#223 - A CISO Primer on Agentic AI

[00:00:00] **G Mark Hardy:** Here's a quote for you. Artificial intelligence is the future for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes a leader in this sphere will become the ruler of the world. Wanna know who wrote that? Stick around. I'm gotta tell you that and a whole lot more as we do this episode on CISO Tradecraft of Agentic AI.

. Intro.

[00:00:35] **G Mark Hardy:** 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. My name is G. Mark Hardy and we're going to be talking about Agentic AI today. And that initial quote that I gave you from September of 2017 was given to a number of students.

None other than Vladimir Putin. Interesting. So as we take a look at the [00:01:00] concept of age agent ai, this might be the threshold that starts to take us into the point where someone may become the ruler of the world or may not stick around. And let's see what we're gonna learn. First of all, let's start with the definition of age agent AI and how it might be different from regular generative AI or basic ai.

And so well, who better to go to than AI to ask that question? So we check with ChatGPT. And we said, provide a concise definition of a AgentAI to a Fortune 500 CISO. Here's what we got. Agentic AI refers to artificial intelligence systems that can autonomously perceive, reason, and act toward achieving complex goals with minimum human intervention.

These systems exhibit advanced decision making, adaptability, and self directed learning, posing both opportunities and risks in cyber security, compliance, and operational resilience. guess that those last three things were thrown in there is a little plus one because we told them that it was for you.

Anyway, one of the things we're thinking [00:02:00] about is the concept of agentic AI. I talked to an expert who'd been on our show, previously, Constantinos Zagantos, and he had said, technically speaking, machines cannot be agents. Now, here's a little quote from his paper he published last year entitled Minds and Machines, Evaluating the Feasibility of Constructing an Advanced Artificial Intelligence.

One of the derivatives of natural intelligence is the phenomenon of agency, which reductively means the ability of the person to choose and act autonomously based on world modeling, planning, and incentives. Lord Anthony Giddens theorized a concept of structuration, which says that there's a constant activity dynamic interrelation among agents as the choice makers and structure in lieu of the chances that they may perform.

He famously introduced humans as knowledgeable agents with the capacity to understand the circumstances and consequences of their actions, thus they could possibly exercise choice. By other accounts, agency, at least as it's broadly understood, [00:03:00] requires that there would be some kind of animating intention behind the observed action.

For example, when I observe a painter painting, I attribute intention to the movement of his hands and the shift in his gaze. But not to his breathing. Breathing is not an action. It's rather a doing. It's something we do. A being is an agent if and only if it can perform action. Furthermore, agency can be represented as an intentional system to which one can ascribe beliefs and desires.

The system may be man, animal, but not an isolated machine because that requires a pre programmed task by a human. Now, even the most intelligent machines nowadays are not capable of agency on their own accord, in the sense that without a commandment, they do not perform actions. Humans, dogs, and even cats are all capable of performing acts because they're capable of intentional states.

Conversely, trees and rocks are not agents, since they're not capable of intentional, that is to say mental states, and thus deliberate [00:04:00] actions. Stones may change color, it might acquire moss, trees might grow leaves, but these do not happen as a result of an action on the part of a rock or tree. hey, with that as a kind of a downer on that, let's go ahead and see what we can find out to say, what are we able to do?

And, maybe Kosto's got some good insights here, but keep that in the back of your mind. And we talk about agentic AI, we talk about AI that could be autonomous. Act toward a goal, and use minimal human intervention. But it still requires the humans to kind of line it up and point in the right direction.

So one way to think about AgenticAI is to really focus on the word agent. Now, agent is a term you might already be familiar with. Like a lot of people use a fiduciary agent to help them with saving for retirement. Let's take a look at what

a fiduciary agent might do for somebody. This agent's primary duty is to make decisions that are in the best interest of the client, not for the agent.

For example, that agent should be making investments that are best for high returns for the client and not just to create high fees for an [00:05:00] agent.

Secondly, the agent is responsible for successfully completing the desired actions for the client's requirements. That means the agent must do things like paying bills, overseeing bank accounts, investing, paying taxes, ensuring that insurance is in place, and many other things. And each of these is likely a different transaction that has to be successfully completed, or it could create a liability for the client.

For example, if you don't pay the client's taxes, there's going to be a day of reckoning at some point, so it's important that the agent meets those requirements. The third thing an agent needs to do is keep good records. This is important since people generally want to see how things are being handled.

It's easy to Monday morning quarterback a football game. It's even easier to judge an agent critically, especially if we don't understand the decision making process that uses. And lastly, we need a feedback loop from the final output to identify agent shortcomings and create compensating controls to prevent bad decisions or hallucinations from being released into the real world.

So let's basically recap. Our objective is for agentic AI to make [00:06:00] decisions that align with our client's strategies and act on them. For instance, instead of going to a website to find plane flights that we do manually, we're seeing agents act in a way that this action could become Obsolete. For example, you ought to be able to go to an agentic AI and say, book me a flight from Washington DC to New York in the afternoon of Friday the 13th, and make sure I get a window seat.

The agent should go beyond the capabilities when Orbitz or Kayak or Travelocity, and then go ahead and recommend specific airlines to book the flight for you. You might even have some pre programmed information saying, hey, I love this airline, but not so much that one. This would be a huge time saver, provided But the agent doesn't produce a hallucination and maybe book you on a flight to Timbuktu.

Also, if it did something like book you on an airline that you hated or stuck you in the middle seat, there ought to be a way to see what it books so you can cancel your reservation, assuming you have time to do Now, this area of agentic

AI is really taking off. And we're, here's just a few companies that have made announcements in the last six months.

In October, [00:07:00] Amazon backed Anthropic announced its AI agents capabilities to complete complex tasks. And oh, by the way, we're going to have these links in our show notes. In December of 2024, Convergence AI created their agent Proxy to compete in the agentic AI marketplace. January 2025, OpenAI introduced Operator to complete browser based tasks for users.

Also in January, ByteDance announced UITARS, User Interface Task Automation and Response Systems, you pronounce it any way you like, that can take full control of your computer to complete tasks. That's scary. January 2025, Zapier created AI agents in February 25. Microsoft launched OmniParser to turn any LLM into a computer use agent.

A computer use agent is a new type of AI agent that can interact with the computer operating system. It can enter keystrokes, mouse clicks, and so forth. And Google is set to release Project Mariner, which will add agentic AI capabilities. The point here is that pretty much all the big players in tech have started putting a huge [00:08:00] focus on agentic AI in the last six months, which means cyber really needs to get a quick handle on how it works and how to secure it.

If we don't We probably aren't going to be able to provide much value to the developers who start playing with it or our employees who go ahead and provision themselves through shadow AI and start to invite these things into our enterprises. now that we know at a high level what agentic AI does, let's get into the nitty gritty.

There's an excellent article by Rajiv Sharma of Markovat. Which does a deep dive on agentic AI infrastructure, again, in the notes. Rajiv creates a helpful paper on agentic AI that highlights three important sub functions. Perception, Cognition, and Action that agentic AI tries to simulate. To see what that looks like in terms of agentic AI, let's walk through the example of creating our March Madness brackets and place a bet on the teams that we think, or it thinks, is going to make it to the final four.

the first sub function of Perception, It's all about taking sensory input. Now, just as humans have sensory input, like sight, [00:09:00] smell, touch, taste, hearing. Agentic machines have their own input senses. In our example, the agent is going to take in news articles and other information about the 68 teams that are going into the competition.

Once the AI collects the news articles, it needs to perform feature extraction. You can think of this like breaking a sentence down into noun, verb, and object, and parts of speech. And this allows us to store a noun that tells us who are the what, a verb that defines an action, an object or the recipient of the action, and so on.

And then we can store those inputs into discrete objects we might call object one, could be the Auburn Tigers, object two, the Duke Blue Devils, and so on, all the way up to object 68, the Florida Atlantic Owls. now that we have objects, let's proceed with performing object recognition. We can essentially tie in facts about all the teams.

This could be player stats, who's healthy on each team, analyst projections about teams, history of the coach, assistant coaches, things like that. All sorts of other important data [00:10:00] points. If we do that right, it successfully modeled the important elements of all the teams. Which is going to take us to our second sub function.

That's our cognitive module. Once we have collected all the information, cognition helps us define our goals, plans, and decision making process. In this case, we want to make a bet on which teams are going to make it to the final four. To place a bet, we need to decide on who's most likely to win four games in a row.

And a likely score of those events so that we have a confidence factor. Perhaps we look at previous games or how each team scored against a similar opponent during the current year. Ultimately, we create a knowledge base of assumptions that we use in a decision making process. For example, we can evaluate all the previous data about teams that made it to the Final Four in prior years.

And think of this as our money ball statistics that correlate the most with success. Was it the number of points scored, the number of points allowed, who followed out early, the record of the coach, something else. Since we have historical data, we could try millions of variations and measures to see [00:11:00] which one has the highest likelihood of correlating with the ultimate winner.

This is essentially a probability mass function, or PMF. The thing is That it serves a little bit better than doing a randomly guessing yourself what the outcome will be. Systemic prediction likely will be better, but it's not going to be perfectly accurate in any case. All one can expect for it to be is, good enough.

But once we find the metric with the highest accuracy rate on the historical data, we can then use generative AI's natural language processing capabilities to state our answer. And a human readable way. For example, we think that Auburn, Duke, Arizona, Michigan State will comprise the final four. Now, yes, I know the brackets aren't even published yet, and they might end up in the same row with me on this one.

Which now takes us to our third sub function, Action, or what we might consider complex automation. We want the agent to obtain our credit card information and use a service like DraftKings to bet money on the site. It queries us about our payment information, how much you want to wager. Next, we're going to create a DraftKings account.

[00:12:00] Logs in through the various API calls to place a wager with our credit card as well as provide all other account information. As a result, we can receive an email confirmation, a bet has successfully been made for a specific amount for a specific outcome. Now these last two sub functions, cognitive and action, are where we're seeing a huge improvement over generative AI.

The ability to represent goals and make decisions from the ability to automate human tasks. That's great. Because ideally, if you do it right, then you shift from requiring humans to write every little instruction with a piece of code to being able to go ahead and just describe what you want. Now, there's no guarantees here, though.

Humans have a world model. As humans, we know whether it's correct according to a real life perception and when something just can't be Machine only simulates it, and it guesses it, and that could be a recipe for disaster if we completely take the human out of the loop. So in this case, the AI agent does a lot for you, which means you start saving humans a lot of time, which could result in organizational cost savings.

Not just for the [00:13:00] betting, but think about this across multiple business functions. Also, the action part reduces the requirement for the human, in this case, to place the bet. It uses a non human identity approach to mimic a human action. And as a result, agents allow humans to automate the mundane and become process owners.

As we create more and more agents, we also need something called agent orchestration. like Kubernetes. You need something that will coordinate spinning up agents, coordinating agents, and supervising each task to

completion. Now previously, the system that could do that effectively was Multiple Neighborhood Cellular Automata, NNCAs.

And they need to be very well connected. You also want the system to be transparent. and accountable, and ideally, non hackable. Remember, if one agent crashes, you don't want the whole process to fail. You want to just restart that one agent and continue the process. You're also going to need data storage, and each agent may do temporarily cache or store their data as it moves around these various models.

And once you have agentic AI functioning at a technical level, you should [00:14:00] also expect data privacy, legal, compliance organizations to bring their own requirements. and require certain safeguards. For example, the AI agent might be prohibited from performing actions like calling 911. An AI model might be precluded from basing decisions based on protected data fields, such as age or race or gender.

And we have to make sure we restrict AI from uploading intellectual property, PII or PHI, externally, outside of our control in violation of any agreements that we have or regulations. There might even be these regulatory and compliance related requirements like documenting decisions and requests to the agent and logging actions into a SIEM.

This might be particularly helpful to see if bad actors try fuzzing the agent with malicious queries to reveal sensitive data inside the agent used to train it. There's going to be a lot of overhead here. But if you're looking for specific examples of emerging standards, documenting the various controls required to safeguard AI and agents, check out the following.

NIST.AI.600-1 [00:15:00] Artificial Intelligence Risk Management Framework Generative Artificial Intelligence Profile. MITRE ATLAS Adversarial Threat Landscape for Artificial Intelligence Systems. OWASP Top 10 for Large Language Model Applications. And ISO 42001 Information Technology Artificial Intelligence Management System.

Again, links in the show notes if you want to go ahead and dig into the pubs. First three you can access at no cost. Alas, the ISO one, I think you have to partner with 199 Swiss francs. Okay. Given all this complexity, what's the real value of creating agentic AI? Let me share with you some use cases that I think we're going to encounter in our industry.

How about autonomous vulnerability management, automated incident response, adaptive red teaming or custom awareness training, autonomous IAM, and third party risk management. Let's take a look at these. Autonomous Vulnerability Management. Imagine if we could automate vulnerability scanning, code remediation, and ticket tracking into a single [00:16:00] process.

We could really detect and fix a lot more threats. For example, your agent uses a vulnerability scanning tool like Qualys to scan your environment. Actually figure out what line of code the vulnerability might have been found in. Then query GitHub Copilot to create the code fix. Then you create a ticket to create a production change along with creating a GitHub pull request to make the change.

And if we put a human in the loop. To approve that GitHub pull request, the code can go forward, the vulnerability gets fixed, the vulnerability scan happens, if it successfully goes away, ticket gets closed out automatically. Think about how much time savings would occur by minimizing the time it takes to find vulnerabilities, write code fixes, track those vulnerability remediations, and then close out the tickets.

There's really a lot of potential here. Now let's think about automated incident response. What if you could document every action a Tier 1 incident responder took when investigating an EDR alert? You might think this is hard, but it really isn't. For example, most people do everything in a web browser today, which gives us the ability to record it.

There's a helpful google [00:17:00] Chrome extension called Scribe AI that can easily record everything you do in the web browser and create a step by step screenshot. Leveraging that type of capability, you record the most common actions that your incident responders take when they investigate malware.

Perhaps they always use the same three tools to figure out if something is a false positive. Then they'll look in various handling guides on what actions to take, such as disabling an account, make a snapshot, etc. Then doing this could reduce the incident response activities, which might take humans an hour to do, could now be completed by AI in mere seconds, allowing your team to stop potential.

Issues like ransomware, much, much faster. Our third example is custom red team exercising. What if you want to perform a red team exercise to mimic a fraud actor, but spearfish as a user? Today, organizations perform generic

phishing tests and use one template pretty much for everybody in the organization.

However, what if you could customize it for each recipient? For example, let's say you use a tool to query LinkedIn. Find everybody in your organization. Then you can take the last 10 [00:18:00] posts those individuals made as an input for a chat GPT request. The request is, create a custom phishing email that is tailored based on what the employee is commenting on.

After that, the agentic AI uses LinkedIn APIs to reach out. Could that employee with this ChatGPT message and a link to a malicious website. Now capabilities like this can drastically improve human awareness against this type of attack. are you listening 0x4? Are the other companies out there? But nope.

We've also laid out a strategy for a highly effective spear phishing campaign. Don't share this episode with the bad guys. Our fourth example is autonomous information and access management. Today, managers and companies must review all the entitlements and requirements for their employees.

However, it takes a lot of time. It's largely a waste of time because most of them are going to be approved because they've gone through some sort of process. For example, approving Salesforce access requests is probably not a good use of a CEO's time. Instead, what if we looked at the usage of every entitlement and every access to say for any system or tool not accessed within the last [00:19:00] 90 days.

Disable user access. Next we use AI to analyze what remains and see if it makes sense. Does an accountant need access to this folder which DLP detects as accounting type data? makes sense. Does the accountant still need access to a folder containing legal data which they haven't used for 8 months?

Probably not, but they still might need access to everything for the year so they can do end of year financials. So the tool is going to go through and make all these account access removal recommendations and present them in one approval request for the manager to look at, instead of 200 individual requests coming in one at a time.

And depending upon your confidence in this model, you might even grant full autonomy for this capability if you see that it's making better approval requests than most of your managements are.

Finally, let's think about using automate third party risk management. Imagine if you had the ability to reduce all the labor costs involved with reading hundreds of policies, questionnaires, pen test reports, etc. by employing a tool that could successfully [00:20:00] analyze those submissions. Create a report. Your third party analyst team would just do a simple audit to ensure the findings are accurate, not hallucinated, and as a result, your third party risk assessments could be done in minutes after submission.

This will improve your business efficiency by providing vendor approval faster. Your team is happier because their job analyzing all that paperwork has been reduced. Now, as a ciso, I gotta tell you, I want something on the other side of that's gonna prepare all that info input for me so I don't have to spend hours and hours filling out these custom forms.

But I digress. For digital tasks, agentic AI can be a game changer. It has the potential to revolutionize the many repetitive tasks in cyber. Things can become cheaper, faster, and better if we learn how to harness its potential. Although a large amount of training can increase our confidence in its ability, Where it might be lacking is the ability to interact autonomously with real life events.

think self driving cars. I'm in Phoenix this week, and I see Waymo self driving cars out here in downtown, and [00:21:00] maybe I'm a little bit old school and maybe have been doing cyber security for way too long, not quite ready to hop in the backseat of one only to see watch the locks go down. Sorry, Dave, I can't do that now.

Also consider that Ultimately, all these agents do have a human initiator who technically was the real agent. So let's take a lesson from threat modeling to understand. What could possibly go wrong? For example, automated incident response. Imagine if we use agenic AI to automate the following. EDR detects malware on endpoint.

Just disable the account till a team of incident responders can look at the system. Unfortunately, an alert just happened on the CEO's laptop, and it was a false positive. EDR saw a huge spike in files being deleted, so it's flagged the CEO's machine. However, this was by design, as the CEO is just performing a year end cleanup after hearing from the IT team to remove unused files.

Now, if CEOs would really do that, would be great. But the idea is, if this happens, do you have a quick way for someone to roll back that action when you get a nasty call from [00:22:00] somebody saying, I'm in here with a client

ready to close the deal, and my laptop's not working. Now, every technology It has the potential to have false positives, right?

But don't let that dissuade you from implementing it. Remember, don't let perfect be the enemy of good. If agenic AI has, for example, a 90 percent accuracy, but your team only performs at an 80 percent accuracy, that's a 10 percent improvement. Now, don't expect the agents to be perfect. You wouldn't complain to get 10 percent more headcount to perform your mission, would you?

What you really need to look at is how can we begin to trust these agents, or more specifically, trust the prompt engineers that are initiating the AI tasks. Let's make a comparison. How do you trust a new intern? you give them a project and watch the results. Essentially, how well they complete the task.

If they did well, give them a bigger task and see how that plays out. And if not, then you go back and do some retraining and some feedback. The same can be done with prompts to AI agents. And to a certain extent, if it's configured incorrectly, the AI [00:23:00] agents themselves. If you take this measured approach to building trust in agentic AI agents, I think you'll see them consistently making the organization run a little bit better, a little bit more efficiently.

And people will be happy about the reduction in the amount of boring, repetitive tasks that could be handled quite adequately by this system. Now don't forget about government, governance, transparency, explainability, accountability, because it brings up some interesting questions. If you connect two IT systems together today, Does that require an access approval and firewall approval and everything else?

Or is that considered to be a brand new application all by itself? And if someone creates a new Power BI visualization based on a Google form, is that a new application that must be formally registered in a configuration management database? Understand where to draw the line, especially as organizations create thousands of these agents that will test your ability to decide what needs to be routinely governed.

Get her wrong! You'll have agents that could upload sensitive data [00:24:00] to places that you can't protect. let's not do that. So I hope this gets you thinking about what decisions you're going to face soon with agentic AI. If you think of scenarios where it's likely to go bad, including those in your incident response plans, maybe have an alternative.

Having an effective plan and rehearsing it in advance will ultimately help you respond faster and better to potential incidents. thanks for listening to today's show. And just a quick recap. We talked about agents, how to think about them like fiduciary agents acting in the best interest of their clients.

We talked about the rapid pace of innovation, which is why we need to get an understanding of them, went over three major sub functions, perception, cognition, and action that will enable agentic AI, and we even gave five examples of how they could transform cyber and things such as that. So think about that.

And by the way, I just found something that came out on the last couple of days called Manus, M A N U S dot I M. It's a new agentic AI tool. It's still in beta. If you go take a look at, I'm not necessarily advocating for it. It is out of the people's Republic of China. You [00:25:00] decide if you want to play with it, but it looked extraordinarily promising and it looks like a big potential disruptor.

So keep your eye on that site and go take a look. And who knows, if you're lucky, you might be able to go ahead and get an early access code. And if it works out for you, send me a message. Let me know on LinkedIn, how this is going. So if you've enjoyed this episode, pay it forward. Share with your colleagues at work.

Post a link to our show on LinkedIn and repost it so people can find it. The more people that apply good security in their enterprise, better for the rest of the world, because we're all going to help keep our information and our business processes safe. So thank you again for listening. Wish you only the best as you master your CISO Tradecraft.

This is your host, G. Mark Hardy. Again, till next time, stay safe out there.