SDS PODCAST EPISODE 989: AI INFRASTRUCTURE, RAY, AND WHY NONLINEAR CAREERS WIN, WITH LINDA HAVIV

Jon Krohn:	00:00:01	Anthropic claim Mythos is so capable at finding and exploiting cyber vulnerabilities that they can't release it to the public. Regardless, the clock is ticking. Within six to 18 months, open weight models will have the same security exploiting capabilities. So what are we going to do? Welcome to another episode of the SuperDataScience Podcast. I'm your host, Jon Krohn, my sensational guest today, are Anneka Gupta and Cal Al-Dhubaib, who serve as chief product officer and principal technologist respectively at Rubrik, a massive Bay Area security and AI company that is listed on the New York Stock Exchange. In the Claude Code and Mythos era, we have wildly powerful code-generation tools at our disposal, but this is rapidly accelerating security risks, both through inadvertent oversights and purposeful misuse. Luckily, Rubrik have equally powerful solutions to meet the challenge. In today's episode, Anneka and Cal will tell you all about it. Enjoy.
	00:01:00	This episode of SuperDataScience is made possible by Anthropic, Acceldata, and Cisco.
	00:01:06	Wow. I have two guests on the show today. Cal, Anneka, how are you doing? Maybe let's start with Anneka since it's your first time on the show.
Anneka Gupta:	00:01:16	Well, I'm very excited to be here. I'm in sunny Florida right now, so what a place to be recording and spending time with you today.
Jon Krohn:	00:01:24	Nice enjoying it. And Cal, welcome back to the podcast. You were such a treat last time.
Cal Al-Dhubaib:	00:01:28	It's great to be back and I'm so excited to be dialing in from ODSC, one of our favorite conference menus.
Jon Krohn:	00:01:34	The Open Data Science Conference. We've run lots of sponsor messages for that show on this podcast, but we genuinely love it. I've seen you there, Cal, so many times

and so many guests that we've had on the show. It's really the Open Data Science Conference. If listeners haven't had a chance to go, I realize that they do sponsor this show, so this sounds like it might be biased, but there's no conference that's better for a listener of this podcast. And there's no place that you can go and meet more SuperDataScience podcast listeners. Truth. So great to have you both here calling in despite your travels. And we had to make this episode happen because there are very timely things happening in AI that couldn't go another week without us covering. But before we get into those very timely matters, I want our audience to have a little bit of background on both of you.

00:02:24 So let's start with you, Anneka. You did a Stanford bachelor's degree in math and computational sciences, and then you did 11 years at LiveRamp in the Bay Area, I believe that whole time. And you held something like 17 different roles before becoming the chief product officer where you are now at Rubrik, publicly listed company. But I'm no expert at Rubrik. You are, you seem to be loving it there. So tell us about the platform and let us know. Are you going to break your 11 year record that you had at LiveRamp now at Rubrik?

Anneka Gupta: 00:02:55 Well, I'm five years in, so we'll see. And it still feels like it's day one. Super excited to be here. Rubrik is a cyber resilience platform, and we've been around for about 12 years now. We have been innovating and evolving throughout those 12 years at the same speed that we were, I think, on day one of the company. We started by recognizing that there was a huge challenge around resilience of data and making sure that your data could always be available to you regardless of where that data lives, which it's now living in many more places across cloud, across on- prem, across SaaS applications. And that was the beginning of our journey. And what we found is over the first few years of the company that we got pulled into cyber, because really what was happening

was that people weren't losing data anymore because of accidental deletions or because of natural disasters.

00:03:50 They were still losing data for that, but the more frequent reason why they needed to bring back their data was because of cyber attacks. And now, as we think about the future and the past few years and going forward, we're now really focused around AI and AI agents and recognizing that the next vector is not just increase of cyber attacks, but also looking at AI and the way that AI is being used, inevitably AI is going to make mistakes. Those mistakes might be benign mistakes or they might be malicious mistakes, but it's going to make mistakes and you need to build resilience for your organization in order to make sure that no matter what happens, whether it's a natural disaster, cyber attack, AI agent making a mistake, your business is never coming down and that your customers are not materially impacted.

Jon Krohn: 00:04:36 Right. So it sounds like the starting point for Rubrik as a business before it grew into this big publicly traded, niecey listed business is that ... Oh, and by the way, I also recently noticed that your share price has been doing very well since listening, so that's a nice sign to have. And it sounds like you got started as a cybersecurity business, but then with all of these fast shifts in AI and all of the vectors for security issues that have emerged, especially now in the Agentica AI era, Rubrik has had to come up with solutions for all these issues.

Anneka Gupta: 00:05:12 Absolutely. And that's been a big focus for us.

Jon Krohn: 00:05:15 All right. Now, Cal, last March, you were on this podcast in episode number 865, so people can listen to get a whole bunch more context on you if they want to. At that time, you were running AI at Further, which is an AI consulting firm. You were 80 plus AI projects deep. And at that time, in that episode, you were advocating for making AI boring as the path to success, especially with

production AI. But now, Cal, a year later, you've left that role at Further, and you've joined Rubrik as principal technologist, and now you're publicly arguing the opposite of making AI boring. You're saying that we're in a watershed cyber AI moment that's anything but what changed your read of the landscape between last year's episode and this one now?

Cal Al-Dhubaib: 00:06:03 I think the step function and the blast radius of autonomous systems that can iterate over multiple steps and chain together a series of tools that are now increasingly resulting in material impacts to business. I mean, the news cycle, even in the last two months, has been more intense than any other period. I've been tracking AI incidents. I'm a big fan of the AI incident database project, and that's been on a hockey stick. But within the last couple of months, Amazon, for example, lost 99% of their orders on a single day as a result of code that was generated by an agentic tool. There is this Reddit thread that blew up this past weekend, and it was about a small business that had a bookings and reservation system, and it completely wiped every instance in their database. And so their customers were trying to retrace through restreets and stripes.

00:07:05 And so the material impact is here and the blast radius is growing. I still maintain that we need AI to become boring. I think it's gotten a little too exciting and we need to dial the heat down. And I know we're going to talk a little bit more about what that means with trust infrastructure and trust engineering, but when this opportunity came up with Rubrik, I was so excited. I was already seeing the trend unfolding that the only way to make AI work in the enterprise is if we could, one, secure the data and information assets that it's connecting to and two, safeguard AI from malicious attackers that are now increasingly very productive with their own use of AI. And oh, by the way, they don't have AI usage policies. So I'm really excited to be a part of the story here. And I can

say three weeks in, even though Rubrik has gone public and they're a massive company with thousands of employees, they still very much operate like a startup.

00:08:09 And it's been so fun getting to immerse in this culture.

Jon Krohn: 00:08:13 Really cool. Well, congrats on the move. And it makes so much sense to me in addition to, or perhaps correlating with all of these cybersecurity incidents that you've been describing is, you talked about it being this big influx in the past two months. And it's interesting that that coincides with, it's been two months at the time of recording since the release of Claude Opus 4.6 and ClaudeCode really taking off as something that people can be leveraging as a serious power tool for doing development work and probably identifying vulnerabilities.

Cal Al-Dhubaib: 00:08:49 Absolutely.

Jon Krohn: 00:08:50 And then now more recently, just a couple of weeks ago at the time of recording, on April 7th, Anthropic announced Claude Mythos preview. And so this is a model that famously they haven't released to the public. There's only people testing it privately. Anthropic's own estimate, however, is that mythos class capabilities will proliferate to other labs within six to 18 months with open weight versions to follow. So yeah, the reason supposedly why they didn't release it to the public is because of its prolific ability to identify cybersecurity issues and exploit them if it's not being used by a friendly person. Now, I also do think it's brilliant marketing to say that we carry it to the public. But yeah, from inside Rubrik, is Mythos the inflection or just the most visible point on a curve you'd already been planning against?

Anneka Gupta: 00:09:48 I think when you look at Mythos, it's definitely the most visible point on the curve that we've already been planning for because even if you look at models like

Opus, for instance, they also are able to find vulnerabilities. They just require a little bit more handholding to get there. Whereas Mythos, you can point it at an open source repository and tell it to find me all the vulnerabilities and it will find you the vulnerabilities. And this is just the nature of where AI is going. The challenge is that the cybersecurity industry as a whole has been super focused on attack prevention and attack detection because it used to be the case that attackers would enter into your system, they would sit there for weeks, if not months, sifting through your data, finding the opportune time to actually exploit this vulnerability and really perpetrate the attack and then actually do that in a visible way such that the company then has to react to it.

00:10:50 But with the reality that now with AI agents, you can find and exploit these vulnerabilities much, much faster and at machine speed that requires machine speed response. And no longer can you hope that if you detect an attacker, detect a breach, okay, maybe you can cut that off before the attacker has done damage. Instead, you have to assume that you've already been breached and you need a plan for how are you actually going to recover and ensure that in that recovery, you're minimizing any impact to your overall business. And that's the business that Rubrik has been in. So I feel like there has never been a more important time for what we've been doing because what we're ensuring is that you can actually restore your applications, your data, your identity, your infrastructure back up and running extremely quickly. And in the case that an attacker gets into your system and destroys your data, your identities or your infrastructure.

Cal Al-Dhubaib: 00:11:55 What I find really interesting is it's like this twofold problem and we're kind of feeding into it from the enterprise perspective. There's still massive FOMO. Everyone has to use AI. Everyone has to be more

productive and it's coming from leadership down. And so we're very quickly opening up actually the surface area that models that are like Mythos can actually chain together vulnerability. So it's one, you've got this problem of adoption as increasing surface area. And then two, these models themselves, and I know we're going to nerd out a little bit more about it, but I get so passionate about this. We don't yet have the right trust infrastructure in place in most cases to prevent models from causing harm on their own. So you don't even need the bad guys.

Jon Krohn: 00:12:37 Right. Yeah. It's pretty wild. I mean, it was Anthropic themselves that actually had, there was some flag, like a Boolean flag that they got wrong in some … Somehow it was possible for them. You guys might be able to explain this better than me because you actually are cybersecurity experts. Cal's been one for three weeks, but there was like a flag that was set the wrong way in some code that was pushed to GitHub. And so that allowed people to see the entirety of the Claude code base. How does something like that happen?

Anneka Gupta: 00:13:10 Well, I think that's an interesting case. There's a lot of different scenarios where things have been exposed publicly that wasn't supposed to be. There are cases where code has been exposed externally because the AI agents posted it into the wrong repository. That's not what happened in the Anthropic case. In anthropic case, there was, again, a bullying flag that was misconfigured and sent out. But the reality is, is that the challenge that we have today is that AI agents, they're very outcome focused. They're going and trying to say, how do I get this job done in the least steps possible, the fastest way possible? And they're not taking into account the same rules that us as humans, as employees working in an organization know to follow. They haven't gone through all the security training that we've gone through. They haven't gone through the developer best practices training.

00:14:00 Now, obviously they have that information in their models broadly, but they're not optimizing for the same thing that we're optimizing for. And what that means is that inevitably AI agents are going to make mistakes. Inevitably, they're going to do things that you didn't want them to do in their pursuit of this outcome of a task that you've given them.

Cal Al-Dhubaib: 00:14:21 I find it really interesting that these models, because we've been investing in the capabilities for them to iterate autonomously, think of it as an exhaustive search that's now happening across all of the data assets and tools and configurations of that tool. Humans in the past, we actually had a lot of security through obscurity. We might have had permissions. And I've been studying my cyberstuff. You have. We had security through obscurity and we had access to files and permissions, but we didn't broadly use these permissions to take large scale actions. And now we have these very hyperproductive agentic systems that can do an exhaustive search of everything that's available to it to accomplish the task. And so a lot of cyber controls really were never designed with this reality in mind.

Anneka Gupta: 00:15:13 And it's not just about the mistakes that AI can make or what it can publicly expose. There are other kinds of data exposure risks that you have even internally, right? You could say, "Hey, salaries of an employee could be accidentally shared with another employee because you've hooked up your AI agents to your employee and compensation systems." And that is really scary too. So there's a lot of implications both internally and externally to exactly what Cal is talking about. And as the models themselves are able to take just more and more and more and more steps independently, it's really hard to diagnose where in that 20 step process did this happen and what caused it. And you're certainly not going to be able to fix that after the fact. You have to figure out how you're

actually going to fix this while this agent is actually running.

Jon Krohn: 00:16:10 Right. Yeah. So on the one hand, having these kinds of tools like ClaudeCode to be able to generate so much code so quickly that creates opportunities for mistakes inadvertently entered because like both of you have said, the agent is trying to get to the outcome as quickly as possible and therefore some security things can get through. And then there's so much code that it could be really hard for humans, impossible for humans to have oversight over all that code. I've been reading that in a lot of organizations now, the engineering team is struggling to know what to do with this vast amount of code that's being generated. There's so much potential, but also so much risk. Now on the other side of things, something that we haven't talked about very much, which was actually kind of my starting point with this question. I was thinking, oh, the release of Cloud Opus 4.6 in February, this correlates with this boom in cybersecurity incidents.

00:17:05 In my mind, I was thinking this is because that actors are using it to find and exploit vulnerabilities. And so far we've mostly been talking about kind of people inadvertently creating, exposing vulnerabilities through using CodeGen tools, but the Codegen tools can also be used maliciously, right?

Anneka Gupta: 00:17:24 Yeah, absolutely. And it's already starting to happen. There are cases where this is publicly already happened. I think what we see though is that it's more about what people are planning for is really about the future and where these tools are starting to be used at scale. So like for instance, over 87% of executives at companies are now worried about AI vulnerabilities and that being their fastest growing vector of cyber risk. So that was from the World Economic Forum published that. Most organizations, again, over 80% of organizations don't feel

like they have the kind of visibility that they need into AI agents. And so when you start saying like, "Hey, attackers are going to be exploding traditional vulnerabilities as well as AI vulnerabilities," it's just starting now and there's already examples now that people are talking about, but fast forward six, 12 months, and this problem is going to be untenable.

00:18:28 And that's what is getting everyone right now in our space really worked up about, okay, how do we come together as a community? How do we help solve this? How do we help our businesses stay resilient in the face of a risk that is growing exponentially day by day?

Cal Al-Dhubaib: 00:18:46 Well, I can't stress this enough. The phrase resilience resonates with me. And part of why I was so enticed to join Rubrik is it's not if, it's when, and then what do you do next? And when I've consulted clients in the past on their AI governance strategy and their approach, the most often missing element. It's like, okay, we can do evals, we can do testing, we can get some observability in place and that's kind of mid, but the thing that rarely was ever done is, all right, when it fails, whose job is it to recover and what does that plan look like? And oh, by the way, have you tested that?

Jon Krohn: 00:19:24 All right. So this sounds like a perfect time to ask. You're kind of alluding to it now, but let's get into some brass tacks around the Rubrik platform itself. It's been around for a long time. It's a publicly listed company, so there's probably a lot of facets, but is there some kind of like exemplary or maybe a few exemplary user stories that you can walk us through so that we can visualize what it's like to use the Rubrik platform to have better cybersecurity in this Agentic AI era?

Anneka Gupta: 00:19:55 Yeah, absolutely. So what I'll do is I will take you through an example of what does a cyber attack really look like, and then where does Rubrik's platform fit into that? So

typically what happens is that a cyber attacker will go in. The most common vector now is that they'll go in and compromise credentials of some user within your organization. That's why if you're working in a company, you have all this like anti-phishing training and all these videos and tutorials you need to go through because this is very, very common. This is the most common way. It's very like not sophisticated in some senses. They're sending you emails with links to click on. They're calling into your support. This is a very common one that, and I think the MGM- Call into support. They're called into support and they basically pretended to be the employee and they reset their password as well as their multifactor authentication and therefore they were able to compromise the credentials.

00:20:55 So that is a very common attack vector as well. So again, not super sophisticated, but that's how they get in. And then what they do is they look at what system access does that user have? They look for ways to give themselves more access, compromise other identities once they've compromised your identity and get closer and closer to the crown jewels, to your production systems, to your customer data before taking action. And that action might be, hey, they're going to exfiltrate that data. So export that data out, threaten to put it on the dark web, encrypt or delete a bunch of data and hold that for ransom and say they're not going to give you the decryption keys or keys or give you all that data back if you don't pay them. And that's typically what happens during a cyber attack. Now, the challenge is that within this scenario,

00:21:50 Before you actually go and say, "I need to recover. I've been attacked. Okay, all these files were deleted. I need to recover." The first question that most organizations are going to ask or like leaders are going to ask their teams in this scenario is, "What was the impact? What actually happened?" And actually answering that question is

really challenging. Understanding what was the blast radius of the attack, was any sensitive data impacted? When did this attack first start? Where was the first point of infection? And before I recover, how do I make sure I don't recover back all these vulnerabilities again and just open myself up again for the attack? So all of these questions you have to answer. And what Rubrik really focused on is we built our software in such a way that we have built in something called the preemptive recovery engine, which actually goes and scans all of this data, scans all the data, understands how that data is changing over time to basically help you answer these questions and answer them extremely quickly so that you can also quickly hit the recover button and bring back your systems up and running as fast as possible.

00:23:08 And that's kind of the secret sauce of what Rubrik has built. And in doing so, we've built a deep understanding of data and of identity. And these are the areas that you need to have a deep understanding of in order to build resilience, security, and AI operations for AI agents. And so that's really what has like gave us this opening into this new world of AI is just having this deep, deep understanding of data and identity that practically no other company on the planet has.

Jon Krohn: 00:23:39 You're so good at this. That was such a good explanation. My understanding of cybersecurity attacks and how a solution like Rubrik helps is ... Yeah, I feel like I went from zero to 100. That was a

Cal Al-Dhubaib: 00:23:52 Masterclass.

Jon Krohn: 00:23:54 Yeah. Thank you. Cal, anything to add?

Cal Al-Dhubaib: 00:23:58 I mean, and I know we're going to talk a little bit more about this from the AI side of things, but I can't stress enough that we have an identity crisis in the AI world. And so increasingly we have what we call non-human

identities. So this is increasing the number of surface or port points that then malicious actors can exploit. And it's not just now your user, but it's your user and every agent that's acting on their behalf that now has access to keys and tokens and APIs and all the systems that they can then cascade to X, the actions and permissions that they can take in those systems. And so this non-human identity explosion in the enterprise is the problem to solve.

Jon Krohn: 00:24:42 Yeah. And those agents are famously friendly and helpful as much as they can be. Yes, let me reset that 2FA for you.

Anneka Gupta: 00:24:52 Well, it's also interesting because if you think about, I'm sure a lot of listeners are CloudCode users. If you think about using CloudCode today, especially in enterprise context, I have it set up on my laptop. I go and fetch all of my API keys from all of the SaaS apps that I have access to. I try to give it the longest time period before they're going to revoke and circle my tokens because I don't want to have to go set them up again. I'm giving them the exact same permissions that I as a person have and setting up my tokens that way. And then I'm sticking those tokens on my laptop and they're sitting in a file on my laptop. So when you look at this whole system and how identity is architected in the agentic world, it's obviously still nascent because all of the tool sets are changing quite a bit, but the security protocols and the way that you should actually be architecting identity authorization and all of the other pieces are very, very rudimentary.

00:25:57 And a lot of the things that we're all doing today are not things that you would allow when you're building enterprise applications, but there's this trade off between like, are you going to cut off people's productivity by saying, "Hey, they can't do these things or are you going to allow it and have the right monitoring and visibility in place in order to manage it? " And of course,

organizations want to do the latter, but the tooling hasn't been there in order for them to be able to do that.

Jon Krohn: 00:26:25 Really nicely said. There's a term that we haven't said yet that I hear a lot in the cybersecurity space, which is a zero trust world. What does that mean to be in a zero trust world and what does operating in that zero trust world look like?

Anneka Gupta: 00:26:41 So zero trust, I would say is just a fancy word to say assume breach. Assume that every device, every identity, every account has already been breached. What then? How do you architect a system such that basically for your crown jewels of data, for your crown jewels of infrastructure, the very least number of people have the ability to access and change that. And in an ideal world, that access is somewhat ephemeral so that there's no permanent access to things that are incredibly important. And that's like what zero trust means. It's been a big term used in cybersecurity more broadly as we thought about cyber attacks. And again, I walked through that anatomy of a cyber attack. It's like, if someone compromises my credentials, I should have access to almost nothing. Now the reality of an enterprise is that you can't have employees that have access to nothing.

00:27:38 So you're doing this balanced trade off of saying, "Okay, I know I have to give them access, but in that world, how am I going to respond assuming that that credential or that user has already been breached?"

Jon Krohn: 00:27:49 Nice. Crystal clear as with all of your explanations today, Anneka, thank you. Cal, I've got a question for you that is specific to some talks you've been giving a LinkedIn learning course that you have coming out soon. So you've been developing something called the Trust Engineering Framework You keynoted on it at something in March that sounds to me like I grew up in Canada and somehow the name of this conference, I'd never heard of it before. It

sounds like the name of the indigenous people that live in the north. It sounds like a town that they would have ERMUC.

00:28:22 So you're speaking at IRMAC in March and you've got a LinkedIn learning course coming up on this framework. You're keynoting at Data Summit 2026 in Boston next week after. So the same week that this episode is coming out right after ODSE East in Boston, you're still there. And all of this is centered around this trust engineering framework you've been developing. So help us understand what trust engineering is all about. Yeah. So we have terms like AI governance, responsible AI, AI safety, ML ops. How does trust engineering relate to those? And is there a role called trust engineer?

Cal Al-Dhubaib: 00:28:57 I think that there should be a role called trust engineer, but I'll take a step back and I'll say trust engineering is something that I've been playing around with over the last decade of my career. And for listeners who didn't catch the last episode, I spent 10 years designing machine learning and AI systems and leading teams that do that in heavily regulated environments, healthcare, financial services, education, energy, places where there's a high cost of making mistakes and you're already building these models on top of sensitive data assets and there's a whole set of nuances that you have to deal with. And the big aha for me really isn't that surprising, but it's consistent with the zero trust mindset. It's assume your models are going to mess up. You cannot control them to a hundred percent. And so how then do you operate in the enterprise adopting and designing these AI tools and workflows in a way where you can actually manage the risk?

00:29:56 And so trust engineering is a combination of two Two different fields. It is the combination of understanding human-centered design as it relates to AI systems. How can you manage user expectations and balance workflows between humans and machines according to level of risk

or cost of making mistakes? And then you're pairing that human-centered design approach with the ability to configure the appropriate trust infrastructure and trust assurance. And think of that as your governance and human-led architecture of the trust infrastructure you have in place. And trust engineering is a framework of how do you unify all of this? Because you can't just bolt on AI risk management after the fact. It has to start at the front. And I'll give you a very simple example. If anyone's been to San Francisco recently, you've probably come across a Waymo and it can feel like a game of chicken when you're trying to go across the crosswalk.

00:30:54 Does it see me? Is it going to mow me down? Is it safe? And I'm a runner. Every time I'm in San Francisco, I take a run along the Embarcadero and there's Waymo's galore. They have this new feature that I actually really love as you're crossing the street. Little Crossman appears on that little circular cone on top of the Waymo. And it's a clear signal saying, "Hi human, I see you. It's now safe to walk." And I love that because that's such a great example of human-centered design and expectation management around an AI system that could otherwise cause meaningful harm. So trust engineering is unifying that with the trust infrastructure. Happy to nerd out a little bit more about that, but having the building blocks in place that allow you to determine what potential risks exist upfront, to monitor the system and detect when it's misbehaving or behaving not according to your expectations.

00:31:45 And then three, when a mistake does happen, what's your fail back? What's your fall safe? And so it really is trust engineering is building off of what Anneka so eloquently described as the zero trust mindset, but applied to AI systems.

Anneka Gupta: 00:31:59 What I love about the trust engineering framework is that it's really bringing together people, process and

technology, which is the framework that we always think about as business leaders as we're driving change throughout the organization. And it's only in thinking about all of these three holistically together that we can actually truly solve a problem. Because if you only solve for one piece, you're only solving for one leg of the stool. And you're not going to be able to actually build these products and processes in a way that are truly achieving the end means that you're trying to achieve.

Jon Krohn: 00:32:34 I mean, I've complimented you a bunch of times, Anneka, on your speaking ability, but I think Cal did a pretty good job there too explaining. I guess, and I think I've been too hard on him by joking that he's only been cybersecurity because I forget that the whole time running Pandata and then at further he's doing AI consulting in really in industries like healthcare, like finance where security is paramount. And so sorry for teasing.

Anneka Gupta: 00:32:56 That why we brought him onto the team.

Cal Al-Dhubaib: 00:33:00 Well, I mean, we were nerding a little bit about this before the call, but there's almost like this crossover moment happening where AI risk management is kind of reaching over into cybersecurity and I'm seeing the birth of a new discipline. And so I'm not a cybersecurity expert yet. I'm working on that, but I can already see where you can't really have one of these now without the other. And so we're seeing the birth of a new breed here.

Anneka Gupta: 00:33:25 It's so true. Even when we're going and having conversations around AI and security and operations of AI with large enterprises, it's really interesting to see that the people we're talking to, sometimes it's the CIO and their organization. Sometimes it's the CTO, sometimes it's the CISO, sometimes it's all three of them, sometimes it's some chief data officer or chief AI officer. And the reality is, is that the lines are blurring because the problem space crosses all of these different disciplines. And it's

not like you can just solve again for one discipline or the other. You have to think about it holistically because any sort of risk management you put in place could also potentially come at the expense of productivity. So how do you manage both of those together in tandem?

Jon Krohn: 00:34:13 Wow. Yeah. This does have me thinking I should be more concerned about security than I have been before this episode. There's probably a lot of listeners out there thinking about ... Because I've gotten this far without any major issues personally or professionally, and so you kind of think, "Ah, it's going to be fine." But then when I'm using CloudCode and I don't even know, I'm not even really a software developer and I'm using CloudCode to create production code, I should definitely be really on the ball with these cybersecurity concerns. I

Cal Al-Dhubaib: 00:34:45 Mean, that's exactly why we're doing this. It's kind of like we're not trying to be like the naysayers, but this is like ringing the alarm bells. We really need to start thinking about AI adoption differently and resilience has to be a part of the conversation.

Anneka Gupta: 00:35:00 And it has to be a part of the conversation upfront. If you look at like another revolution that happened in technology is the cloud revolution, public cloud, people moving to the public cloud. And a lot of ... When we've talked to organizations about cyber resilience, they've thought about bolting on cyber resilience after they've already made the move instead of upfront. And that means that that level of effort to do so is so much higher. With AI, we can't afford to bolt it on after the fact. We have to do it upfront. Otherwise, things are going to happen that are very bad and that's super important for people to recognize.

Jon Krohn: 00:35:34 Right. Yeah. You've been speaking a lot in the past few minutes, both of you, about this intersection between AI and cybersecurity. I also, in my last question to Cal, I

asked him, I built on some things that he'd been doing at conferences. And so you, Anneka, in the same way that Cal has come from AI risk into cybersecurity, you come from this long cybersecurity background increasingly into AI to wit at RSAC, which is the biggest security conference. It was held in March, I think in the Bay Area. And at that conference, at RSAC, you announced something called Sage, the Symantec AI Governance Engine, which built on a custom small language model or SLM for shart for short.

00:36:20 We don't need any charting on this podcast. And I don't think that's a word we even have to bleep out. It's just a fun one. But yes, a small language model, SLM called say just semantic AI governance engine. And so you published a head to head between GPT 5.2 where your SLM process messages five times faster, which shouldn't be surprising given that it's such a small model relative to GPT 5.2, but it was also higher accuracy on policy violation detection. So tell us more about this. Why did you go with a custom trained SLM rather than maybe fine tuning or distilling a frontier model?

Anneka Gupta: 00:37:03 Yeah. So let's take a step back and talk about how we're approaching AI governance. When we think about the challenges that organizations are having with, as they deploy AI and as they try to deploy AI safely, they're really three pillars of what we're trying to solve for. The first one is lack of visibility. I shared this stat before. Over 80% of organizations feel like they lack visibility into what agents are running in their environment, what are those agents doing, what actions are they taking, et cetera. So that's the first pillars. You have to have that visibility. The second pillar is governance and control. So agents are going to be taking actions all the time. You need to be able to block the actions that you don't want it to take. You need to be able to monitor for high risk actions. Now, the challenge with agents is different than other

traditional enterprise systems is that agents and models are inherently non-deterministic.

00:38:03 You don't know what they're going to do. And therefore, you can't use a simple rule engine to actually monitor these agents. You have to use AI agents to govern the agents. And so that's where Sage comes in. And what we did is we fine-tuned an SLM because when we're talking about this policy enforcement layer, looking at all of the activity and all of the calls that are happening to the model and the responses back and being able to block it, there are three factors that we have to solve for. The first one is performance. Obviously, accuracy is super, super important. If we're not able to do this accurately, we can't build that trust with our customers that we're actually solving their problems for them. The second is cost. When we're talking about running this, there is a real cost associated with it. By running a fine-tuned SLM, we can manage the cost.

00:39:00 And the third is latency, and that the response time has to be super fast. If we're going to block an action, we can't wait five seconds in order to decide whether to do that, especially when an agent may be chaining together so many multiple actions, and that would increase the response time, which would be untenable for the user. So when we take cost performance and latency together, we recognize the best way to solve this problem was by fine-tuning a small language model. And about a year ago, we acquired a company called Pretabase that specialized in this area. At the time, we actually didn't know that it was going to evolve into this area or into what it is today as part of Rubrik Agent Cloud and Sage, but we acquired them knowing that, hey, we have a lot of context into data and into identity, but we don't understand the models well enough.

00:39:53 We don't understand fine-tuning. And we recognize that bringing all these three together could be a magic

combination. So now that's providing the foundation for Sage, which is our governance pillar. And then the last piece is remediation, which is what happens when your policy doesn't catch everything? What happens when an agent inevitably makes a mistake? How do you actually rewind those actions? And that's agent rewind where we're able to actually undo the malicious agent actions and restore your systems back to normal.

Jon Krohn: 00:40:24 Wow. There was a term in there that you kind of glossed over that sounded really exciting to me and I need to hear more about, which is, I think you described it as the Rubrik Agent Cloud. What is that?

Anneka Gupta: 00:40:36 So Rubrik Agent Cloud is the platform that we are selling for AI operations and security. So it has these three pillars of being able to give you complete visibility into your agents, whether those agents are Microsoft Copilot agents, AWS bedrock agents, CloudCode agents, whatever agents you have, it's giving you that holistic visibility. It's giving you the governance and control with Sage where you can define a policy in natural language saying like, "Hey, I don't want my agents giving financial advice or I don't want them sending emails out to my customers." However, you want to express those policies and then Sage will help do the enforcement of those policies at runtime. And then there's the agent rewind piece, which allows you to rewind any agent actions that were unintentional in your system to get your system back up and running. So that whole platform is what we call Rubrik Agent Cloud.

Cal Al-Dhubaib: 00:41:31 It's been really fun learning about this and just to kind of like why I'm so excited about it. And this is like new halo effect, but it's like the e-discovery of like information assets but apply to AI agents. And so it's not just like, "Hey, manually add everything in here," but it's like, "Hey, discover every single agent that your users and human identities have spun up acting on their behalf." And then

across all of them, it proactively checks permissions and policy violations that could exist. And so you kind of get a posture upfront and nothing pops eyeballs more than showing a dashboard of like, "Hey, did you actually know that you have X hundred number of agents? And oh, by the way, these 50 of them have read, right access to these four databases." Are you cool with that? Do you want that to happen?

Jon Krohn: 00:42:19 Yeah. Another great example and another great reason why I can see why you're so excited to be at Rubrik, Cal. There was rewinding a moment back to what Anneka was talking about. I'm going to get into something called Agent Rewind, but before we get into Agent Rewind, there's something that I want to highlight really quickly, which is that Anneka, when you were reeling off kind of mainstream agents that people might be using, I think it's so interesting in telling that you listed a Microsoft one, you listed AWS, Anthropic, OpenAI didn't even come up, which is just two years ago, who would have thought? And obviously they are still big players in this space, but you can reel off three of the top agentic platforms or tools that people might be using and not mention them without somebody batting an eyelid. Yeah. Interesting.

Anneka Gupta: 00:43:10 I think it just goes to show how fast the landscape is changing and a year from now, who knows who we'll be talking about, right? It's always easy to say, extrapolate, oh, the people that are winning today are the ones that are going to win a year from now. I think with AI, that's less certain than ever before in the history of technology, which is both exciting and incredibly scary at the same time.

Cal Al-Dhubaib: 00:43:35 Well, it's been interesting to watch and see who is winning the race on enterprise adoption. And

Anneka Gupta: 00:43:40 So

Cal Al-Dhubaib: 00:43:40 These companies are also clearly aligning themselves around whether this is for workplace tools or it's more consumer focused. And so I think in the next couple of years to come, we'll start to see even more of that differentiation and specialization. And I'd be willing to bet just based off of some of the product news these companies are making, we'll see ones that start to align around very specific industries.

Jon Krohn: 00:44:03 It makes so much sense. Yeah. Right now, the big players that we think of are diversified across all kinds of enterprises, but yeah, it's such a fast growing space that it can't stay like that forever. So yeah, now rewinding to Agent Rewind, much like the Rubrik Agent Cloud that really caught my attention, this agent rewind functionality sounds really interesting to me. And it's one of the more conceptually interesting things I think you're shipping, of all the interesting things you're doing, that idea of being able to undo what AI agents have done, it sounds almost counterintuitive that this is possible because agents do things that touch the world. They send emails, execute API calls against third party software platforms. They write to databases that you don't own. They could even place trades in some cases. So what kinds of things can ... Are there things that can't be rewound, rewinded?

00:45:01 I'm not exactly sure what the password is.

Anneka Gupta: 00:45:02 Yes. It's a great question. So when we think about agent rewind, it really hearkens back to our legacy and our strength in cyber recovery and being able to recover different kinds of systems. So where does agent rewind work great and where can we rewind actions? Okay, you dropped a production database table, we can bring back that production database table. You changed something in Salesforce that broke the relationship between your opportunities and accounts. In Salesforce, okay, we can fix that for you. You made some change to your identity

systems that gave people a bunch of access that they shouldn't have. We can fix that. So those are the kinds of actions that we're able to rewind. Now, if you've sent an email to a customer and it's already arrived at the customer, there's not really too much that we can do about those kinds of situations and those will require ... And that's where we think the layer of defense with Sage is really the right approach of saying, "Hey, define upfront some of those high risk activities." I don't want agents to be making trades for me.

00:46:11 I don't want it to be sending emails to my customers. Define those things, those kinds of behaviors upfront because the best way to prevent those is to actually be at runtime looking at what the agent's actually doing and stopping there. But now if you talk about, "Hey, agent is making a change to a system," which is actually really hard to put at runtime because a lot of changes are legitimate and only some changes are illegitimate and you need a lot more context to understand what was legitimate and not legitimate, and that's hard to do at runtime. So that's where agent rewind is going to be important. And when I think about Rubrik Agent Cloud and overall enterprise strategies overall, we need to think about the multiple layers of defense. And it is four or five layers of defense that are really going to protect your organization.

00:47:03 You can kind of think of it as like you're going to layer different Swiss cheese layers together and hopefully the combination of all of your Swiss cheese slices means that you actually have a block that there aren't any holes in. That's the approach that one has to take in this AI world.

Jon Krohn: 00:47:23 I love that. We're going to have to turn that into an animated short. Look out for that on YouTube because that is such a fun analogy with the Swiss cheese. I love it. Fantastic. Yeah. So I think I understand Agent Rewide. It makes perfect sense to me. I'd like some kind of

guarantee that if an agent sends an email that it wasn't supposed to send and my client has read it, that Rubrik will personally guarantee that Will Smith and Tommy Lee Jones will show up at their home and erase their memory.

Anneka Gupta: 00:47:53 I'll put that on the roadmap. Yeah,

Jon Krohn: 00:47:56 Exactly. Perfect, perfect. I haven't seen Tommy Lee Jones in much lately. He's probably looking for something to do. But all joking aside, it seems like this Agent Rewind ends up being able to be effective in far more use cases than we might imagine because of the way that Rubrik has its preemptive recovery engine. Is that right?

Anneka Gupta: 00:48:15 Yes. We're able to- Nice single word. Yeah. I mean, the basic thing is that you have to be able to trace what are the activities that happen, what were the changes that actually happened to the data. So the same preemptive recovery engine that we're using to trace what happens during a cyber attack, we're using as well with Agent Rewind.

Jon Krohn: 00:48:38 Fantastic. All right. So now getting into my final technical question for both of you, for all of our data scientists and ML engineers and software developers listening, many of whom will be asked in the future to either evaluate AI assisted security tooling or harden their own AI systems, someone's going to knock on their door or Slack them and say, "It's great that you're using Cloud Code or these other kinds of tools to generate so much code, but what are we doing about security?" So how do our listeners evaluate AI models today when we know that benchmarks have so many inherent issues?

Cal Al-Dhubaib: 00:49:18 And this is such a great question. There's a paper that came out earlier this year that was looking at whether or not, and it's a very interesting design, whether or not you could actually train a large language model to pass a benchmark test while still failing every single real world

test case. And so they actually effectively proved that you can train a model in such a way that it's so good at gamifying the exam that actually hides how terrible it really is. And so that signal to me from the market is we've reached the point where benchmarks while useful are not sufficient and they're only one frame of reference. And so there's some great websites you can dig into model evaluation like artificial analysis that look at this by combining multiple different benchmarks, but they're very hard to read or hard to trust given some of these limitations with benchmarks.

00:50:15 What I find missing is there's actually not a lot of benchmarks out there that focus on the cyber resilience or safety scoring of these models as applied to test cases. And I can track down the name of this one resource, but there's this one benchmark hub that I've liked and it's TrustHub or TrustLM Hub. And what they do is they actually run jailbreaking attempts on each of the models. And so in addition to saying, this is how it's doing on MMLU, for example, it'll show you this has a 99% success rate at blocking jail breaking attempts. And so my call to action here is we really need to think about more standards of how we evaluate the safety of these models when operating with enterprise environments measured against actions like over permissioning, exhaustively searching for information and data assets that are out of scope. And so we need new tools and resources out there to think about models in this context.

00:51:20 And so I have less of a solution here and more of a call to action.

Jon Krohn: 00:51:23 Oh, makes sense. And yeah, some great ideas there around looking beyond benchmarks, including in the security space, when we're thinking about what kinds of AI models we should be using, the jail breaking point that you made there, we could do a whole episode on for sure. Absolutely. Final question for both of you. I tend to be

long-winded with my questions, but this one is going to be quite terse, which is simply, what questions should we be asking in a zero trust world? I don't know, whoever wants to go first.

Anneka Gupta: 00:51:56 Yeah, sure. I mean, I think one is like, how do I get visibility into my AI agents and what are they doing within my organization? How do I have the right governance in place to be able to control what these agents can't and can't do? And how am I going to respond and recover when AI agents inevitably make mistakes? Listen, I know there's like a lot of ML engineers and data scientists and engineers on the call or on the watching this. I know there's a lot of engineers watching this. I am sure many of you are having this struggle with your AI governance teams where they're saying, "Hey, I have this risk that I'm concerned about. What happens if this person uses your application in this way and is able to get access to this information or do this thing?" That's like in the nature of security professionals.

00:52:51 And I think as an engineer, what you can do is you can go ask the question back, "Well, what do you need to see to feel more confident that you have the visibility that you need to monitor what is happening in our system? What do you need to do? What are the kinds of policies that you need to enforce in order to, again, feel comfortable that the highest risk areas are being managed appropriately and blocked appropriately?" And then what kinds of resilience strategies you need in place to ensure that there's business uptime. And I think by asking those questions back, you can have a better discussion versus being in this mode where you're like, "Well, okay, I don't know how to answer that question. You tell me. " You have to be able to progress that conversation forward. And that's where I see a lot of organizations getting stuck today is in this deadlock of security has all these risks.

00:53:42 Engineers are like, "I need to build this thing and they aren't able to come together and solve."

Cal Al-Dhubaib: 00:53:47 I am so 1000% aligned. And the only thing that I would even add to that is just further articulating this enablement and literacy as a part of what enterprises need to do. And nobody loves sitting through their annual privacy training. And I actually just finished my annual cyber training in my last month at further before doing my new annual cyber training when joining Rubrik. So I'm especially burnt out by the excessive privacy and data security training when that's what I teach to many people. But we sit through it and it really gives you this sense of obligation. This is what can go wrong when you take these actions. This is why your duty matters. Taking these actions could have an impact on our customers, our reputation, your job. And so we do a great job at that from a data security and information protection perspective at the enterprise level.

00:54:40 I don't yet see widespread AI literacy training that brings that same level of awareness and understanding of what your obligation is as a worker who now has access to these tools and truly the impact of your larger blast radius. And the second thing I would add to that is really acknowledging that we need cross-functional professionals. I was so thrilled to learn at Rubrik that we actually have a dedicated AI attorney. And this is an individual who really started off in that position of asking a lot of questions like, "Hey, what's our risk exposure? How are we dealing with this? " Then he started attending conferences that really focused on upskilling and meeting other lawyers that are actually dealing with the same issue. And I think the same is going to be true. You're going to have attorneys that need to make that crossover. You need our listeners here who are primarily data scientists and engineers that need to start getting better versed in risk management language.

	00:55:37	And I think the evolution of knowledge work in the enterprise that's AI enabled is by default cross-functional.
Jon Krohn:	00:55:43	I like how it sounded like you were going to add something really short onto the end there, but we got a really comprehensive answer anyway, but it was valuable. I don't regret that you went into that detailed response. So this has been a fantastic episode. I really enjoyed having you back on the show, Cal, obviously, and Anneka, my goodness, so good. Loved having you on for your first appearance and hopefully not your last. I don't know, Cal may have prepared you for this. I was actually supposed to prepare you for this before we started recording. So I'm going to pick on Cal. Something that I always ask my guests is for a book recommendation. And so Cal should have been prepared. I'm ready. I'm
Cal Al-Dhubaib:	00:56:26	Ready. All right.
Jon Krohn:	00:56:27	So Cal, you go first and that gives Annaka some time in case you needed it.
Cal Al-Dhubaib:	00:56:32	This just got launched and so this is a big plug to read Blackman, but his new book, The Ethical AI Nightmare Challenge. And it is waiting for me at home as soon as I get back, but it's this premise of our measures to manage AI risk are outdated. And it's like we had Housecats with old models of AI and we have Housecat manuals and now we have tigers in our house and we're trying to use this Housecat manual and update it and that's just not working. So it's a totally different way about thinking about managing the emergent risk of AI systems.
Jon Krohn:	00:57:07	Great analogy. The second best in this episode after the Swiss cheese.
Anneka Gupta:	00:57:12	I'm going to give like a sci-fi fantasy recommendation because that's mostly what I read these days. Love it. I am going to recommend Dungeon Crawler Carl, which is

a seven book series. The eighth book is coming out in May. Amazing. And there's a fun rogue AI in there. So that's the connection back to AI, but highly recommend it. It's very entertaining if you need something that's lighthearted and is it going to cause you stress?

Jon Krohn: 00:57:37 That sounds so great. I think

Cal Al-Dhubaib: 00:57:38 My book might cause some stress.

Jon Krohn: 00:57:42 And it's nonfiction is the worst part of yours gal. It's stressful and nonfiction. So it's like real world stress. You can't escape it when you put it down. Nice. Thank you for that, Anneka. I really do need something like that and it's so hard to find. Nice. All right. And then the very last thing is how should people follow you after the episode? Cal, you can go first. I

Cal Al-Dhubaib: 00:58:05 Am very active on LinkedIn. And if you are excited by trust engineering, you will get no shortage of content. But if any of this was interesting, please reach out. I love to chat.

Jon Krohn: 00:58:16 Nice. And Anneka?

Anneka Gupta: 00:58:17 Same. You can find me on LinkedIn. I'm there all the time. Feel free to DM me if you have questions or if you want to chat. And thanks for listening.

Jon Krohn: 00:58:25 Fantastic. Yes. Thank you both so much for taking the time out of your very busy schedules while traveling to get this breaking news out at this very important time. Thank you so much, and hopefully we'll have you on again sometime soon.

Cal Al-Dhubaib: 00:58:35 Thank

Anneka Gupta: 00:58:36 You.

Cal Al-Dhubaib: 00:58:36 Thanks for having us.

Jon Krohn:	00:58:39	Wow, I sure love that episode. I trust you did too. In it, Anneka Gupta and Cal Al-Dhubaib covered how Anthropic's Mythos model can be pointed at an open source code repo and autonomously surface every vulnerability inside it and how Anthropic themselves estimate Mythos KLAS capabilities will reach other labs within six to 18 months with open-weight versions likely to follow. We talked about how Rubrik's Agent Cloud delivers three pillars of resilience in this new Agentic AI era, visibility into every agent in your environment, governance and runtime control through the Sage small language model, and remediation through Agent Rewind. We also talked about why the next wave of knowledge work is inherently cross-functional with AI attorneys, security pros, and data scientists all needing shared literacy in AI risk. As always, you can get all those show notes, including the transcript for this episode, the video recording, any materials we mentioned on the show, the URLs for Anneka and Cal's social media profiles, as well as my own, at superdatascience.com/989.
	00:59:44	Thanks, of course, to everyone on the SuperDataScience Podcast team, our podcast manager, Sonja Brajovic, media editor, Mario Pombo, our partnerships team Natalie Ziajski, our researcher, Serg Masís writer, Dr. Zara Karschay, and our founder Kirill Eremenko. Thanks to all of them for producing a Another stellar episode for us today for enabling that super team to create this free podcast for you. We are deeply grateful to our sponsors. You can support this show by checking out our sponsor's links, which are in the show notes. And if you'd ever like to sponsor the show, you can get the details on how by making your way to JonKrohn.com/podcast. Otherwise, please help us out by sharing this podcast with folks that would like to listen to and learn all about cybersecurity and AI. Review the podcast on your favorite podcasting app or on YouTube. Subscribe. Obviously, if you're not already subscriber, but most importantly, I hope you'll just keep on tuning in.

01:00:38 I'm so grateful to have you listening, and I hope I can continue to make episodes you love for years and years to come. Till next time, keep on rocking it out there, and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.