models. We can build models locally, custom them.

The idea of a virtual agent, something that is able to connect and carry around with you, I think is pretty great. But we've got other things concerned about. [00:12:00] Regulation is not catching up. There's some talk about it, but I don't think that the lawmakers who are responsible for building the regulations fully understand the implication of AI.

And it's not really their job to do so, because they've got a lot of other things that they're working on. It's really kind of our job to inform them. And to be able to communicate this information to our elected officials to make sure they make correct decisions. And that does not mean that you have to go ahead and register as a lobbyist, but you are certainly entitled to contact your elected officials.

And quite honestly, they're not the people writing the regulations. They don't write the rules. It's the staffers working all kinds of crazy hours for relatively low wages doing it. And if you can make their life easier, you got to win. What about Shadow AI? We've heard of Shadow IT. And as CISOs, we worry about that when you have exposures to your risk profile and you increase your attack service by people incorporating apps and feeding in information [00:13:00] and you don't even know they're there because they're paid for with 74.

99. Taxi fare that didn't require a receipt. And then someone just puts that bill on their monthly expense report. And of course, with a shadowy eye, the concern is what? That somebody is going to leave the company and the bill doesn't get paid and this thing shuts down. And it became a critical process because everybody had been using it.

They didn't realize it was procured through straight AI or IT. Imagine that same thing happening with AI as well. It becomes an interesting concern. And then, of course, the most important trend may be one that we haven't even figured out yet because we're not even halfway through the year. But all these different things that we look out and all these different pieces of information don't tell us what we should do with AI, or more importantly, perhaps, what we should not.

So as a result, welcome to the world of ethics. Now when we talk about ethics, let me, first of [00:14:00] all, tell you what ethics is not. It's not the same as feelings. If all you do is make my decisions based upon my feelings, it's a fairly immature approach toward life. It's not the same as religion. There are ethical implications that someone assigned to a religion, but it's not a religion in and of itself.

There's no deity, for example, in ethics. It's not the same thing as following the law. Laws are made by humans. They're not perfect. Anybody who's ever gone to court understands they are not courts of justice. They're courts of law. And what might be a very unethical and unjust outcome is can occur within the rule of law.

It's just how things work. And ethics is not the same as following culturally accepted norms. I'm doing what everybody else is doing, therefore it must be right. That groupthink could be dangerous. Think of Germany in 1938. That's a very dangerous type of a [00:15:00] groupthink. And yet as a result, we say that wasn't ethical, but it was not following culturally accepted norms because what was okay back then was different.

It's not science. But even so, feelings, religions, laws, culture, and science are part of being human.

Essentially, ethics are a framework for our behavior as humans. And there are many different ethical frameworks that have come out over the years, and I'm going to go into five of them as examples, but there's a lot more. But essentially, we take these ways of thinking and these ideas and these concepts and then we'll come out with some actions.

We'll look at the idea of rights or duty, a deontological approach for ethics, justice and fairness, a utilitarian, common good, and for those who are Greek philosophers, virtue. So let me start with a concept of [00:16:00] rights based ethics. Think of the Declaration of Independence. We have certain unalienable rights.

We have certain unalienable rights. Life, liberty, and the pursuit of happiness. Doesn't mean you'll get it, but you can at least pursue it. And Immanuel Kant wrote about this, and it turns out that Kant was born in 1724, and so Declaration of Independence about 1776. You're pretty much right on schedule for that, thinking to be Front and center for the U.

S. Declaration of Independence, but we talked about a rights based ethics What is a right? A right is a justified claim on others Now Immanuel Kant said that each person has a dignity because they're a human and that dignity shall be respected Now, that creates both positive and negative rights. A positive right, or a welfare right, is basically a claim on others to assist you.

Doesn't mean getting a welfare check necessarily, but it could be a claim on society, such as the right to an education. Other people have to provide you with the [00:17:00] education, but we say from a rights based perspective that you have that positive right. There's also the concept of a negative right, which suggests that there is a negative duty.

That is to say, a right to privacy. Leave me alone. That's something that I could go ahead and expect out of a rights based education. environment. Now, what

do you do when rights conflict is where things get interesting, where you say, I've got a positive claim and a negative claim. And that's why philosophers get to philosophize a lot, trying to think about what happens on these border cases, boundary cases, and when things conflict, if we go back another 20 centuries or so, actually 24 or 25 to Plato and justice and fairness ethics, it's a basic element for a lot of public policy, justice, fairness, because Plato had his concept of the theory of forms that ideas is. Are timeless.

They're [00:18:00] absolute, but things are ephemeral. This is a thing. It exists now in time, but there's an idea that exists that's always there, and as a result, if we can understand these series of forms, those that always exist, that's what we should aspire to. Now, the principle of justice will resolve conflicts of interest equals should be treated equally.

Also, unequals should be treated unequally because it doesn't necessarily force everybody into an absolute egalitarian state. But you should be able to approach others with the same level that you brought if they're at about the same level. Now justice can be distributive where I have what we call shared burdens where I allow other people to do things like paying your taxes.

Corrective justice like the criminal code. You were wrong, you go to jail. Or compensatory. Something went wrong and you need to make up for it. Well, ambulance chasing, if you will. But as I said before, all that justice gets dispensed through courts of law. [00:19:00] But they may not be ethical in terms of the outcome from a law perspective.

So what's another way of thinking about an ethical concept? How about utilitarian? Jeremy Bentham wrote about this back in the late 1700s, early 1800s. Morally right action provides the most benefit. Benefits to the most people with the least harm. So if you do the most benefits to the most people with the least amount of harm, that's a moral and ethical thing, according to utilitarianism.

Now, what happens is, to figure that out, you assign the benefits and the harms to a course of action, you chart it all out, you add it all up, and you see how they sum up. Notice that in the utilitarian perspective, it's the ends, not the means, that are important. Meaning you could lie or steal, but as long as you're reaching an ethical goal.

That allows you to go ahead and justify your actions. And as we look at world events today, we can see some pretty nasty means that [00:20:00] have potentially a just end in mind, but do the ends justify the means? And that's of

course the key question here against utilitarianism. And then what you want to think about is can people using this framework claim that they're going after a moral outcome Yet using immoral mechanisms to get there.

So not all of these concepts are necessarily perfect. There's a concept of the ethic of the common good. Common good, which are in the writings of Plato, Aristotle and Cicero. Well, what is common good? And Pope Paul VI gave us a good definition back in 1965. The sum of those conditions of social life which allow social groups and their individual members relatively thorough and ready access to their own fulfillment.

It's your common good to public health care a common good. You may or may not need it, but on the day you [00:21:00] do need it It's nice to know it's there. Public safety adds value across society A just legal system adds value across society. That is a common good. The danger with the common good is that if everybody's chipping in to make something happen, you've got the free rider problem.

And the free rider problem complicates that. For example, there's a drought, and so therefore we're all going to not water our lawns to conserve water. And someone can say, well, all that water is being saved, there's plenty for me, so I'm going to go ahead and wash my car and water my lawn. And you can't do much about it.

Or, you go ahead and said, everybody contributes, they chip in a little bit as a donation. I don't have to donate, I just contribute. Good along for free. That's a free rider. So it creates unequal burden sharing and thus is the danger of the ethics of the common good. It's really the reason communism just never succeeded is because when you tell everybody, you're all going to get the same input [00:22:00] and the same output.

People don't do that. They say, Hey, I can get the same output and not have to do the same input. If it's good, I'll sit back and collect. And last ethical framework I wanted to look at was that of virtue. And we're going to go back to Aristotle back in the fourth century BC. That moral principles are based on ideals, the concept of excellence, or arite.

It's, you express your virtue, your ethics through virtues, which becomes a universal standard of excellence. Things such as honesty, courage, compassion, generosity, fidelity, integrity, Fairness, Self Control, Prudence, items such as that. You see, there's no objection to that, but that then becomes the expression

over the years because you improve virtue through practice, according to Aristotle.

And our distinctive [00:23:00] function is reasoning. And if you have a life that's worth living, it's one in which you reason well. So develop your character and live a moral life. Okay. Enough with the ethics lesson. What does that have to do with AI and artificial intelligence? Well, to quote a friend of mine, Wynn Schwartel, who said, we created AI in our own image and we don't like what we see.

Why? Wynn is working on a series of lectures. I've attended one of them and I'm going to get him back on the show. I've had him before and he's just a brilliant thinker. And he's talking about his thesis on metawar or basically how AI can go off the rails. And in a nutshell, as we look at, the impact of artificial intelligence and how it can change our thinking as a society and change our action and behavior.

He suggests that as humans, we start out with stories. This is why we still have the Iliad and the Odyssey, because it was a story that was [00:24:00] told even before there was a written language to go ahead and record it. People would memorize big, long elements and stories and things like that, because humans think in stories.

That's why I say, instead of doing a password that is uppercase, lowercase numbers and special characters. Use the length. Tell a story. The little girl went to the store to buy some food for her cat. NSA can't crack that password. A little divergence on security there, but you get the idea. We think in terms of stories.

But from there, we want to make it an immersive experience. So we had first, what, books we could read that, and then the talkies, the movies, and then all of a sudden we're getting to the point where we have this immersive experience where I go to watch a movie and I lose myself in it. You come out there and you suspend reality for a couple hours.

But now we have reality distortion where we can focus the audience's attention and we can take you into things and make you believe things that aren't real, because for the sake of. Being part of this immersive experience, we suspend our disbelief, which is why science fiction and a [00:25:00] lot of fiction movies work, because, well, it's, we'll go along with it.

But then at that point, because we've got our attention focused here, we can have a reality confused with too much information, a real problem, the TMI, or

disinformation, which is deliberately designed to mislead us, which then creates the manipulation of our emotions. That we could be persuaded and influenced to go ahead through strong emotions that are being inserted to us by this misinformation, disinformation, immersive experience to cause us to think and do things you wouldn't normally do otherwise.

And as humans, we need rewards. We respond well to rewards, as well as BF Skinner pushing a little button on there for a mouse, or a rat, or his humans. You do something, you get rewarded, you do something, you get rewarded. Potential danger with something like a TikTok. The algorithms in TikTok, brilliant! To the point where they've realized that you could go ahead and provide this little Digital dopamine feedback when somebody likes something and they get something else that's just [00:26:00] like I wanted.

And so we're doing this feedback loop where we're going like dopamine, dopamine, dopamine, next scroll, scroll, scroll, scroll, scroll. Creating an addiction, digital addiction. Some of us have lost loved ones to social media and now potentially into AI as it creates more and more of this immersive environment.

And then at this point, when we've got everybody hooked, we can get their compliance. We can get them to do things we want them to do. And ultimately it creates a sense of belief. This is where people go ahead and they're totally convinced that this is the truth. Now this is Wynn's thesis in a nutshell.

He's writing a whole book on it. It's going to be brilliant. Hopefully he doesn't mind a little bit of sneak preview here, but it gets you thinking along the lines. Now if that's a concern that we're gonna go ahead and we're gonna create humans and we're gonna enslave them into this mindset through AI, how do we get there from here?

I mean, what is AI? To a lot of people AI is just sort of, well, we call a black box. You've got inputs, you have outputs, [00:27:00] but you don't know what happens inside. It took a long time to program that stuff and we sort of kind of think we know, but unless you're programming it, you don't know. But even those who built the models were surprised at what they could do.

And they were getting outputs. They said this thing is inventing molecules. It's inventing medication. It's doing stuff we never even programmed it to do. How is it so smart? How is it able to, is it taking off? Is it moving past stage two in our seven layer stage to the next? Not really. Because what happens with AI is that it's only going to be as good as the information you train it on.

And the old saying garbage in, garbage out doesn't apply anymore. Now it's garbage in, garbage stays. And it gets pregnant, and it gives birth to triplets. If you put bad information, disinformation, do bad training on your model, you're going to get bad output. And it might be subtle, you may not realize it, but there's been a couple things where I've seen posts where people [00:28:00] asking if you could go ahead and have generative AI say, can you do this thing, which is Stupid medical things or the like.

And people go, Oh yeah, absolutely. You can do this and this and this. And I'm like, no, you can't. And it turns out that some cases it was trained on the onion. Or some other satire and the like. They can't tell the difference, it's just inputs. And some of these things might be too subtle to notice at first.

Some people are worrying about AI from, it's unethical, it's going to take away my job. You know what? AI is not going to take away your job. Somebody who knows how to use AI is going to take away your job. So you better get on the bandwagon, figure it out. And some are worried about AI destroying humanity.

Now Skynet was science fiction, but there's a non zero probability of this. That we might have a Terminator type future. In fact, Elon Musk estimates it at around 10 to 20%, but he said it's worth the risk anyway. [00:29:00] I don't know about you. I fly a lot, but I don't think I get on an airplane that had a 10 to 20 percent chance of crashing.

Yet we're talking about doing that for a whole society. That's a big gamble. The idea of deepfakes, we're seeing more and more of them to the point where it's going to potentially change people's perceptions of elections. And I've heard it said that 2024 will be the last election decided by humans. In the future, everybody's going to be manipulated so cleverly, so precisely, that you won't even know why you reached that belief, according to Wynn's meta war thesis, but you'll believe it.

Now, if you're a scientist, and you have a belief, it's based upon experimentation, and it should be reproducible results. And if somebody goes through and produces a different set of results, and brings the scientific evidence to say, you made an error here, this is the correct, things like that, I get it.

You change your conclusion. But if your [00:30:00] beliefs are not based upon a scientific principle, but you believe it because it's an emotional foundation, or somehow you've been fed these things by some AI enhanced process, where people have figured out how to get you to think the way they want you to think,

you're going to defend those. Why? It's not so much that your ideas are brilliant, But you defend them because you have no way to back them up. And as a result, you want to admit the fact that you don't have the scientific proof for it. So that's a general ethical dilemma that we face. So challenge yourself.

If you believe in something strongly, this politician is a jerk or this politician is brilliant, or this is true. Why? Where's your evidence? Is it firsthand? Did you actually meet this person? This person cheats you out of something in a business deal? This person stiff you on a dinner check? They punch you in the nose?

If not, think carefully. Someone is trying to change your perception of things. Now AI also, from an ethical perspective, is [00:31:00] Can you trust it? I mean, if you ask AI, a generative AI question, you will get an answer. It's like asking a seven year old sometime. It just might not be the right one. And we've seen situations where a lawyer last year went ahead and came up with a whole bunch of citations on a legal case.

through ChatGPT, and they didn't exist! And the judge ended up sanctioning the lawyer and his law firm. And in his defense, the lawyer said, I just thought it was Google and steroids. No! It trains on a bunch of stuff. But these hallucination episodes, that they call, will continue to occur. It hallucinates because it tries to fill in the blanks.

It tries to give you an answer based upon the input, although the input may be flawed or incomplete, and yet it's programmed to go ahead and generate an answer. And so as a result, we're always told trust, but verify, and ChatGPT has a little warning down there at the bottom of the screen telling you it might be making up the answer.

Now as we look [00:32:00] at these tools, these large language models or LLM like ChatGPT, by the way, here's your 500 point Hacker Jeopardy question, what does GPT stand for? It's a generative, pre trained transformer. You're welcome. It trains on large database, petabytes of data, huge amounts of information, and we first of all, as we train this thing, allow what's called unsupervised learning.

It's unstructured data, Just scroll through it. It's like a big box of stuff, but you start to identify patterns. Hey, that looks like one of these, and this looks like one of those, and so in letting it to do the unsupervised learning, it starts to draw associations between different disjoint things and say, well, that's common to that, and it's common to that.

A lot of horsepower, a lot of memory, a lot of And then after that we do supervised learning to correct [00:33:00] some errors in labeling the data. You know, that was not quite correct. You got that wrong. Okay. This is, uh, different than that. And then we'll create a transformer part to assign weights or scores to the data.

information, which we call tokens. And then the predictive models are going to assign the most likely relationships pretty much going token to token to token, or hand over hand doing one word at a time. Essentially, it's all a mathematical model. And if you get the numbers wrong, you get the model wrong.

And when you LLM, you pre train it at a high level, what you're going to do is feed in all of your text, all these petabytes of structured data. Unstructured data. Start to tokenize some of this information, and then you begin to find out that after this word often comes this, but after this comes that, and after this comes that, and you start to figure it out.

So let's try it. If I give you a word, give me the next word that comes to your mind. Peanut. Most of you probably [00:34:00] thought peanut butter. Okay, well if we want peanut butter, give me another word after that. Peanut butter. Not Anjelly, that's two words. Peanut butter and. Okay, peanut butter and jelly. One more.

Peanut butter and jelly sandwich. And what we find out is that as we go through these things, it's just picking the most Likely output. Now, it could be peanut allergy or peanut perfume. Pretty weird things, but as you tell it to try again and again, it's going to try lower probability associations.

They're non zero, but they're low probability. And as a result, when we get a generative AI output, it's higher probability things. We say, that makes sense. But this is why we think AI is creative. It's not creating anything. We just traded on such a large volume of information. It's going to generate very likely and less likely and then not so likely correlations.

And the obvious ones are likely. We like what we see. This looks pretty [00:35:00] good. But the not so likely correlations have no meaning. They're just garbage. But somewhere in between or right at the end, occasionally you'll stumble across something that goes, Whoa, that actually does make sense. And what we mistake for creativity is really just a string of low probability correlations that nobody has ever bothered to come up with yet.

Well, what's unethical about this? You see, generative AI is going to answer the way it trains. And it's been trained to lie, to trick humans. But what they found out with some of these models where they've gone ahead and trained it to lie is you can't get it to stop. It becomes embedded in it. Remember, we had the pre training first, where you had the unsupervised learning, and then you do the supervised learning, where we tried to do that.

Well, because it's permanently rewired, if you will, the connections of the DNA or the weighting in there, you can't straighten it out. It doesn't change. Now, we could set poisons into these AI [00:36:00] models, little triggers that will be based on time or certain behaviors, such that This reinforcement learning and the supervised fine tuning and adversarial training will be resistant to this because we've had something to say, Hey, if it's 2024, answer this question this way, but if it's 2025, answer it another way.

Now, all your testing that you do is in 2024. You don't notice it. Anything goes out, it comes out in there, but it's like a time bomb. Sort of a novel supply chain attack, isn't it? Plus there's other AI related ethical concerns. For example, privacy. What about all this big data? There's a huge amount of concern that right now people are talking about.

Well, you trained on my proprietary information, or it was copyrighted or trademarked, and now we have artists suing Microsoft and generative AI organizations saying you train on this stuff without our permission. Well, you put it on the internet for crying out loud. What did you expect was going to happen?

In any case, we'll see how that works out in the courts of law. [00:37:00] Bias and fairness. I want to mention a couple of things that concern like their bias in hiring or law enforcement, even loan approvals. We find out that there can be some distortions there based upon the information we train on. How about accountability?

When an AI makes the wrong decision, who stands up and said, I will take the hit for that. I will pay the fine, or I'll go to jail, or worse. No. Blame it on the computer. Didn't that happen with Air Canada where the chatbot, someone had said, Hey, I got a bereavement fare. I got a problem with my family. So with death, I need to get, okay, yeah, you can have a refund.

And that wasn't part of their policy. And Air Canada said, well, we're not going to give you a refund. But your chatbot said it would and it's speaking on your behalf. Well, we disavow that. Went to court, they're kind of lost. I don't know

why you would spend tens of thousands of dollars in a courtroom and come across as being an embarrassing organization like that.

Went for 700 bucks and it was only 700 bucks Canadian. You could have just made the client whole. And then how about [00:38:00] transparency and explainability? You go in and you get a decision from a system. And it doesn't give you the outcome you want. I got denied, I got turned down, why didn't I get this job?

Well, we don't know, the computer just said no. Or why didn't I get this or that or the other thing? We can't explain it anymore. It's not a human who can sit down and say, Well, let me tell you what my criteria were and why this went this way. We're losing that. And of course, job displacement and the concern about people saying entire classes of jobs are going to go out of style.

But if you think about it, think of all the jobs that we have today and what percentage of workers didn't exist 40 years ago. You'd be surprised. There's an awful lot there. And I found an article on May 29th. It was from Tristan Harris and it was published in The Economist and it's behind the paywall, but I pay for it.

So I can share a couple of ideas and things like that. And what I found out interesting in this particular article is that he is talking about the concept about the transformation of society, [00:39:00] that social media transforms society. And if you go back in 10 years ago and look at the taglines of different social media companies, Instagram, Capture and Share the World's Moments, Facebook login page.

Facebook helps you connect and share with the people in your life. And Twitter was, what's happening? Right? Well, we've found out that over 10 years, social media has gone a lot away from, Hey, what's happening to coming to have a little bit more insidious impact on people. Charlie Munger, who is a late partner of Warren Buffett said, if you want to.

See what's going on. Show me the incentive and I'll show you the outcome. So what's the incentive on social media to capture and retain your attention? Get the eyeballs, get the clicks. Why? So you can sell advertising, we could go ahead and get more mind share. But what's the legacy of that? Attention spans that have been shortened down to only a couple seconds.

Go back and watch movie scenes from the 1940s. You might have a one minute scene. Today you're lucky to get five to six [00:40:00] seconds. Click, click,

click, click, click, click, click, click, click. Political discourse has turned into a matter of just outrage on one side and the other. People with loneliness, anxiety.

This is the legacy that we've got from the way we've allowed social media to develop without really putting controls on it. And the part of the reason is the incentives, what they call perverse incentives, where the things that benefit you for putting something out can actually be harmful to the people who are consuming it.

You want to maximize the engagement. You're creating addiction in our societies and they're being distracted and they want to be, you know, Outraged and they want to be validated and things such as that. And what Tristan says is that social media is really our first contact with AI. And as Winn Schwartau has observed, We didn't do very well with it.

And we're not going to get a chance to do it right the second time if we screw it up. We got to get generative AI, our second contact with AI, correct. And we're starting to see already how it's being [00:41:00] misused. Financial scams against the elderly with voice clones of loved ones sounding exactly like grandma.

I need some money Weaponizing against teens by having apps that go ahead and allow them to digitally Disrobe them and go ahead and threaten to show that to other people or deep fake audio for blackmail and doing business email Compromise and we've heard of organizations that if people don't they Zoom call and all the other characters are fake, but it's so believable.

Write a big check.

Generative AI reduces friction and it redemocratizes the ability to do bad things with these tools such as creating scams, inflicting damage on other people. We could build images instantly using generative AI instead of having to be a master of Photoshop and manipulating and changing and modifying stuff like that.

You just click and then boom, there's your image and it's believable. And we could create infinite [00:42:00] streams of totally believable content, targeting messages very precisely based upon our knowledge of others to influence them and change their beliefs. There is a huge role for government to intervene. AI executives will go ahead and proclaim that need to Congress and then perhaps go back and write checks for a lot of money to go ahead and lobby against those changes.

We did a wait and see with social media and I submit we can't afford to do the same thing with generative AI. Wow. So this is not to be gloom and doom. I think there's a lot of things we can look forward to with respect to artificial intelligence, but if we look at it from the perspective. of are we being ethical?

Well, what's our result has been. We've seen things such as algorithmic bias, facial recognition, being much worse for dark people, black people, people of color than white because they train on white people. Um, Amazon hiring system, they went ahead and they said, we'll train on 10 years of applicants. You have these [00:43:00] millions of applicants.

We hired these tens of thousands of people. Let's go ahead and turn it loose. Turn out that you go back 10, 12 years or whatever, a lot more white males are applying. Not as many minorities are women, and so it had a built in algorithmic bias based upon the data it was training on to go ahead and prefer white males.

These weren't deliberate violations. They weren't designed to be unethical, but they happened. And as a result, we could end up with these inadvertent things, such as loan approval. Now, if you're a banker and you're approving a loan, what do you want? You want the highest probability of repayment and the lowest default rate, and do it at a profit.

So what do you do? You only give loans to people with an 850 perfect credit score. Well, what about the other 98. 7 percent of the population? Well, there's a little bit of risk involved. And so you try to go ahead and train your model to go ahead and reduce those risks. But if you have historical inequities in your data, you're going to get future inequities in your decisions.

And you're going to [00:44:00] penalize people based upon those generalizations, regardless of their individual credit worthiness. For example, zip codes have been used as a proxy For, well, racial demographics. And historical evidence has shown that there's been higher interest rates assigned to inner cities or other neighborhoods, typically things like a black neighborhood or a Hispanic neighborhood, as compared to a white neighborhood.

But zip codes are not a monolithic representation of socioeconomic status, of their ethnic backgrounds, or even creditworthiness. So you got to be sure, for example, in that area that we're not using these algorithms as a proxy or data harvesting all the data that's required to make a generative AI model is huge.

I don't think the average customer understands the implication of being part of that data set. And can you opt out? We've seen how difficult it has been to opt out of things such as a. Facebook or whatever, because this data is not stored just for you. They don't have a bucket on there. We'll say, Hey, that's Joe's hard [00:45:00] drive right over there.

You know, let's take it out of the array and throw it out. It's spread throughout there. And as it came out with these right to be forgotten in GDPR created fantastic Logistics problem, digital logistics anyway, for how do you remove somebody from a system when they're splattered and scattered all over the place.

And then of course the PII, the Person Identifiable Information, if it gets trained into the model, it might be leaked out of the model. Someone might be able to tease that and get that out. And so as a result, we're going to find people blaming the computer. And even companies like Zillow, a couple of years ago, lost over 500 million by trusting the AA model to go ahead and buy houses and overpaid on an awful lot of stuff trying to flip houses.

And they had to lay off almost a quarter of their staff. So, a lot to think about. I'm at the end of the episode, but I hope I gave you a little bit of ideas in terms of ethical frameworks. Some of the dilemmas that we face in artificial intelligence and things that go sideways. And I want you to consider [00:46:00] carefully before you implement AI in your organization, just where the guardrail should be and how you should manage and control that to avoid potential problems that you could have going forward with AI being used in an unethical manner.

So thank you very much for being part of our audience. This is CISO Tradecraft. I'm your host, G Mark Hardy. Make sure you follow us. If you're not doing so already, we'd love to go ahead and give you more information. Oh, by the way, on our LinkedIn page, we have a lot more than just podcasts. We have a steady stream of high information, low noise, great stuff.

It'll help you benefit in your cybersecurity career as well. So take care. And until next time, stay safe out there.