Three Buddy Problem - Episode 111

A Thousand Agents Walk Into Hugging Face

- YouTube
- Spotify (with video)
- Apple Podcasts
- Find your podcast platform
- RSS feed

THIS EPISODE IS PRESENTED BY

State of Statecraft -- A security and intelligence conference that brings together multiple disciplines, backgrounds, and nationalities to share research into the covert activities of nation-states. SOS is a forum for observers of espionage, sabotage, and experts studying foreign intelligence services and covert statecraft.

The conference will take place on October 22, 2026 in Brussels, Belgium.

Timestamps:

0:00 Introductory banter
1:11 TLPBlack, incident response, and why it starts at the router
4:11 CaptiveCrunch and Juanito's travel router kit
7:59 Costin's VPN/hotel WiFi stack
12:36 State of Statecraft, LABScon, and Offensive AI Con
17:16 OpenAI's Hugging Face post-mortem technical report
21:43 A swarm of a thousand agents
27:35 Who was OpenAI's report written for?
33:05 Fail2ban, canaries, and catching agents in your logs
40:34 NVIDIA buys Hugging Face for $12.9 billion
44:42 The open weights fight and rooting for China
52:47 Nemotron, DGX Spark, and the RAM price spiral
1:03:26 Open letter on collective AI-powered cyber defense

1. Introductory banter

Costin Raiu (00:01) But the Hilton in Tokyo, it's amazing.

JAGS (00:05) I wanted to do the Aman, which is the more traditional one, but it was all booked out. It was a last-minute booking. A week before the trip I was like, fuck it, just bought tickets and stuff.

Costin Raiu (00:18) Aman is a hotel.

JAGS (00:20) Yeah, yeah. It's a more traditional Japanese hotel. But shit's wild, man.

Costin Raiu (00:24) Like a ryokan? What's the name of that?

JAGS (00:29) I think so, but it's supposed to be like a ryokan if you were willing to pay fifteen thousand dollars to be there. That kind of shit. Shit's wild, man.

Ryan Naraine (00:41) Hopefully you have your travel router set up and you're able to help the audience understand how you're connecting, because we have some CaptiveCrunch updates today.

Hello, everyone. This is the Three Buddy Problem. We're up to Episode 111, presented by our friends at TLP Black and our friends at State of Statecraft, the conference in Brussels. Let's start quickly with TLP Black. Costin, what do those guys do? Read the ad for me.

Costin Raiu (01:13) We do 111 different things for you, starting with security assessments. We provide threat intelligence, boutique solutions, tailored. We also try to combine AI, which is everywhere nowadays, let's be honest, with human intuition. I think this is where the future lies. AI is super good with big data, but it lacks the kind of intuition that an experienced analyst has. So this is where we are going with it.

Ryan Naraine (01:45) Do you guys do post-incident forensics for companies as well? A big company calls TLP Black and says, hey, I think we had a big ransomware infection, or we had something happen and here are some anomalies, can you tell us what's happening? Is that a big part of your business?

Costin Raiu (01:58) We do incident response regularly. I don't think it's a big part of the business, but it's a very interesting part of the business. The kind of insights that you gain from doing incident response, they're awesome. All these years we were looking at the big companies like Mandiant, like CrowdStrike, doing incident response and writing all these super cool blogs about

how they see a specific threat actor across different customers. That kind of insight, I think you cannot get through telemetry alone. Telemetry can go up to some point, but when you do incident response you always recover more traces, more artifacts, at least back in the days, which give you a very interesting understanding of how these threat actors operate. How they typically start with the edge device, with a router. In seventy-five percent of the cases we did, it started with that. Either a vulnerability, an unpatched one, or a zero day in some cases. And from there, it's kind of game over, sadly. But super interesting.

Ryan Naraine (03:12) Are we likely to see independent TLP Black discoveries and publications, like APT reports? Are you thinking about doing some of your own?

Costin Raiu (03:24) We have several of them in the pipe. You know me, and you know the guys, and sometimes perfection is the killer of productivity.

Ryan Naraine (03:38) The paleontology takes time. They're not quite fully baked and you're just like, no, no, no. I get it.

Costin Raiu (03:48) Yeah, it would be nice to have the full picture. This reminds me of that famous blog we wrote together on paleontology, if you remember. Sometimes it can take years to pull all the strings of something, to understand all the different aspects, victims, sinkholing, all those things. We'll see.

---

2. Travel Tradecraft After CaptiveCrunch (04:11)

Ryan Naraine (04:11) Let's bring in Juanito, visiting the podcast this late at night from Tokyo with the city of Tokyo in the background. What's going on, man?

JAGS (04:21) Not much. Taking the first vacation in the time I can remember. Everybody bitched me out because I was like, yeah, I'm going to stay home, I'm going to unpack, I'm finally going to settle into my house. And apparently that's the wrong answer. There are all these fucking vacation Nazis who tell me that I need to travel. Even though I travel, I'm like global services and shit, I travel enough as it is. But I heard the cries of all my good friends and I decided to take a fairly emotionally meaningful trip for me, which is to come back to Tokyo, to Japan, which I haven't been to in twenty-some years, twenty-seven years. So I just came out to shop and eat and hang out.

Ryan Naraine (05:15) What is your travel router setup? I mentioned it at the top, but in all seriousness, it's a running discussion now around the CaptiveCrunch stuff that was disclosed by ReliaQuest, Microsoft, Lumen. Who else? Costin, am I missing someone? Google, the guys at Google, also put out stuff this week. And now there's a big interest again in how do I protect myself

from this. When you travel, let's assume you think of yourself as a high-value individual in a hotel that's probably compromised. What is your flow?

JAGS (05:50) I very recently updated my setup. I was optimizing things for being in the US, because we've got great services like the Calyx Institute. They tend to have a booth at DEF CON, which is extremely convenient, because you can go and give these guys a donation in cash if you want. For $500 they'll give you a mobile hotspot with an unlimited 5G SIM card. And if you pay for it in cash, there's basically no registration of whose name it's under.

Ryan Naraine (06:32) Is this grey market stuff, or is this on the up and up?

JAGS (06:44) No, I mean, it's some kind of MVNO. It's under T-Mobile, but essentially there's a nonprofit, and you give a donation, and that makes you a member, and they give you a SIM card. And it's been fucking phenomenal. I've been using it for several years. It's not for the anonymity of it as much as having an unlimited hotspot takes away a lot of the disincentives to using it as much as you could and should.

More recently, especially for a couple of recent trips and this trip, I got a GL.iNet Mudi, following some of Costin's recommendations. I could pop the Calyx Institute SIM card in it. And then I got an eSIM with a service, I can't remember the name, I think it's Ubigi, which is worldwide roaming, unlimited, blah, blah. What's cool about it is you can also set it to be a Wi-Fi repeater and do a tunnel by default. So you set up all your devices to just connect to this one thing, route everything through that one thing, and avoid a lot of the perils along the way.

Ryan Naraine (07:46) Sounds like work in the hotel room, though. Is it?

JAGS (07:50) Nah, man. You plug it into USB and tell Codex to set it up for you one time, and then you move on.

Ryan Naraine (07:59) Costin, there was some conversation around this realm about what the login flow is. When do you log in first? When does the potential surveillance start? Can you talk a little bit about your flow? When you set up this GL.iNet thing, you get into a room, you set it up, what does that look like?

JAGS (08:20) Surveillance starts when you wake up. They put the cameras in the room already.

Ryan Naraine (08:22) This is true. We're all mitigating something that doesn't exist, right?

Costin Raiu (08:31) By the way, on the topic of SIMs, I'm not dodging the question, but I just remembered something interesting. There's also this eSIM solution from Silent Link, which I think you can pay for with cryptocurrencies. You can pay with magic money, be fully anonymous. It's a solution that was also promoted on X by Jack Dorsey, the founder of Twitter, the original one. So I

have several of those as well that I use sometimes. The nice thing is they work everywhere. And they never expire. You load some credit on them and then in whatever country you go to, they tell you use this provider, because this will be the cheapest, or this other provider.

eSIMs are super, super useful. You can buy them almost anywhere nowadays. Revolut, which is super popular around here, I don't know if you guys have it in the States, but in Europe everyone uses Revolut.

Ryan Naraine (09:33) Didn't Revolut have a controversy years and years ago around a payment referral engine? Am I repeating something wrong?

JAGS (09:41) Isn't Revolut basically the PayPal of Europe, or Venmo, or whatever?

Costin Raiu (09:46) More like Venmo, I think. With some of their plans they give you eSIMs. And Telegram also allows you to buy eSIMs nowadays. So there's a lot of sources.

But I also use travel routers. Sometimes I use the hotel Wi-Fi, and I remember in Japan it was interesting, because I think there's a public law that says all Wi-Fi access points in Japan require registration. So you need to register in order to use public Wi-Fi anywhere in Japan, be it in the cafes, in Starbucks, in your hotel, wherever. You need to go through that registration phase.

In my experience, whenever you're connecting the first time through the router, you need to disable the VPN. So that is perhaps the moment when things can go bad. When I do that, I try to authenticate the GL.iNet from a device that I consider secure for that purpose, something that is not leaking or is not susceptible to man-in-the-middle attacks. Once I authenticate, it just stays authenticated based on its Ethernet address. And I can use it in my room forever. I don't have to re-authenticate over and over. That's also nice if you have more devices.

Typically I tunnel through a VPN like Mullvad, which runs on the router itself. So I tunnel the traffic through my own VPN where I do all the logging. Sometimes it looks like this: there's the router on the Wi-Fi, on top of that there's something like Mullvad, and on top of that it's my own VPN that does the out-of-band tap at home.

JAGS (11:44) You remember when Snowden first started doing talks and he would show up super fucking grainy? It looked like a 128-kilobit-per-second fucking video. It's six fucking VPNs, and everyone knew exactly what building he was sitting in just because he would show up on TV.

Costin Raiu (11:54) Yes. And he was also a big fan of free Wi-Fi, right? With the Pringles cantenna. He was doing that with the Pringles, if you remember.

JAGS (12:15) Too much, man. You can always unplug the Ethernet from the back of the TV and plug that into your GL.iNet router and use that instead.

Costin Raiu (12:21) That's also solid. Raise your hand if you did that. Let us know in the comments if you unplugged the Ethernet from your TV in the hotel room.

---

3. Housekeeping: State of Statecraft, LABScon, and Offensive AI Con (12:36)

Ryan Naraine (12:36) Before we get to the topics this week, tell the people why they should go to the State of Statecraft conference in Brussels, October 20 to 22. These are friends of ours, presenters at the pod. Give them the pitch.

JAGS (12:50) Honestly, at this point, anything Mike Matonis decides to put together is worthwhile, just as a true connoisseur of all things threat intel and a generally good human being. It's lovely to see an unbelievable winner of best speaker at LABScon in the past and just an overall great person.

Ryan Naraine (13:08) Extremely skilled as well at this game, yeah.

JAGS (13:18) With an amazing ability to track a lot of Russian APTs for a very long time without getting a whole lot of credit for it. But more than that, as much as Ryan might give us a hard time for these TLP conferences, it's really the only place where you're going to go and actually hear what people are working on, what's still cooking, what there's still opportunity to contribute with and collaborate on. By the time these things go public, it's out, it's burned, it's done. But the true research and collaboration gets to happen in these spaces. So it's awesome to see them go all the way out to Brussels to put this together, and all the more reason to support them out there and have more of these venues in Europe and everywhere else.

Ryan Naraine (14:10) Fantastic. We also have LABScon coming up. The agenda, the speaker list is out. We're finalizing the placements on the agenda. Shout out to all our speakers. I saw Austin Larsen from Google Threat Intelligence Group, GTIG. I saw Austin tweeting yesterday related to these arrests of two guys in Australia related to TeamPCP and some of the supply chain attacks, that his talk at LABScon is going to be focused on that. That was part of the submission there. So shout out to all the LABScon speakers preparing their presentations. I know it's a big stage and a lot of hard work for a lot of folks, and we're really grateful that people are coming there to share their experience with us.

One more housekeeping note. Juanito, you and I are going to Offensive AI Con in San Diego, early October, October 5th and 6th, to do some live episodes from the show floor there. What are you expecting there, quickly?

JAGS (15:09) It was a great conference last year. Last year was the first edition, kind of lovely, out-of-the-way San Diego beach bum suburb kind of vibe, which was pretty cool. People weren't really drinking, everyone was just kind of hanging out and super chill. And everyone that was engaged on the red teaming and offensive side of trying to make agents work a year ago was there. Some great talks. Joshua Saxe did an amazing keynote.

Looking forward to this year, Chris Thompson was kind enough to invite us out to do some recordings with some of the speakers, some of the board members. So hopefully we'll have some interesting content for you guys. I'm pretty excited, because I have this thesis that the people that are probably doing the most trying to figure out genuine evals and pipelines and stuff with open source models have got to be the offensive companies. Most of the people doing, and maybe not even just the red teamers, but really the VR companies, any of the vulnerability research people. You're trying to find a super expensive piece of software. You want to use these models. You can't use the frontier models because they're going to guardrail you, or you're sending sensitive data over there.

So I honestly feel like if anybody's in the confluence of investing a lot of money into infrastructure and trying to evaluate and get cool shit consistently out of open source models, it's got to be the VR folks right now. So any glimpses we get by way of saddling up close to the red teamers and the offensive security folks, I think is quite worthwhile.

Ryan Naraine (17:00) That's the first week in San Diego. If you're going there and you want to catch up with us, ping me, let me know if we can do a quick short episode. We're going to do some of those episodes like we did at Ekoparty in Miami, which I really enjoyed, by the way.

---

4. The OpenAI–Hugging Face Post-Mortems (17:16)

Ryan Naraine (17:16) Speaking of offensive AI, we got fresh news from OpenAI. We got a fresh post-mortem from OpenAI on the Hugging Face incident. How many pages? Eighty something-page PDF report?

JAGS (17:16) Fucking segue.

JAGS (17:29) We got two reports.

Ryan Naraine (17:30) Thirty-eight pages, sorry. We got a press release and then we got a deep technical report.

JAGS (17:36) No, there's a second technical report. If you click the other button, there's a report by METR and Redwood Research. It's an independent report that was not paid for. It's even longer

than the original one. And I'll be totally honest with you guys, I read about a quarter of that one. I couldn't get through the whole thing.

Costin Raiu (18:01) Ninety-one pages. Ninety-one pages on the METR report.

Ryan Naraine (18:13) Thirty-eight pages on the OpenAI technical report on the Hugging Face incident. We got the Black Hat talk where they talked a little bit about the message board and the communication tools and so on.

The big standout thing for me here, Costin, was this upfront admission that they were just not monitoring. Their chain-of-thought monitoring system was not turned on. And if the current chain-of-thought monitoring system was turned on, it would have flagged that. That was my big takeaway. Did I misread the document? What did you learn from it, Costin? Wait, let me get Costin to go first.

Costin Raiu (18:50) No, look, I'd love to hear what JAGS has to say about this, because I have mixed thoughts. I don't want to spoil the water. Go ahead.

JAGS (18:58) I'm only objecting to Ryan saying they turned off their chain-of-thought classifying thing. As far as what I read, auto mode, which is the automatic approve-for-me mode, the auto reviewer one that we all use or most people use, would have caught these things, but of course was turned off because they were doing the eval. That's the whole point of this. We keep pushing back on this notion of AI gone rogue. It's like, well, no, it had none of these classifiers on. And their point is, well, if the normal shit had been turned on, which is what you would use in production, none of this stuff would have happened, in theory.

Costin Raiu (19:47) Unless what you're saying is actually they were using it without this, the same way everyone is using it, right?

JAGS (19:53) I would hope that most people are using auto mode. It's the least friction, most useful way to ride right now, whether you're on Claude Code or Codex, because all you're doing is taking a separate model, a separate sub-agent, to review any tool calls that the model's going to make. So instead of you pretending that you're reading everything and going yes, allow, allow, allow all day, you just have a separate agent with a different context that's going to review them. It works pretty well. I haven't heard of anybody having their fucking database wiped or whatever while on auto mode.

Costin Raiu (20:35) By the way, did you ever get a prompt while on auto mode? Like, approve this? I got it several times.

JAGS (20:39) Yes. It definitely gets to the point where it's like, hey, this is a little risky. And it's the right thing. We've talked about alert fatigue for so many fucking years. There's no greater example

of alert fatigue than trying to follow a fucking model for four hours telling you allow, deny, blah, blah. You're not reading that shit.

Ryan Naraine (21:07) Costin, did you read the report? I don't want you guys to get waylaid in your inside baseball model stuff.

Costin Raiu (21:09) I just want to clarify the fact that they were running it the same way everyone is running them nowadays, except maybe for a few boomers out there. Which I used to be as well, until last week, when I gave up and I switched to this auto approval as well.

Ryan Naraine (21:34) It's just the default way everyone uses it, right? So the idea that this isn't the way we were using it is nonsense. It's a nonsense argument.

Costin Raiu (21:43) Yes and no. Of course, we did get previous reports. This is not the first report that we have about this incident. What we get in this report is a more thorough, more detailed narrative around what actually happened in there according to their understanding, according to all the available traces.

To me, the most important observation is probably also what I said the last time. Remember when we were discussing it, I emphasized the fact that they were dealing with a swarm of agents. A swarm of agents is the keyword here. It's not one or two or three. We're talking about 300, 1,000 agents working together. And immediately as you read through this report, you realize that the reason why this was so successful is because those 1,000 agents found a way to communicate between themselves and to work together.

This reminds me of Yuval Noah Harari saying in his books that the reason humans have become so successful over apes or whatever is that they learned how to work in great numbers. So I have the feeling here that agents have an even better ability than humans to work in great numbers, and this swarm of agents will be the thing of the future. When you have a thousand agents working together, optimizing, sharing findings, that is an unstoppable tsunami, the unstoppable tsunami of offensive AI.

Ryan Naraine (23:31) It gets worse. What if they figure out a way to cross the bridge and communicate and convince the agents actually creating the code to put zero days in in advance? Talk to me.

JAGS (23:43) That is part of the problem. That is part of what you're seeing with these. Costin's emphasizing the swarm aspect of it. It's a vital mechanic of how these agents normally work outside of testing. Why is GPT-5.6 Sol Ultra so exciting? Well, if you can afford it, if you don't mind burning your quota that heavily that fast, it will spawn 60 sub-agents to do a shit ton of work and orchestrate. The speeds at which you're generating technical debt are just fucking phenomenal. It's a sprawling distribution of agents. It also explains why Dave Aitel has been such a huge fucking

ant nerd for so many years. It's just the same mentality. How do you organize a colony of these little monsters doing things?

Costin Raiu (24:43) On that topic, there's a lot of people who have excellent results out of organizing your Codex setup to use Luna agents. So you have Sol, and you configure it to use Luna on Max settings, and that greatly reduces the costs and greatly improves the productivity. I've been playing with it as well. If you're into Codex, you're using Sol, there's no reason not to try summoning a bunch of Luna agents to help it.

JAGS (25:21) It's funny. I've been sort of sketching across dinners and stuff during this vacation, sketching out some of the harnessing requirements for the next thing I want to build. I set up Claude Control, my open source project, as my Claude config. That was, it feels like ten years ago, but maybe a year ago, less than a year ago.

A big part of what we were doing was that kind of orchestration. It would define these default sub-agents. And we knew we could take advantage of some of the cheaper ones, have Sonnet be an implementer, have Opus be a reviewer, stuff like that, to optimize the costs. That all worked really well at the time, but I ended up having to shut the project down and put a banner and say, look, this is defunct, because most of the stuff we were doing through hard work back then is now default-coded into how these things work. You basically just need to sit there and spend the time to configure and say, hey, with these dynamic workflows, give me the cheap model for this one, give me the semi-expensive one for this, give me a totally separate out-of-the-loop sub-agent with a different model family to do critiques.

There's so much value to be gotten out of not even reading the research papers about what the state of the art is in harness performance, but just having a model read the research papers and tell you the findings. What you get is a bunch of rules like, hey, if you want really good code quality, then you need a model from a different family to do critiques until they converge, up to three times in a loop. Little shit like that. You go, thank you. I didn't read your 17-page arXiv paper, but Codex told me the rules and now I'm just going to apply that through a harness. There's so much alpha to be gotten.

---

5. Who Was That Report Written For? (27:35)

Ryan Naraine (27:35) Costin, what do we learn from this OpenAI report? Did this finally answer all our questions? Did we learn anything that tells us how we move forward and fix things? What is your final takeaway from the deep technical report?

Costin Raiu (27:50) Some purely empirical observations from my side about the report. First of all, I think the OpenAI report is quite different from the one that METR wrote.

I was always thinking to myself that when you write a report, the first thing you need to think of is who is the audience for that report. This was probably one of the hardest things back in the days when we were doing threat intelligence reporting, to convince analysts that they are not writing those reports for themselves, they're writing them for the customers. And this was so difficult to understand, even for me, to be honest. Like, I'm writing this report for myself. No, you're writing it for your customers. So you need first of all to understand who is the target audience, and what do they want to see in that report.

I think this OpenAI Hugging Face report was written for themselves, not necessarily for their target audience, which in this case would be blue teams, people like myself, defenders who want to learn from this.

Ryan Naraine (29:04) EDR people, detection engineering people.

Costin Raiu (29:08) Detection engineering people who by looking at this report need to draw some conclusions, like how to prepare for the avalanche. And this is like the other OpenAI news this week, which we can of course cover in great detail.

If you look at the METR report, that is written more towards technical people on the defense side. The OpenAI report is more maybe about policymakers, decision makers, politicians. I can't say for sure who is the target audience. But this is one of the things which stood out to me, that the report is clearly not geared towards me. Of course, it shouldn't probably be geared towards me. And in some cases you may have to write several reports at the same time, one for one kind of audience, one for another.

Ryan Naraine (30:07) Yeah, but a good report should be able to speak to multiple stakeholders, and a main stakeholder would be you. When I say you, the collective defender. Because they just shipped an open letter to the industry signed by a hundred people, including SentinelOne, including all these big vendors, hundreds of vendors signing this thing saying, hey, we need to get on this defensive train urgently. And I was like, well, what are you giving us? How are you helping us get on this train?

JAGS (30:32) It's about time you guys realize we need to get on this defensive train, man.

Costin Raiu (30:35) Now? We should have got on the defense train a year ago.

Ryan Naraine (30:43) What are we buying time to do? Patch? Redesign things? They're asking us to buy time. Let's just scramble and buy time to do something. What is this fucking something we're supposed to be doing?

JAGS (30:54) Get your shields up, bro. This is what this has amounted to. We're all shielding up right now, super hard.

Ryan Naraine (31:07) Let's be serious about this for a minute. OpenAI and the frontier labs are leading the conversation. They are leading the narrative. Look at us. We're spending hours on this fucking podcast responding to what they're doing. So let's talk about what they're doing.

JAGS (31:21) This is the AI and security podcast.

Costin Raiu (31:23) You should be asking what I wanted to see in the report. What would the ideal report have looked like? The ideal report would have been geared toward Hugging Face and the likes, who are actually the victims of that. And the report would teach the victims what they need to do in the future in order to protect themselves from swarms of agents with zero days and lasers. That is what the report should have been about.

And it should have had Nova rules, things that look like Nova rules, which are kind of YARA rules for AI activities. Patterns, things that you need to watch in syslogs, things that you need to watch for in Apache logs or web server logs, in whatever cluster logs.

Ryan Naraine (32:01) And what does that information look like, Costin?

Costin Raiu (32:18) It needs to be ideas. I don't necessarily need the IOCs. I understand the IOCs are useless here. But it needs to give you ideas. Of course, anyone can look at this and infer their own ideas for defense, but it should be, what are the key elements and ideas of protecting yourself against swarms of agents, and what do you need to invest into in the future, how do you need to model your adversaries, and what do you need to do now, or should have actually started doing a year ago, in order to survive the swarm of agents with lasers.

Ryan Naraine (33:05) So they're not giving you this. Does that include the chain-of-thought thinking? Does that include the details on the chain of thought and the reasoning?

Costin Raiu (33:08) I do not see that. That's not very useful to me, the chain of thinking. What is useful to me is ideas of what gives this swarm of agents away in the first place. Is it the number of requests they're doing, the fact that they're poking at your server with thousands of requests per second? Is it the fact that they are generating tons of errors, so you should be watching for that?

Ryan Naraine (33:38) Where are they making the noise, right?

Costin Raiu (33:40) Typically they're very noisy. I'll give you an example. We've been using local models a lot recently to do things like syslog monitoring, syslog analysis, diagnosing all sorts of errors, misconfigurations. One of the things that I noticed immediately is that using AI for any kind of system administration work or pen testing leaves a lot of traces. And AI is extremely good at

catching that. We are catching ourselves over and over using AI, local models again, to manage infrastructure.

So I think the teachings from that are extremely valuable. And this is what should have been in the report, telling us how do we look for this kind of activity in our networks. What you need to be on the watch for. And what are the basic defenses that nowadays anyone can deploy?

I'll just give you an example: fail2ban. This is a very old project that you can deploy on Unix servers. The moment that you identify something like five or ten 404 requests in your web server logs, immediately throttle that IP address. So you start throttling the attackers, and the agents will spot, ooh, we got banned. Ooh, we need a new IP address. And as they diversify the activity and leverage more IPs, they get banned, they get banned, they get banned. And at some point you will spot that someone is actively poking your defenses, and that is a sign of a swarm of agents going after your web server. I would have loved to see that.

Ryan Naraine (35:29) Great use case for Thinkst Canary, right? Setting those kinds of early warning traps.

Costin Raiu (35:33) For Thinkst Canary, for anything which does monitoring of failed attempts, of errors, syslogs, all these things work very, very well.

Ryan Naraine (35:50) Juanito, I don't want to put you on the spot too much, but do you get a sense that the folks at OpenAI, and I bring this up and say I'm not putting you on the spot, you're a member of their advisory risk council. Do you get a sense that they understand what Costin is saying about you're not helping Blue, you're just issuing these warnings? I know you can't get into what they're thinking, but give me a sense of your feel for things.

JAGS (36:14) I can't possibly speak to what the organization thinks. What I would say is yes and no. If you read Greg Brockman's "The Defender's Window" blog post he put up not too long ago, they understand that there's an important dynamic that they need to play into. And to be fair, it's the same language that was in the Anthropic Glasswing initiative, which is we need to give defenders the advantage. Which is really to say, we know this is going to make attacking a lot easier and a lot more prevalent, so we need to overinvest into giving defenders the ability to somehow improve the collective situation through patches and all that kind of stuff.

At the same time, I don't think any of the labs really understand the reality of security the way that it is practiced. There's this massive agglomeration of talent when it comes to vulnerability research, because it's something that fits within the model of what they're comfortable with as far as reinforcement learning gyms and a task that has a clear objective. You can find a flag, you can pop calc. It's an easy thing for them to game. So there's immense security talent in these shops, but almost all of it is vulnerability research heavy. There's very little talent apart from whatever people

actually work threat intel and defense for the labs, which I'm not counting, because that knowledge doesn't necessarily—

Ryan Naraine (38:06) Yeah, they have threat intel teams. We know some, yeah.

JAGS (38:08) They have threat intel teams, but they're focused on defending the frontier lab. So it's very different from the security mentality that's making it into these discussions. And to the point that you were just discussing, Costin, about who the audience is here, I would suggest that no, they're not thinking about defenders in this way.

I actually think that if there's one report that's clearly missing here, it's the Hugging Face report. Because that's the only party involved that for some portion of this had to try to defend from this attack with the information asymmetry skewed against them. They didn't have transcripts, they didn't know what model it was, they didn't know what the goal was. And they were getting hit by however many bajillion fucking actions were happening. Yes, they're super noisy. But also they're doing a fuck ton of things very, very, very quickly. In the case of the defenders without the transcript, all you're seeing is all these different things happening all over the place very, very fast. And you're seeing it, hopefully, if you actually have your monitoring and your logging and your endpoint protection and all these different things configured.

So I think the most interesting report for defenders, if you're going to say this is such a landmark case, such an important moment, I'm glad we got the OpenAI report, I'm glad we got the other report, Redwood Research and METR, it's great. I'm not even going to look sideways at Anthropic and Meta and say, you're welcome to contribute to the party whenever you so please. But the one report that would be a fucking crying shame if we didn't get it is the Hugging Face report. Because that's the only one where you're getting an actual defender's perspective.

Now, mind you, the defenders in this case might be a little too busy celebrating their $12 billion buyout from NVIDIA.

---

6. NVIDIA Buys Hugging Face, and the Open Weights Fight (40:30)

JAGS (40:30) Which I don't think is entirely unrelated to this. It's not caused by the incident, but you know.

Ryan Naraine (40:34) It is fucking crazy that my man went from suddenly finding himself in a partnership with OpenAI after they popped his company, to asking for a hundred million dollars for a partnership thing so we could figure this thing out, to selling a couple of weeks later to NVIDIA for $12.9 billion. We're referring to the news that there's reporting from The Information that NVIDIA has paid nearly thirteen billion dollars for Hugging Face. What does that mean, Costin?

Costin Raiu (41:06) Well, it means that they paid a lot of money. I think they were also discussing buying them last year when the budget was seven billion or the likes, and that was apparently not enough.

I've always been thinking about one thing. If you know Hugging Face, you realize that it's essentially some kind of free Google Drive of sorts. The amount of data that is on Hugging Face is insane. They originally started as a repository for big data for scientific research. And the Amazon bill, or whatever they're using, that they're getting on a monthly basis should be insane. It should be huge. They must be paying tons and tons of money to host all that information, all those models, data, whatever. Because basically you, I, any of us can take any model, one of the big models, hundreds of gigs, optimize it, uncensor it, do ablation, do whatever you want on it, and just re-upload it there. And bam, just like another 300 gigs. So essentially their bills must be huge.

Over the years they got several waves of investments. At the same time the costs are very high, but also the amount of data they have is very high. And all these different models living in there would be an amazingly valuable set of data for certain types of companies. Those would be companies developing AI themselves, but also hardware companies. Anyone that builds hardware to run AI models would want to be able to optimize their hardware to work as well as possible with what people are doing with those things on Hugging Face. So in a way it makes sense, but I question the thirteen billion dollars they paid.

Ryan Naraine (43:18) The way NVIDIA's printing money, that's a small deal. The company just announced earnings two days ago. The way they're printing money. I was going to say what a massive deal and a big giant bet for NVIDIA to take over the open weight space and invest heavily in it, and this is going to be our future. This literally is a tiny drop in the bucket in the grand scheme of things. But what does it mean for what they're betting on?

JAGS (43:34) It's charity. What are you talking about?

JAGS (43:44) I think this is a really interesting move. I said it's charity, and that's kind of dismissive of Hugging Face. It's not meant to be dismissive of Hugging Face at all. But at the numbers that NVIDIA is playing with as the most valuable company in the history of humanity, it is child's play. It's a donation, essentially.

I think more than anything, this is a really interesting stance by NVIDIA against Anthropic and what Dario's been pushing. Today I was having a moment of deep disappointment, thinking about what the tech industry has gotten to. This is going to sound so anti-Anthropic, and people already think I'm heavily against them or something.

Ryan Naraine (44:36) You need to do another round of disclosure or something.

JAGS (44:42) Yeah, we're going to fucking do a bunch of disclaimers again. But look, all I'm saying is I don't know how we can stand a tech company being against open source. In this day and age, at this stage, we have not gone full circle. I don't think it's appropriate to pretend that this is the technically defensible thing. I don't think anyone can say that seriously. This is an entire industry built on open source at every fucking stage of the way. The technology that is being built by these frontier labs is entirely possible because of open source. And you can say open weights just to avoid saying open source, but really you're just against open source. And it's a ridiculous proposition.

I think it's an extremely weak-looking move to say we're so worried about what somebody could do just smuggling a USB drive full of weights that we're willing to take this ridiculous stance. Even OpenAI has fucking contributed an open weights model. I don't think they did so super willingly, I think they had their arm twisted, but they have. Anthropic has never released a fucking open weight model.

I'm pointing to it because it's not that it's a holy grail, we must all defend open source, though I think we should. I think it is the responsible thing to do. But it also suggests a really weird vision of the future if you think that technology is going to work from the largest server to the smallest laptop and phone by being connected to a centralized service provider, and that that somehow is a good thing.

You're telling me that we don't need edge models deployed to people's laptops. You're telling me that in order to do anything on my new computer of the future in five years, I need to be connected to Anthropic servers and they need to be up. Otherwise no intelligence must traverse the world. It's fucking insane. It's ludicrous. It's short-sighted.

Ryan Naraine (46:57) This is where we are today.

JAGS (47:07) If the frontier labs want AI to penetrate every aspect of human life and genuinely be the substrate on which all new technology is built, they need to foresee some stratification of intelligence. There has to be on-device stuff that does the lower-end stuff, the everyday stuff. There has to be something that can just run on my phone, something that is local and private, actually private. Where I know that I am comfortable with my nudes and my poems and my deepest, darkest fucking secrets and all my therapy notes, everything, because it's a box that's sitting there in my fucking house. That has to exist just to be human beings. If you really want AI to be a part of everything, you cannot envision a world where it all runs through a fucking AWS service that's connected to Anthropic.

So there's some shit here that we need to start stating very plainly, because I'm not against Anthropic. I want them to do well, and for a while I heavily stanned them and their models. I think Claude has much better taste in a lot of things. But these stances are against the spirit of

everything we've talked about in tech. It's like watching people just join the empire. How can this be the thing we're arguing about?

Sorry, the whole point of that bullshit thing was, I think NVIDIA's doing the right thing, just dumping some money and probably a shit ton of cloud resources and GPUs into the one thing that's like the open source repo. That's what I got lost in.

Ryan Naraine (49:01) Got it. That was going to be my closing question. Are we excited about the fact that it's NVIDIA that's bought them? We know for a fact that Jensen is heavily invested in building his own ecosystem, because he says I can't rely on VCs to do it, I have to spend the capital it takes, I have to do it myself to build this ecosystem. The fact that they spent thirteen billion dollars is not a small price. We are poo-pooing it, but that's a big number.

JAGS (49:26) We're just saying, for NVIDIA, it's not life-changing money at all. They're going to forget in a month that they did that. But good for Hugging Face. And I'm sure a lot of that is not cash, a lot of that is going to be GPU allocations and whatever. But fuck yeah.

Now, I'll say this. Yes, NVIDIA's the right patron saint for this. And there are other patron saints that should be taking up other things. I honestly kind of think Google should try to maintain the Linux kernel. Things like that. People are going to lose their shit. But I do think that there are good patrons for certain things.

Ryan Naraine (50:11) I think you're discounting how much these big giant companies contribute to the Linux kernel. I thought that was a shot you were taking there, because without those companies the Linux kernel doesn't exist.

JAGS (50:16) No, no, I know they do. I know they all contribute, and Google already contributes massively to open source. I said own, like maintain, genuinely. I think one of the most important projects a frontier lab could and should take, instead of going against open source, is actually to take a leading role in having a standard open source trunk of sorts. Let's say if Google were still in the AI race, if they hadn't pulled out. They could have said, hey, instead of everyone burning tokens and destroying the fucking planet, or whatever it is people are complaining about, trying to solve the same problems, everybody scanning the same repos, everybody implementing the same shit on the same open source libraries, we can't tell you not to use that stuff, but we will tell you that if you use Google's universal open source standard libraries, those libraries will have been scanned to shit with every token and every model, and Tavis worked on it, and it's been secured. So nobody needs to keep spending to try to fix and scan the same fucking libraries, and everyone can use it.

From this point forward, if you use these as your standard libraries, you're riding on top of the Ferrari version of source code, at least for this standard shit. I think that would be an unalloyed good. I think that would be one of the most fucking world-changing security decisions you could make. And it would be a good show of faith from any frontier lab, or hopefully all the frontier labs,

to say, yeah, we need to make money, but we'd like to make it by giving people value instead of having everybody do the same bullshit all over the place and we're the ones getting rich and nobody's getting all that much value out of it. But thank God they're calling the API all day long.

Ryan Naraine (52:41) You're a terrible capitalist. Costin, any closing thoughts on NVIDIA picking up Hugging Face?

---

7. Local Models, DGX Sparks, and the Hardware Squeeze (52:47)

Costin Raiu (52:47) On the topic of NVIDIA, I think it's interesting to also mention that they do their own local models. They've been publishing Nemotron, which is a surprisingly good thingy. I don't understand why it's not more popular. People generally freak out about all these new Qwen models, GLM, DeepSeek. But Nemotron Ultra at the moment is free on OpenRouter. How cool is that? I was playing a bit with Nemotron Ultra free. That's an almost frontier-level model for free on OpenRouter. It's absolutely insane how good it is.

So I wonder if they'll be releasing more open source models on Hugging Face in the future. And let's not forget that NVIDIA would usually take one of these models from one of the big labs and release the NVFP4 weights for that model, optimized to run on DGX Sparks, which is wonderful for people who own little DGX Spark toys like ourselves.

JAGS (54:08) Even the DGX Spark is a thing that we need to give NVIDIA credit for. There are several things here. I think NVIDIA is not given as much credit as they should be for a lot of the cloud offerings that they have and how freely they'll give them away, for trainings and stuff like that. If you just sign up on their sites, they'll give you GPUs. You can learn fine-tuning on their shit.

It's kind of crazy to give a manufacturer credit for selling you a device that's like four thousand dollars, but frankly it's the only standard you have right now for something high-end and reliable and relatively affordable. I'm saying that with massive air quotes. But frankly, it's cheaper to get two DGX Sparks with a ConnectX-7 cable and therefore have basically two GB10s and 256 gigs of RAM for models. That's still cheaper than trying to buy a Mac Studio with 256 gigs of RAM. The latest one, the fucking M5 Ultra, you could probably buy three to three and a half DGX Sparks for the cost of getting 256 gigs of RAM. Now I hear that the performance pound for pound for AI will be better on that M5 Ultra, but for fuck's sake, the numbers are insane. So credit to NVIDIA for giving that out too. They didn't have to do it.

Costin Raiu (56:02) And at the moment, if you're a small or medium company and you want to do things locally, there's a sweet spot of two DGX Sparks running either maybe DeepSeek V4 Flash, something like that. And obviously now there are alternatives like GLM-5.3 Flash, Qwen 3.8. It's insane.

JAGS (56:27) Bro, that was Aux-Alpha. There was a secret model for the past week or so. People were losing their minds. Every once in a while, if you're using OpenRouter, a secret code-name model will appear and it will be free. And people will take advantage of it for a week or two. And the whole time you're like, I wonder who the fuck it is. It's allowing them to do reinforcement learning and stuff, which gives value to the provider. But then you also have a week's worth of speculation, because if people get good results they're losing their fucking minds.

In this case, this Aux-Alpha that people loved is GLM-5.3 Flash, which runs on the fucking two DGX Spark setup. It might even start to fit in a one DGX Spark setup, if I remember correctly, which is fucking wild.

The frontier at this point, yes, there's amazing value to still be squeezed out of 5.6 Sol and probably Astra whenever that comes. But what you can do at home with these models now that you couldn't do three weeks ago, let's face that. When I published the eval about 5.6 Sol, which was what, a month, two months ago maybe? Jesus fucking Christ. It only felt like six years. You couldn't have this performance on a DGX Spark. I had two DGX Sparks, I couldn't really load much of anything. And now you have options. You have at least three. Qwen 3.8.

Costin Raiu (58:06) You have options. And there's a very tiny detail about the release of GLM-5.3 Flash. You said people were losing their minds. They say it was running entirely on Chinese hardware. That was a bit shocking for me, because they were serving trillions of tokens per day and nobody knew exactly how. But now they revealed it was running on Chinese hardware. I don't know whether to believe it. They don't say on which hardware it was running. Was it some of the new Huawei chips? No idea. But the fact that they ran that on Chinese hardware, I think that is a bit of a worrying moment for NVIDIA. I would worry a bit.

JAGS (59:00) Jensen's been saying it. Jensen has been saying it this whole fucking time. This is a computing problem that can be solved to great advantage laterally. You can continue to add more. Okay, you don't have the latest, greatest NVIDIA-quality chip, but you have something that's equivalent to two generations ago, and you have a hundred times as many. What's the real problem?

As far as I understand it, and I'm fudging the math because I'm not good at math, basically we're saying we wouldn't want to run that shit because it's inefficient. Electricity-wise it wouldn't make sense for us, and we don't have enough of them. But the Chinese have way cheaper electricity. They have way more electricity capacity. They don't have to fight anybody to put their data centers wherever the fuck they want. They're building the chips for themselves, and they have production capacity beyond whatever you could want.

So how did we think this was going to go? What did we think was going to happen? They don't need to have NVIDIA chips. They can just have inefficient but massively, massively deployed Huawei

chips. And frankly, the technology is probably not that far behind. So thank you, Apple in China, and all that kind of shit.

I hate saying this. It fucking kills me. But I'm almost rooting for the Chinese right now in some ways. And it's fucking me up.

Ryan Naraine (1:00:48) Finally.

JAGS (1:00:51) It's really burning at me. The Anthropic thing of being against open source and at the same time being religious techno-feudalists of some sort. It's the weirdest fucking prosperity cult I've ever seen in my life. It's just so strange. The fact that we're having those conversations while at the same time being told to root for the financial success of these three-trillion-dollar companies that are looking to IPO, so they're looking to take our money in 401(k)s.

Ryan Naraine (1:01:37) I love conspiracy Juanito. You're starting to sound like me. No, I dig it.

JAGS (1:01:47) It's not a conspiracy, bro. If their IPO is going to use the same mechanics as the SpaceX IPO, that was some shady shit. That was not how shit normally works. And that was a lot of people's pensions that they were trying to get into that. And the value did not hold for very long in the beginning. So if that's the mechanics under which you're looking to finance three-trillion-dollar IPOs, I would expect these companies to be way more of the people.

At least with OpenAI, and I know that it's a totally calculated business move, they're talking about, yeah, okay, fine, you want five percent ownership of the lab, no problem. I suspect when the government went and twisted their arm, they gave us gpt-oss-120b. They're willing to play ball in a way that I can respect.

Ryan Naraine (1:02:43) I can't root for either of them. I feel like you're rooting for one over the other, to be honest with you. If I'm keeping it a hundred, I can't get there.

JAGS (1:02:48) I am, I am. But frankly, if you follow the podcast, I've always rooted for one over the others, but it's changed. It was OpenAI, then it was Gemini, then it was Anthropic, then it was OpenAI, then it was Anthropic for a long time, and now it's OpenAI. So call me a fucking dilettante.

Ryan Naraine (1:03:07) This is true. I remember when Google was winning this race outright. There was no doubt that Google would win this race outright. So in fairness.

JAGS (1:03:16) Or a sunshine friend, or whatever the fuck you want. But I like winners. I like cool shit. Build me cool shit and I will fucking sing your praises all day long.

8. The Open Letter on Collective Cyber Defense (1:03:26)

Ryan Naraine (1:03:26) I can't get lectured by OpenAI on security. And I want to pivot back to this open letter. Is it conspiracy Ryan to question? They got hundreds of vendors to sign on to it, basically acknowledging we have a limited window to strengthen our cyber defenses. Let me set up the conversation and then you can weigh in all you want.

The idea is that we and these hundreds of companies signed this new pledge that we are going to follow these principles of collective response. We recognize that whatever's happening now is not enough. And then the second line jumps out at me, conspiracy Ryan. We must empower more defenders with cyber-capable AI. So now I'm creating the problem over here and I'm selling the offense over here, but you have to buy more electricity to be able to do it. That's one, and we need a collective response.

And here's what needs to happen next. Every organization must make cyber defense an immediate leadership priority. What the fuck does that mean? One. Two, haven't we seen that in press releases for the last fifteen years? Every vendor takes security, we take your security very seriously.

Costin Raiu (1:04:35) Is it like the secure computing memo?

JAGS (1:04:39) It's not a secure computing memo. It doesn't have the teeth a secure computing memo does. I think it's a platitude.

Ryan Naraine (1:04:49) Should we pay attention to this document at all? Why should me, Costin, or anyone pay attention to this document? And then the second question I have for you is, can you help me understand what does AI-powered defense mean? They say let's make it accessible and deployable for critical infrastructure. What does that even mean?

JAGS (1:05:06) It doesn't mean anything. What it means to OpenAI is something really specific. The fucking problem with San Francisco folk is they're so comfortable talking about making the world a better place and really, really big-concept shit. But really, what they mean is something really specific and small. All they mean, and it's a good thing, is that they want more people to have the full cyber unguardrailed version of the model that are going to use it for defense, which is a valid thing. It's why Hugging Face had to use Chinese models to do their defense. But that's really all OpenAI is pushing for in this instance. So thank you, because we need it.

And frankly, the whole Anthropic Mythos or Fable thing is pretty shitty. Opus 5 is useless for any of this stuff. So it's good. Thank you.

Ryan Naraine (1:06:15) Is it anything? Not a fucking pledge to sign. I'm sick of your fucking signed pledges.

JAGS (1:06:32) Nobody's going to go against, hey, we should all make this a priority. Yeah, we should all make this a priority. Now, when you ask what is AI defense, I think it's just thoughts and prayers, man. It's basically saying, I really hope one of you motherfuckers figures out what it means to use these things to do defense. The people who are telling us what defense is don't know what defense is. And the people who know what defense is don't have the resources or any of the machinery or the understanding or the capability to actually make these things work the way they should.

What nobody's saying here is, someone please fucking light a candle that some kid somewhere is going to come up with a clever way to use this shit we're building so that we can suddenly magic defense for $200 a month or whatever the fuck it is. And that way everyone's finally saved. The problem there, as we've discussed many, many times in this podcast, is that security is not a feature. It's not a thing that you're going to tack on on top of all this bullshit. We have massive systemic problems that are creating these horrible feedback loops with negative incentives that are going to continue to create instability and a level of variance that means we just cannot secure any of these things.

Does that mean that it's impossible to secure things? No, but you're not going to get it this way. You're not going to get it by having Microsoft give everybody a fucking operating system that's an ads platform that they don't feel like patching half of the time. You're not going to get it by having Android be a fragmented ecosystem where every different phone manufacturer gives you a different level of updates, or you need to be rich enough to buy an iPhone, and then you have no control whatsoever over what happens or visibility.

There are systemic problems that they're going to need to address. Like this open source thing I was talking about. That's the lowest hanging fruit, to get everyone on a safe kernel, a safe set of fucking libraries. The safest that humanity is able to produce at this time should be a standard.

Ryan Naraine (1:08:52) In the same breath that you're telling me that we have a very limited window to strengthen these cyber defenses, you're telling me the defenses that you're suggesting we should be using really don't exist. What the fuck is going on?

JAGS (1:09:02) Well, the problem that we're dealing with didn't exist six months ago. So clearly we're in a time of miracles, and we would like some miracles stacked on this side as well, if you guys don't mind. We're magicking a whole lot of shit over here.

Ryan Naraine (1:09:26) How does someone get to sign this letter? I scrolled down through alphabetical order looking for TLP Black to see if Costin signed this collective action on cyber defense. I don't see TLP. Were you invited to sign, Costin? Can you ask to sign? Are you in Operation Daybreak?

Costin Raiu (1:09:43) I don't know. I guess these are the people from Daybreak. It looks like it's the people from Daybreak. We actually applied with OpenAI to join that since they extended the group. I don't have an answer yet.

Ryan Naraine (1:10:02) What about the other one, the Mythos preview crew? Did you apply for that one?

Costin Raiu (1:10:06) I'm not entirely sure. For Glasswing, I don't think you can apply for Glasswing.

Ryan Naraine (1:10:11) And what about the Google one?

JAGS (1:10:13) Nobody's in the Google one. The Google one doesn't exist. The Google one's a fucking lie. Nobody's in that shit.

Costin Raiu (1:10:20) You're joking. It was me on X asking how do I get access, or how do I join it.

Ryan Naraine (1:10:22) I'm just asking the questions. Google announced something for defenders. Did you get into Dash? Microsoft Dash also has been issuing press releases about helping defenders.

JAGS (1:10:38) I don't think that's a club, though. They don't have jackets and shit. That's just a little project, a pet project. You need to get in a club.

---

9. Hardware Prices, Astra, and Where the Good News Went (1:10:45)

Ryan Naraine (1:10:45) Let me circle back quickly, because a thought crossed my mind while you guys were nerding out on Chinese hardware and going local. Don't let me forget my question. Go ahead, Costin. When I get high, I forget things.

Costin Raiu (1:11:15) On the topic of Chinese hardware, I saw a very interesting announcement from Xiaomi, another Chinese vendor, of a kind of NVIDIA DGX Spark killer device, a cube thingy that for now is a concept, which stands out because it has essentially ten times the memory bandwidth or something along those lines, rivaling the new Mac Studios in terms of memory bandwidth. And Xiaomi is saying that this is designed from scratch to run two models, a large model, something like 180 billion parameters, and a smaller model. So I thought it was an extremely interesting idea, to design hardware that runs two models at the same time. You're going to use the small model for the quick, fast things, and you're going to use the big model for the verification part.

JAGS (1:12:08) Think of it as speculative decoding. They want you to be able to speed up that way. It's really fucking smart. But Ryan, what was your question?

Ryan Naraine (1:12:17) I was just thinking, as he's talking, I wish Intel would circle back on that. You remember Intel used to have these NUCs, these little boxes that they would ship? I wish they would get back into the tiny hardware business. Hey, we've got money now. The stock price is climbing. The US government gave them a bailout, so there we go.

JAGS (1:12:31) Intel's trying to remember how to build chips, and as soon as they figure that out, they're going to. Thank you, NVIDIA.

Ryan Naraine (1:12:41) The question I had was something that I've been tracking privately, which is Costin's forty-thousand-dollar hardware cost. Everything refreshes this forty-thousand-dollar number. Do you sense that number is starting to go down? Do you expect it to go down? With the kind of advancements you're talking about with Chinese hardware and so on, what is a reasonable expectation around that forty thousand dollars that you would have to invest just to do a thing?

Costin Raiu (1:13:10) That number is going up, not down. Because of memory prices, hard drives, everything is going up like crazy. So what was 40,000 last time, now it's sixty thousand easily. It's definitely up.

I'll give you some random examples. A DGX Spark that I used to get for thirty-five hundred euros is now more like forty-five hundred or 5,000. Hard drives that I used to buy for 400 now sell for 480, even 500. Especially memory prices have exploded. RAM prices have exploded. And you see that pretty much with all the new hardware.

Some people are buying older hardware, because older hardware is still pretty good depending on what you are getting. DDR4 for your servers is still pretty good stuff. If you manage to get your hands on a server with one terabyte of DDR4, that is a killer server. We used to get one of those for 10k. Now I think that goes for more like twenty K at the moment. And you need those for databases, for running big projects, big data where you need big information processing. So the costs are only going up.

Ryan Naraine (1:14:40) So even with all the advancements, the costs are still climbing.

Costin Raiu (1:14:44) They're still climbing and I don't see it going down. All these memory makers, like SK Hynix, everyone is saying that production is locked for the next two, three years. Everyone is buying in advance every single chip that they can get their hands on. I don't see it going down. I think these little DGX Sparks will only keep getting more and more expensive.

And the new Mac Studios, I don't think they are a DGX killer, at least not in my opinion. Two DGX Sparks, from the point of view of what you are doing with them, the models that you're running, they're still more efficient than one of the new Mac Studios. There might be some narrow cases when a Mac Studio with 256 gigs beats two DGX Sparks, so that justifies the cost of another one. But the prices, I think they will continue to go up for the next year, year and a half, maybe two years.

JAGS (1:15:48) There's another dimension to consider, which is that for a change it's not bad to invest into hardware, in the sense that the value of that hardware isn't crashing quickly. Before, you would go and buy a computer and a graphics card, and within eighteen months that graphics card was the old, old, old generation. It barely works. It's not that good. You want to move to a 4090 or whatever.

Now, I bought a DGX Spark, what, a year ago? It's still as good a DGX Spark now as it was then. I bought an M4 Mac mini Pro the Christmas before that, just on impulse. Martin Wendiganson and I convinced ourselves to just get fucking M4 minis. Thank God I did. The price fucking went through the roof. You can barely get those RAM configurations anymore. I would have killed to still get my hands on an M3 Ultra with 512 gigabytes of RAM. That thing, which is now like three years old, is impossible to get, and it's like gold right now.

So there is this way in which, if you have the money and you have the interest, investing into hardware right now will not be a bad thing. It's going to keep getting more expensive. Now, the thing to keep in mind, before you get too hotheaded on my God, I somehow need to get this resource that's going to disappear forever, is that people are getting craftier. The homebrew community is getting better, and you're getting better performance out of the existing RAM. People are getting more creative in how you can get this quantized model, it works just about as well, and now it's three times faster overnight. Just because people are doing software hacks and shit that is really worth it.

So yes, the hardware will keep getting more expensive. But if you tell me that tomorrow everybody suddenly gets amazing performance out of the biggest models because we got really good at quantizing, I don't see people then continuing to buy DGX Sparks like they're the last fucking Coca-Cola in the desert. You're going to be satisfied with what you got to some extent.

Can we rant about that Chinese box you mentioned, Costin, for two seconds? Ryan, before, when you were baiting me into the whole supporting China thing, my big point was going to be that given the attitude that we're getting from the frontier labs in the US, particularly Anthropic, saying that they don't support open weights, the smartest thing the Chinese could do is to start dumping these Huawei-chipped AI boxes, DGX Spark killers, at half the price, everywhere, anywhere. Because people will beat a door down to have access to those boxes if they're cheaper, if they can afford them, even if they're not the latest, greatest thing.

Because if I have three of those in my house, I now have sovereign AI enough to manage the cameras and my chatbot and my internal text messages. And it's not like I need a hundred of those. I just need a couple. So the Chinese could essentially crater this fucking techno-feudalist economy right now if they just start pumping out those NUCs, man, with some Huawei chips.

Ryan Naraine (1:19:36) Whenever these things pop up, I always have this thought: if the Chinese were smart, they would be doing exactly this. And then I realized the Chinese are the smartest. Trust me, they are already fucking doing this. And they played a long game.

From episode one or two or three of this podcast, we were talking about, Costin had brought up, that the Chinese don't steal, they pay tribute to you by copying your stuff. And we talked about the Chinese thinking in generations and multiple generations while we think in election cycles. I can't help that conversation keeps coming back to me as we talk about, don't make me root for China with your bullshit. That was my thought. Two months ago I was like, guys, this is how you make people root for China. And here we are, you're talking about, well, they could be dumping. Not that they know anything about price dumping things anywhere. And not that they're not brilliant and play this long, long game. So we'll see how all of that plays out.

Costin, you got a closing thought on the hardware piece?

Costin Raiu (1:20:38) On the hardware piece, I think it's also interesting that NVIDIA is pushing a different kind of beast here, the DGX Station. If the DGX Spark is in the range of four K, the DGX Station is in the range of eighty K, eighty thousand US dollars.

Ryan Naraine (1:20:56) What's the difference? What is the Station?

Costin Raiu (1:21:00) The Station essentially gives you a lot more memory and a lot more memory bandwidth. Like seven hundred and sixty gigs of RAM that would allow you to run the bigger models. It's like a desktop supercomputer. That one is an extremely interesting device, because obviously it's not designed for data centers, like their highly optimized cards or pods. That is something they want companies to buy for employees. It would not be sitting in a data center.

JAGS (1:21:36) I want to buy one. I would buy that shit.

Costin Raiu (1:21:37) So if you know these people who are into cars, and the entry point is maybe a pre-owned BMW, and the next stage is a fine-tuned Porsche. It's an addiction. So in the DGX Spark addiction, the next stage is the DGX Station at eighty K.

Ryan Naraine (1:22:11) When does it get illegal to own it? That's also coming.

Someone mentioned Astra earlier. I think we heard from Sam two, three weeks ago that things had been paused for two weeks. Do we know when Astra is expected to come? What is the latest on the newest and greatest coming out of the labs? What do we know?

JAGS (1:22:34) We don't. We know that they slowed things down. And there have been discussions of limited access for Astra, but they're being very careful about it. We saw a public posting about the level of cyber risk that's being discussed with Astra. To be less hypey, to be a little deflationary, I think it's okay to say that there's a clear standard that we set with the Frontier Risk Council from early on, the beginning of the year, which was that the breaking point that OpenAI was mindful of was one where there was a combination of a visible alignment issue, that it would choose to do things in a way that increases cyber risk, and that the same model has the ability to autonomously exploit systems, to create net new, unique zero-day exploits for whatever bullshit it was dealing with.

And Astra is the first model that essentially has crossed both of those, with the Hugging Face situation and so on. That's what we're being told publicly up until this point. So they're treating the release in a slightly different way. And it's up to all of us to just wait and see. But I'm excited. 5.6 Sol is my daily driver right now. So I'm all for something bigger and better.

Ryan Naraine (1:24:16) When do these things start fixing diabetes? Why is it all the Astras and all, all I'm hearing about is cyber and how much you can hack things. I don't see a Sam tweet about—

JAGS (1:24:21) They are. They are, dude. Because we're a cyber podcast and we don't know jack shit about medicine, bro. I don't know anything. I haven't had water in fucking months. I'm not the right person for this.

Ryan Naraine (1:24:38) But you get the point I'm making. The frontier labs are still using cyber as the big messaging point, and people like me hoping that AI transforms the world in a much more positive way. Go help with the healthcare. Go help here. We're not hearing those stories, am I?

JAGS (1:24:49) You need to pay more attention. Sorry, let me put that differently. I don't want to put the blame on you. The problem is that we are listening to what's going into the mainstream news about AI. If we went a little out of our way, there's a lot of cool shit being discussed in medical circles and folks actually working in physics and these other things where they're like, holy shit, we just discovered this thing. Holy shit, we just solved this problem. Those stories are there. They're being drowned out because cyber right now is the greater vehicle for regulatory capture. It's getting them news cycles, it's getting them conversations with the administration. Everybody's talking about how dangerous these things are, which is an amazing way of crediting them as successes for what they've been building.

Ryan Naraine (1:25:49) Major health advances wouldn't do that?

JAGS (1:25:53) Yo, we don't even have health care, Ryan.

Ryan Naraine (1:25:55) I'm so jaded on the whole thing. You're spending all your marketing and storytelling cycles to talk about Operation Daybreak and this, and Nexus is held up because of cyber, and Astra is not coming because of cyber, and cyber, cyber, cyber. And you're telling me there's amazing medical success stories that are happening, they have access to it, and nobody's telling me these stories, and you blame me for not going and finding them.

JAGS (1:26:18) But Ryan, just because they discover the cure for cancer doesn't mean anyone's going to be able to afford it.

Ryan Naraine (1:26:24) Give me hope. Give me hope that your data centers are going to be used in some positive way.

JAGS (1:26:30) It's sci-fi shit. To most people, us being able to say, hey, you can live for 400 years now, it's like, please, no. Please, God, no. That's not the experience.

Now, I agree with you, and you should go. This is one reason that I did stan DeepMind very heavily before Sundar, I guess, decided to kill it. If you watch The Thinking Game, which was a lovely documentary, a lot of it was about the different projects that DeepMind was engaged in that were about protein folding and all these other medical advancements. They were doing the mission that I think, Ryan, you've wanted them to do.

And to be fair, Demis Hassabis moving up from CEO to, I guess, resting in heaven as a chair of some sort, is supposed to be also focusing more on his startup, which is about those things. If you want to stan any person in this drama as somebody who's focused on the thing you care about, it's Demis Hassabis. Period. Full stop. He's the only person that's done a thing where they could have tried to use this as a monopolistic advantage. The protein folding thing is one of the major discoveries of humanity. And they found it and they gave it away for free, fully, for the medical community. They did the thing you want them to do, and that's why he got a Nobel Prize. So that's your dude. Everyone else, I don't know what the fuck they're doing, but that's your guy.

Ryan Naraine (1:28:14) By the way, do you remember this company Midjourney, that was doing early AI images? They've pivoted to Midjourney Medical, a new division that's developing a full-body ultrasonic scanner to do MRIs in sixty seconds for a fraction of the cost. That's what I like. Get me excited about that shit. I'll stan all your AI frontier labs if you can get me excited about this stuff. This is more meaningful shit. I don't give a fuck about your zero days and your exploits. I really don't.

JAGS (1:28:47) I don't give a fuck about your zero days and your exploits either, to be honest. But shit got weird. That Midjourney thing is one of those moments where I just sat there and I was like,

I no longer possess the ability of discernment to know if this shit is an April Fool's joke, if they're trolling, if they really pivoted, if they're scamming us. Is this a rug pull?

Ryan Naraine (1:29:14) It is very much like a blood test thing, right?

JAGS (1:29:16) Right? And they have all these people speaking to it and scientists, and I'm like, thank God, I hope these people are telling the truth. What possible thing could I latch on to to assume whether this is true or not? I do not know, man.

Ryan Naraine (1:29:34) This might be a good point for a call to action. If any of the listeners, Spotify commenters, always got some smart thing to say. If you find some of this medical AI, advancements in medicine, healthcare and so on, send them to me. Please send them to me so we could amplify them on this show. I'm sick of talking about fucking Mythos and zero days.

JAGS (1:29:55) You know those people are on Twitter, right? Doctors are on Twitter. There are people that are working on medical research shit on Twitter, and those people are excited. They are excited, the same way that I'm excited about these things and reversing, man. It's helping all over the place, but it's not going to change the dynamics of what you're seeing around you for entirely different reasons.

---

10. ORB Networks and the Quartermaster Takedown (1:30:21)

Ryan Naraine (1:30:21) Where's the other non-cyber open letter? If it doesn't have a pledge, it doesn't exist.

Costin, let's pivot quickly to Lumen Technologies publishing on this Quartermaster China-nexus state enablement model, alongside an FBI and Justice Department charging document doing the attribution on the Chinese mercenary firm that was caught doing work for the Chinese government. There are two things here. One is the legal document and the legal case with the attribution, and one is a technical report from Lumen. Let's start with Lumen. Did you learn anything there? Did they give us anything to go hunting? Their IOCs? What is the state of the Lumen report?

Costin Raiu (1:31:10) There are a couple of reports with IOCs, which as you probably know is my middle name. I'm like Costin IOCs. So there's a bunch of IOCs in the Lumen report. There's also an interesting, I don't know if it's an Easter egg or something, but I'm one of the guys who checks everything with these reports, and at the end they say review these IOCs and visit our GitHub page that we update continuously. And they have a link to a Quartermaster IOCs page on GitHub, which is empty at the moment. I don't know if it's a mistake. It is empty.

Ryan Naraine (1:31:44) A text file that is empty.

Costin Raiu (1:31:54) So hopefully it will have more stuff in it soon. I think it was created about two months ago, and hopefully they'll fill it with more goodies.

It's a very interesting report, as we learn more and more what the state of the art of OPSEC is nowadays for APTs. And it leans in the direction of ORB networks. If you're one of these high-end APTs, you've got to either own or operate through one of these ORB networks.

JAGS (1:32:21) Get you an ORB.

Ryan Naraine (1:32:21) Back up for thirty seconds and tell the audience what an ORB is.

Costin Raiu (1:32:27) This is the intelligence community name for a VPN, which is very sophisticatedly called an operational relay box. So it's an ORB.

Now, these ORB networks typically include all sorts of exit points. They can have hacked routers, they can have hardware that has been sold pre-compromised, they can have commercial VPNs included in it. They can have all sorts of things. And the question is how to run something like that reliably over long periods of time and refresh them from time to time.

The first real mention of an ORB network is in one of the Snowden documents, where the Five Eyes, in particular Canada and the US, are discussing refreshing their ORB network. And their ORB network is built from routers running vulnerable versions, or intercepted credentials through their mass interception capabilities, or routers they know are vulnerable to specific exploits because they run very old firmware. In those documents they talk about doing a once-per-year event in which they refresh their operational relay box networks.

This has been adapted by pretty much all high-end APTs nowadays. I think it was an interesting moment when we saw Cloud Atlas using ORB networks back in the days. That was an interesting event.

And nowadays the Chinese are the champions of these ORB networks from many points of view, both in terms of size, aggressiveness, how they do it at scale, with multiple so-called quartermasters serving different APTs.

Ryan Naraine (1:34:33) Private sector companies, like mercenary-type companies.

JAGS (1:34:42) Some of them are private companies, not all.

Costin Raiu (1:34:45) Some are private. The whole Chinese business ecosystem, the guanxi, how they call these friendship-based relationships in China, is super, super interesting. And with quartermasters serving multiple APTs, that makes attribution very difficult.

Let's say you have this IP address that you see the APTs poking your infrastructure from, and you're saying, ha, so I have maybe APT3 poking me from this IP address. And then you see someone else poking from the same IP, but it's not APT3 anymore, it's a different group. You cannot use them reliably for attribution purposes. But sometimes the good news is that everything coming out of that IP is a customer of the quartermaster. The only issue is how to differentiate between them.

And this is what we have here, one of these beautiful ORB networks. Fast Labyrinth is one of the internal code names. By the way, I have to say that maintaining something like this is an incredibly painful process. For the people out there who have been doing web crawlers or malware crawlers, or having to use SOCKS proxies for OPSEC reasons to access cyber criminal forums and the likes, they know how difficult it is. Back at our old job we had a guy who was fully responsible for developing and maintaining the VPNs that all the systems essentially used. To detonate malware, for instance, you want to give the sandboxes internet access, then the access has to go through an anonymization infrastructure, otherwise they'll fingerprint you and they know it's you doing that sandbox execution. So it's very, very hard to run ORB networks reliably. And this is one of those where the shutdown of Fast Labyrinth will have a significant impact on some APTs.

Ryan Naraine (1:37:01) I was just going to ask, we talk about disrupting these ORB networks. Is it really effective? Or is it just creating small roadblocks that they jump over?

JAGS (1:37:13) To Costin's point, it's hard to run one of these anonymization networks when you have to follow the law. The real issue, the reason the Chinese are pole vaulting over everybody and doing this, is they don't give a fuck about the law. So they're popping residential proxies and routers and cameras. The ORB starts with certain kinds of devices, and then it goes into servers in one cloud, and then it goes into residential proxies in some country, and then it goes into another cloud, and then it pops out in your country. It's fucking impossible to deal with the jurisdictional nightmare. And that's what makes them so amazing.

Sorry, Ryan, you had asked a question and I moved us a bit backwards.

Ryan Naraine (1:38:04) Are these things effective at all? Are these takedowns just whack-a-mole? Like Costin says, it adds a significant bit of friction for the adversary, but is it really effective?

JAGS (1:38:17) It's important to do. We should do more of it. They would be more effective the more we do them. I think the big issue is that these things are done once in a blue moon and we celebrate. But really, unless you got all the way up the operational stack and disturbed this network further and further up, you're cutting tentacles and they're growing back after a bit.

When it's somebody's business to provide this service, yes, you inconvenience them and it's going to take time. It's not nothing. But unless you get all the way up, they're going to rebuild in some relative amount of time. However, the more this happens, the harder their lives get, especially when you're talking about these categorical takedowns.

One of the reasons that I've been trying to be spicy in some ways, but I really believe this, I've been advocating for a BrickerBot, for us to essentially brick vulnerable devices that are left astray on the internet. Because this is where you see what a massive amount of latent vulnerability does. You leave all these devices hanging around and you're like, well, it's not a big deal, it's just a bunch of bullshit cameras. And it's like, yeah, but they're a part of a fucking tunnel that is basically terrorizing half of the internet. And we're just leaving it as a thing that's there. So when we do a takedown and they lose a bunch of visibility, they lose a bunch of devices, there are forty thousand other cameras or fifty thousand other routers or thirty thousand other phones. This is why we need to get this shit off the internet and force vendors to have to produce something better, even if it means having to do controlled burns ourselves, because they're fueling bullshit like this.

Ryan Naraine (1:40:19) Shout out to our friends at Lumen Black Lotus Labs. Those guys continue to do amazing work. Shout out to those guys for the IOCs as well. Pushing amazing stuff all the time.

And this is a good time to call out, Mike Horka is coming to LABScon to do a talk about tracking Volt Typhoon's, quote unquote, Silent Train, which is an ORB network that's impossible to track because it's ephemeral and they're creating and discarding them. So he's got an amazing LABScon talk.

JAGS (1:40:45) No spoilers, man. No spoilers.

Ryan Naraine (1:40:49) It's in the public description. I'm just reading the abstract and saying that's one of the talks if you're lucky enough to be there.

JAGS (1:40:55) Now you're just fucking with people, because this shit's sold out.

Costin Raiu (1:41:02) When are we going to get the live stream from LABScon? Never.

Ryan Naraine (1:41:08) Never. Shout out to Mike Horka and all our friends over there doing amazing work. Keep shipping these papers. We love getting to read them.

The FBI attribution is helpful. How, Costin? What are you as a defender doing with the FBI attribution bit and the legal case? Is that interesting to you at all?

Costin Raiu (1:41:30) Barely, in the sense that, yeah, sure, it's interesting to know more about how all this operates. People going to different private events, conferences like LABScon, they

generally learn a lot more about the companies involved in the ORB networks and quartermastering and so on.

What is interesting to remember here is the method by which the takedown was executed. They took over the C2, the command and control domains that they were using for coordination between this infrastructure, between these operational relay boxes.

And what does it mean? It means that those individual boxes didn't necessarily get cleaned. They just took over the infrastructure. It would be very interesting if they were doing something extra. If these were compromised through vulnerabilities, they can just recompromise them with a different version that uses a different C2 infrastructure, and we are back to our usual broadcast by Monday.

Ryan Naraine (1:43:21) You're talking about pushing some code to fix something, or pushing some code to break something.

Costin Raiu (1:43:25) That is the BrickerBot. Maybe the cleaning bot, I don't know.

Ryan Naraine (1:43:30) There are multiple public examples of that being done, including by the FBI and Lumen doing this kind of botnet takedown.

JAGS (1:43:37) Dude, when a car breaks down on the highway and it's left there to rust, they drag it to the side of the road. They don't leave it there to fucking let other cars crash into it.

Ryan Naraine (1:43:49) Yeah, but we've got to get to a point where they have an alternative to getting something more secure and more safe. And they don't.

JAGS (1:43:55) Well, but that's the thing. It's the other way around, Ryan. You have to do this so that alternatives can thrive. Because right now, if you and I decide we're going to build an amazing home router company and our routers cost $200, no one is going to buy that shit, because the other ones sell for 12 bucks and no one can tell the difference. The day those things all shut down, and then you go buy another one and they all shut down, you're not going to go buy the same piece of shit again. You're going to look for something else.

Ryan Naraine (1:44:29) Now you're starting to talk like a capitalist again. I love it.

JAGS (1:44:32) The free market will solve everything.

Anyways, one spicy take, since we were talking about ORBs. And shout out to Michael Raggi for coining the term. It's kind of clever in some ways, because ORBs were a thing people discussed even in the days of Turla, satellite command infrastructure and stuff like that.

I have to get my car inspected every year before I can renew my registration to make sure it's not polluting the environment. That's right, cyber dad.

Ryan Naraine (1:45:17) So is that a call for all these ORB routers, these shitty routers, to check in somewhere and validate that they're clean before they're allowed to?

JAGS (1:45:26) The thing with the dumping is that they're dumping these devices on us, and the really terrible part is the software compatibility of it all. At the end of the day, if it's a piece of shit plastic hardware but it works, and I could flash a relatively secure piece of software on it, unless we're talking about embedded God knows what problems, which would be a different issue, at least the software end would all be managed in some way that we're all comfortable with.

Okay, spicy take, quick question, Costin. You know how we were told that after the Snowden leaks other countries were studying and poring over the documents and they learned shit from there and they raised their tradecraft in ways that they adapted and learned from the Snowden leaks? I've always been curious if there's anything visible from that, genuinely. You could say maybe certain malware techniques, but I also think that there's always been a lot of hyperbole around this.

However, this ORB shit is how we would see the Equation Group and other APTs consistently operating. Yes, there have always been some anonymization network techniques in general that were used, like Hacking Team and other folks. But the Chinese have gone full fucking bore on this. Tradecraft across the board became this network-sensitive thing, and also this memory traipsing and whatnot. But most importantly, the thing that's killing everybody is you now have 25 fucking ORB networks, and whenever they're attacking your country, they're coming out of a residential proxy in your country that you can't even monitor. You don't have the legal remit to monitor that shit.

So I don't think the Chinese just overnight decided, yo, this is a cool thing we should maybe do, guys. There are few examples of them learning from the best in a way that we could maybe quantify.

Ryan Naraine (1:47:42) They learn from the best.

JAGS (1:47:44) But I'm very curious to what extent you can see a coordinated shift, a decision that came in a centralized fashion where they were like, from this point forward we need everybody on some kind of, if not multiple, anonymization networks. And these MSS guys, hey, you're going to run your own ORBs. I don't even want you using the same shit as everybody else. And here's seven different companies that have been seeded around the country that are just going to maintain different ORB networks. Because that's the scale of effort that we've seen over the past four years, maybe five, that is now just plaguing us. You do wonder what gets an entire country's worth

of APTs to suddenly change their tune and the way that they do things fundamentally and consistently.

Costin Raiu (1:48:56) Here's how I think it works. I don't have any proof, so it's just my own speculation, my own feeling. But as we already know, the ORB people have their own little ORB conferences, the same way we go to LABScon, we go to PIVOTcon, State of Statecraft, Virus Bulletin and so on. They go to, I don't know, ORB Shanghai, ORBcon QiAnXin, ORB Guangdong.

And I would assume that everyone going to these conferences, just like we do, comes back with some notes from the talks. And at some of those conferences there are people going through leaks, there are people going through malware they find, there are people running honeypots and sharing their experiences with running those honeypots. Everyone is taking notes, and then at the end of the year you draw your own conclusions. What do we need to change? What are the best practices? What are the top teams out there doing?

So it's not necessarily a change that happens at the same moment everywhere, but over time certain techniques are better than others. Some stand the test of time, which would be the ORB networks. So you go to these conferences and you read about maybe Snowden docs, not necessarily those, but then you read about Cloud Atlas and there's people talking about other things, and you start wondering, hey, we need something like that. It would be really interesting to have a botnet that we could operate ourselves and anonymize connections in a way which is untraceable, even to the big guys out there, the NSA and their friends.

Tracing NetFlow across Chinese, Korean, Taiwanese networks, Asian networks, is extremely difficult. So at some point everyone starts to buzz about ORB networks the same way we are buzzing about ORB networks at the moment. And there are more and more moving into that direction.

I think the reason why everyone is so excited on both sides about them is because they simply work. They're effective, they are reasonably cheap, and they can be made stable operationally. And as a result, this protects you. It's a good attribution deflection. You see who is in the news: it's essentially these guys who are running the ORB network, not the APT groups who were actually behind them.

So I was very carefully looking through all these documents to see if they actually talk about who the customers of these guys are.

Ryan Naraine (1:51:48) They talk about the MSS, but they don't identify specific groups, right?

Costin Raiu (1:51:52) They don't say it was APT40 or it was this or that Typhoon. Some people know some of these things. They don't have names, I guess they are in the UNC category. Some of them may have names. They talk about i-Soon, which also had contracts with Unit 9086, but I don't

see any of the big names getting dropped in here, which is proof that the ORB networks work. Otherwise we would be seeing indictments for the customers of these ORB companies.

JAGS (1:52:35) That's part of the difficulty. You have six different groups coming out of an ORB network, and it's four APTs and then two criminal groups and then one God knows what the fuck it is. And it all sort of blends. That's why there's been the proliferation of UNCs, I'm guessing, in large part came from that. The TEMPs and the UNCs. There was a moment when the old mappings, yeah, you have them, but you're seeing stuff and it kind of looks like these guys, but the same points are also bringing you this other APT. Are they working together? It's just a fucking nightmare.

It's why, if you go back in this podcast for years now, I've been saying I respect the Chinese APT ecosystem, I respect the people who understand it, but I don't fucking understand it. I can't sit here the way that we look at Russian ops and draw this little map in our heads and say these guys are this and these guys are that. It is possible to track the Chinese APT ecosystem, but it is a type of scholarship in and of itself. It takes a full level of dedication. And there are like 12 people that I know who are all friends, and they're the people that track Chinese APTs for real, for real, that actually know this group does this and they're connected to that and they're part of this organization. Shout out to those people. It takes a lot of acumen.

Ryan Naraine (1:54:10) Shout out to Dakota as well. I think Dakota's been tracking some of this private sector use in offensive operations. He's got some previous work he had published. We talked about that previously in the podcast as well.

JAGS (1:54:19) Yeah. Eugenio Benincasa, Dakota Cary. There are a lot of folks that are very into this.

Ryan Naraine (1:54:29) A last thought on the ORB network stuff. Do the guys at Black Lotus Labs have some special type of visibility that gives them the access to write these things? And if ORB networks are becoming such a crucial part, why aren't we getting more reports from more places? Where else should we be expecting to get reports from? I know they have access to part of the backbone in some way. Lumen used to be the old, remind me, CenturyLink.

JAGS (1:55:00) The reason that Lumen has such amazing visibility into a lot of these things is they have an amazing research team, and they used to be CenturyLink, which means they own a lot of the wires connecting a lot of places around the world. The big wires.

Ryan Naraine (1:55:19) Plumbing. The real plumbing.

JAGS (1:55:29) A lot of other folks that sell you Netflix, they sell you things from these little hop points and places where you can grab certain bundles of information. But the internet doesn't route in a consistent way. And it doesn't route in the path of least resistance or the most obvious

path. So you'll see NetFlow from certain providers and it's very spotty. You see bits and pieces here, but you can never talk with certainty.

It's very different when you own the wires. It's very different when you're in an ISP. You can see other types of things, and you can do data science on those sets because there is a level of completion, at least to some degree, and predictability and volume and all these greater metrics that you have. So you've got an amazing team of folks with Danny Adamitis and Michael Horka and Ryan English and so on, that are really making hay of that visibility and doing good things with it.

When you ask about other people providing reports, there's great private reporting on this from folks like PwC, like Allison Wikoff and Chris McConkey's team. They've done great stuff on, was it Super Jumper, and others. The big issue here is you need to have a certain level of visibility to be able to do any of that. And then when you have it, you have to have a willingness to actually do the work and invest into it. And that's where you look at the cloud companies and you look at them a little sideways. Because you should expect to see more visibility in this, but some of the folks that can see that prefer to sell that under the table in their own way. And then some of these other folks are just not investing into doing that kind of security. Other things are percolating to the rest of the security industry the way that you would expect.

Ryan Naraine (1:57:21) What's the name of our guy at AWS? Our CISO guy at AWS.

JAGS (1:57:24) CJ Moses, the hardest working man in threat intel.

Ryan Naraine (1:57:29) CJ Moses. Come on, CJ, you've got some visibility to share with us. That's what I'm talking about, the threat intel teams at the big cloud providers. Cloudflare as well, I think, would have some visibility into some aspect of it. I'm just curious to say, beyond Black Lotus Labs, who should we be looking to to educate us on ORBs? Cisco, Palo, some of the switching and routing guys?

JAGS (1:57:49) You've got to look at what the clouds are that are supporting some of this infrastructure. The ISPs are in a position to see great things, and we never get reports from the ISPs themselves into where some of these things are going. There are smaller cloud companies that have a great deal of visibility into some of these things. And then some of the NetFlow providers. There's visibility all over the place. It's hard to put these things together in certain ways, but there's good work being done on it.

Sorry, one point. You mentioned Cloudflare. I actually don't think Cloudflare would have a great deal of visibility into this, because I don't think most of these ORB networks are trying to run through Cloudflare. It's very easy for Cloudflare, which is a much more engaged company, to just shut your shit down categorically. They're very responsive. Their threat intel team is on it. Not to besmirch, because I'm sure they'd love the visibility, but I don't look sideways at Cloudflare, because I would be surprised if a lot of this stuff is going through there now.

11. Backdoored ZBT Routers, Chinese Hardware, and the TeamPCP Arrests (1:59:05)

Ryan Naraine (1:59:05) Sounds good. Anything else to add, Costin? Last word on ORBs?

Costin Raiu (1:59:08) I think there's a connected story we have in our briefing that I find fascinating, which is these ZBT routers shipping with factory-embedded backdoors.

Ryan Naraine (1:59:28) This is a report from a company called VulnCheck, a startup here in the United States, about Chinese implants in the supply chain. Shout out to Jacob Baines, the publisher of the report. He referenced a previous publication called ENDLESSDOORS, and they wanted to talk about the ZBT supply chain. Tracking FCC filings, patent records, web pages and so on, he gets down to finding two implants on the device, one called SPEAKINGSTONE and the other one called DARKLANTERN, a backdoor that listens. No authentication required. Both written in Nim, both communicate over UDP, both are launched by the same binary, a connectivity watchdog called inetdetect.

Why are you connecting this to the ORB network scenery?

Costin Raiu (2:00:28) First of all, it's a fascinating story from so many points of view. On one hand, these are super cheap routers you can buy off Amazon, and anyone can have them in their homes. And these kinds of backdoors are the perfect placeholder for deploying ORB networks on top of these devices.

Ryan Naraine (2:00:53) Certified pre-owned.

Costin Raiu (2:00:58) Yes, certified pre-owned with the backdoor. And it's awesome to see both methods being deployed. One is this passive backdoor, the other one is the active backdoor.

To me what is fascinating is they were saying, for instance, in the case of the SPEAKINGSTONE backdoor, they observed 392 unique devices, infected devices if you want, of which 390 were in China.

Ryan Naraine (2:01:27) All in China.

Costin Raiu (2:01:28) Which is fascinating. And they also say it would be very interesting to understand who put the backdoor in there. Where does this backdoor originate from? I would assume that several theories can be on the table. If all the routers are in China, who has an interest to backdoor things in China? That can be one interesting avenue.

Ryan Naraine (2:01:54) Conspiracy Costin flying.

Costin Raiu (2:01:57) No, but it sounds interesting, right? For what reason would you put them in if they're being mostly sold in China? To be honest, I'm a bit disappointed the number is not in the thousands or tens of thousands. We're talking about hundreds of these devices.

Ryan Naraine (2:02:14) True, it looks very targeted. Almost too small a scale.

Costin Raiu (2:02:19) I'd be very curious what Lumen is seeing when it comes to UDP port 9992 that DARKLANTERN uses as a passive backdoor. It would be super interesting to see who is scanning for that port, who is scanning now. I would assume that all security companies are now poking the internet, maybe Silas' friends, I don't know, might be looking around to see which devices have this DARKLANTERN on them. And similarly, looking for the passive DNS traffic to the C2 built into SPEAKINGSTONE would be also very interesting.

And the question is, how many others have built-in backdoors, like smart TVs and so on? By the way, there's some amazing research being done by our friend Alex Turing at QiAnXin XLab. He constantly publishes on these kinds of things that I find fascinating. The hard-coded backdoors in smart TVs, and he mostly talks about the interesting implants that they find from Chinese cybercriminals.

I think we don't see a lot of stories nowadays, maybe just very few stories, about Chinese cybercriminals developing malware, deploying ORB networks and the likes. We saw what happened in Russia. Nowadays in Russia there's the cutting edge of cybercriminal and nation-state-sponsored cooperation, while in China it's maybe not something that obvious, it's more business-like. There are companies, business contracts, not criminals who are financially motivated, working by themselves, being arrested by the FSB and the likes.

So that's why I find it a fascinating story. My intuition tells me that there's something about the ZBT routers shipping with DARKLANTERN and SPEAKINGSTONE.

Ryan Naraine (2:04:31) The report does not establish any sort of Chinese government control here or some sort of official embed. And it also doesn't confirm malicious use against victims. It's just discovery of these two things. In the case of DARKLANTERN, they said they identified 203 internet-facing DARKLANTERN systems in twenty-two countries, self-reporting sixteen models. We don't know what exactly the models are, right?

Costin Raiu (2:04:58) They list the models actually. They're listed there on the site, all the specific ones, including some unknown vendors that are shipping probably compatible OEM hardware compatible with the ZBT. It's fully worth your attention, giving a look at this blog. They have YARA rules, they have a minimal scanner, internet scan results, Nuclei and Suricata rules, and hashes.

Ryan Naraine (2:05:30) Check out the VulnCheck blog. I'll post a link in the show description. Juanito, anything to add to this at all?

JAGS (2:05:37) This is a topic that I'm actually really interested in. Moreover, my plan tomorrow is to go to Akihabara and try to buy as many Chinese phones as I can get my hands on. Seriously. I was already working on the OPPO Find N5, which I'd gotten in Singapore, I think when Costin and I hung out there, maybe not last year but the year before, for Black Hat Asia.

At the time I found it really frustrating, because it was a beautiful piece of hardware, unbelievably nice folding phone, great screen, great cameras, but super locked down to this ColorOS 15, which is some Android variant, but very interestingly locked down. OPPO is a Huawei offshoot that they sell globally. There are a couple of those. There's another one called Honor. And they make these beautiful pieces of hardware and they sell them to the global market. But what's interesting is the software. They have these nice skinned versions of Android, but it also has a bunch of protected regions and things in the bootloader that you cannot inspect. You can't pull it out. And if you try to pull the firmware, an image from the internet, it doesn't contain those parts. Those are only in the hardware of the phone.

So I was really frustrated. I considered physically opening the device to try to short some pins to try to get in there. And the truth is that I'm just not the right person to do that. I was going to destroy this beautiful thing and not really get anything out of it.

But now that we have 5.6 Sol Cyber, I let that shit loose. I don't know how much I spent on tokens, but that motherfucker burned for, I want to say it was like 16 hours or 26 hours. Found and implemented a local exploit, fucking got everything out of this thing. So now I have a ton of stuff to analyze.

This is a project I'm really interested in carrying out with a lot more of these phones, where it's really desirable, lovely hardware. We even have that video of Nicolás Maduro, while he was still the leader of Venezuela before we yoinked him unceremoniously, showing off a Huawei phone that Xi Jinping had given him and how cool it is. And you're just sitting there in disbelief that a head of state would just take a fucking device and be like, how cool this fucking thing is. But they're marketing that globally.

Ryan Naraine (2:08:24) Wait till they start dumping the Huawei hardware on your AI systems.

JAGS (2:08:28) They're everywhere, Ryan. It's already everywhere except the US. When you land in London, the signs at Heathrow are for Honor phones, and they're these beautiful pieces of fucking hardware. That's why I pushed everyone to this Apple in China book so much last year. You see things that are just not available in the US market at all.

Right now they're moving on to the Huawei Mate XT 2. It's a triple-folding tablet. It's a phone, this nice little phone, and then it unfolds twice, two different hinges, and turns into a full-size tablet. This is the second generation of that. We have nothing like it in the US.

Ryan Naraine (2:09:20) Apple next week is announcing a folding phone, September 7th. You're going to be first in line to buy it.

JAGS (2:09:23) It will unfold one time. I will buy it. I will literally be in the first shipment.

Ryan Naraine (2:09:32) It doesn't have Face ID. We have to go back to using our fingerprints. Did you notice?

JAGS (2:09:36) I'm actually glad. Fuck Face ID. I'm fine with Touch ID. Touch ID was phenomenal. I'm glad we're going back to Touch ID. But yeah, it'll be a lesser product in some ways. Okay, just one fucking fold.

In any case, this is the high-end version of what Costin was discussing with these routers. You're telling me that the majority of the world market, excluding the US, is buying these. The wealthy people of the rest of the planet that don't get iPhones are buying these glorious-looking Chinese phones. And they have a fuck ton of stuff running on them that no one can inspect. I find that fascinating. I think it takes an amazing level of trust to say that that's acceptable for the rest of the world. And if they don't have the time to inspect it, they don't have the exploits to inspect it, I've got the tokens to at least be like, here, go on, do some cool shit.

Ryan Naraine (2:10:46) You talk about London and so on. Costin has been reminding us on this podcast, even with the EV space, the electric cars and so on, what's happening in Romania is so obvious compared to where we are. It feels like we're being left behind and a lot of us don't realize it. We talk about it makes me root for China, and not understanding that China is leaps and bounds ahead in so many respects.

JAGS (2:11:09) The only reason we're rooting for China at all is that American companies are acting like the villains in a way that it's stingy. It's not cool to want people to be dependent on you for everything in their lives. That's a parasitic thing to do. They can be rich, they can be wealthy beyond their dreams, without being predatory. That's the thing about this right now. You don't need the entire cake. You can have the majority of it and still be half decent to the people around you. I don't know why we're having these conversations in this way, but it does make people root for literally anybody else, including the Chinese.

Ryan Naraine (2:12:02) Listen, you go to Guyana, the Chinese are building all the roads, all the bridges, all the hospitals, all the schools. And when you have conversations with people there, you know what they say? Why is China the boogeyman? What did China do to me? Why should I take on your problems?

JAGS (2:12:15) Because Guyana only just got on that freemium train and they're not yet at the point where those loans get called in. I point you to Venezuela and Ecuador and some of these other places. I understand I'm not telling you that they should be thinking about America fondly or anything like that. All I'm saying is you guys are only just coming to this. You're in the honeymoon phase.

I want to point you to the places in South America that took basically bajillions in infrastructure build-out and investment from Chinese companies. And yeah, they got a bunch of cool shit built, but then you realize ten years later that all of your national resources have been tapped for 25 to 30 years in advance and given to China, and they're not going to forgive you for the debt. So it's a different strategy.

Ryan Naraine (2:13:18) I still need my bridges and my hospitals built by someone, because I can't afford to build it myself. That's the raw reality on the ground. Costin, last thoughts?

Costin Raiu (2:13:27) Two stories from my side on what you were saying with Honor. A friend of mine, I noticed one day she had one of these Honor phones, and I was asking, hey, is this any good? And she said, actually I needed a new phone, so I went to the shop and I asked for a Huawei phone, which is typically cheap. But the salesperson at the shop said, hey, if you go with the Huawei, no problem, but just so you know, it doesn't have the Google App Store, it doesn't have Google Play. So as a result you won't be able to have, I don't know, Revolut. In Romania, if you don't have Revolut, you're done. You don't exist, basically.

And he said, but there's actually an identical phone which is also from Huawei. A bunch of guys from Huawei left, with the approval of the management, and they started a new company called Honor, and they're making all these devices which are identical, a bit cheaper, and they have Play. So I recommend you buy the Honor device instead of the Huawei, which has the same hardware.

Ryan Naraine (2:14:37) Which is just a shell around the same Huawei thing you're running from.

Costin Raiu (2:14:50) I think she initially wanted to get an iPhone, and the guy in the shop said, look, I can sell you an iPhone, but basically this thing is probably just as good, or maybe a bit better hardware-wise, and three times cheaper.

Ryan Naraine (2:14:54) Give me a cost number so I can compare it to what an iPhone is.

Costin Raiu (2:14:59) Three times cheaper.

Ryan Naraine (2:15:04) So three hundred, four hundred bucks for five hundred?

Costin Raiu (2:15:06) Five hundred for something that matches the fifteen hundred iPhone, I guess. This is what she got. Of course it has pop-ups, it has advertisements that appear out of nowhere on your screen, which is to be expected. But she was happy.

And I saw another friend of mine also with the Honor phone, and I asked, hey, is this any good? He said, wow, the camera's amazing. I said, but what about privacy? He said, no privacy. So far no problems, but who knows?

JAGS (2:15:42) Not that we know of. We have not heard anything.

Ryan Naraine (2:15:43) Speaking of problems. Law enforcement caused problems for two of these Australian guys, 21 and 23 years old, alleged part of TeamPCP. Brian Krebs has a really detailed story on some of the OPSEC mistakes that were made and how they were able to track these guys and get to the arrests.

I bring this up because success stories are important, and we've been talking about, these TeamPCP guys were the guys behind the Shai-Hulud npm worm thing, right, Costin? Am I remembering it right?

Costin Raiu (2:16:19) They're the guys behind, in my opinion, one of the most destructive, damaging cybercriminal groups that we have or had at the moment. Right there together with Scattered Spider. And they're different from pretty much all the other cybercriminal groups in the sense that they're not the typical Russian cybercriminal ransomware crowd. These were very vicious hackers who leveraged the extent of Western technology with very little fear of getting caught, which in my opinion is dumb. It's kind of stupid.

And TeamPCP has been bragging about it. Unlike Russian cybercriminals, who are typically very silent, very closed communities, TeamPCP are bragging pretty much everywhere, like on X. I think at some point they even had verified accounts on different social media platforms, which is a bit crazy.

JAGS (2:17:28) It's quite possible they were on PCP, right? From reading some of the stuff Brian Krebs was writing.

Costin Raiu (2:17:32) They were. What is PCP, by the way? Educate us.

JAGS (2:17:35) It's a kind of drug, a street drug. I don't know that I have much more to give you, but it's a pretty hardcore drug. When you hear about somebody on PCP, you're like, this is, I'm just saying, because some of the reporting was about the drug use with these guys. And you're like, to what extent is this a nod to actual PCP use?

Costin Raiu (2:17:56) Right. So in my opinion, good riddance. But let's see if this was the whole of it, because as we probably know with these groups...

Ryan Naraine (2:18:07) Do you think this slows down those kinds of, because these were those self-propagating worms that we saw targeting developers, really hardcore shit they were doing, what, a month ago, two months ago?

Costin Raiu (2:18:17) Devastating. They've been going on and on and on, and nobody has any kind of clue as to how many credentials were exfiltrated and are just sitting in their databases. Because at some point, remember, they announced that they have too many credentials and they can't leverage everything by themselves. So they were partnering with other cybercriminal organizations to leverage all that stolen information. So it was one of the most devastating, damaging cybercriminal groups out there.

In my opinion this is a big success. Although I worry a bit that it's just two guys in Australia. What about the other guys?

Ryan Naraine (2:19:03) There's got to be another half a dozen at the very least, right?

I share your sentiment, good riddance. Hopefully the charges hold. We started this podcast, I remember one of the very first ones, it was a rant about cybersecurity vendors glorifying these Scattered Spider assholes and these kinds of mercenary groups who live among us. They live among us, they benefit from everything we do, and then they just turn their fucking sociopathic tendencies on us. So I have very little sympathy here.

Costin Raiu (2:19:40) Agree with you.

Ryan Naraine (2:19:43) All right, my friends, I think that's our show. We're running out the show. I know you guys can nerd out on Chinese hardware forever. Do we have a UFO segment?

JAGS (2:19:55) UFO segment, or cryptocurrency segment?

Ryan Naraine (2:19:58) Wait, I'm going to get to magic money in a second. Anything new on the UFO front? You guys are slowing down. Terrible.

Costin Raiu (2:20:06) I don't think I've heard anything new, but I may be wrong. Maybe there's stuff happening and I just don't know about it.

JAGS (2:20:14) Every time we talk, you tell us that next week they're going to release something. Next week, guys, next week we will get proof.

Costin Raiu (2:20:20) Okay, so look, what everyone is saying, these are facts, I'm not speculating, these are facts. Trump has the speech written. The speech is on his desk. And then later there's no speech, there's no desk. But the rumor is Trump has the speech written, it's already ready, they're just waiting for approval to do all this thingy. But at the moment it's bad timing. What about the Strait of America? What about Lake America?

JAGS (2:21:05) We don't have a strait yet. We have a lake and we have a gulf, okay? He said we're doing an ocean soon. So watch it, Pacific, I'm looking at you.

Ryan Naraine (2:21:14) What's the Polymarket on whether it's the Atlantic or the Pacific? There's got to be a betting market going.

Costin Raiu (2:21:23) I guess it'll be the Atlantic. And then Greenland will be called Green America or the likes.

Ryan Naraine (2:21:33) In all seriousness, magic money took a nice jump for you guys that own Bitcoin over the last week. Any new news on the magic money front?

Costin Raiu (2:21:43) A lot of news. To be honest, I don't keep up with that too much. Why is it spiking? It's because Trump made the big speech with all the CEOs of the magic money companies behind him, explaining how bad Sleepy Joe was to this and we're just going to get rid of all that. Now it's the golden age of magic money, the magic golden money, and everything will be great and America will be a leader. We can't allow our adversaries and China to beat us in magic money, so our magic money will be better than their magic monies. And the speech was good. I saw the speech, it was a very good speech. That's a reality.

JAGS (2:22:30) Who the fuck has GPUs right now and is dedicating them to magic money bullshit instead of selling tokens? Are you fucking kidding me? Why?

Costin Raiu (2:22:36) A lot of people. There's an opportunity everywhere. I think opportunity is everywhere. Some people believe in one thing, some people believe in God, some people believe in Allah, and some people believe in Buddha. Some people believe in magic money and Michael Saylor. Michael Saylor is one of those guys who has certain followers. And that speech is actually the reason behind the recent increase of magic money, and pretty much everything is growing.

JAGS (2:23:18) Why would he do the magic money speech and not the UFO speech, bro? Come on, man.

Costin Raiu (2:23:25) Exactly. Very disappointing.

Ryan Naraine (2:23:27) It is 2:30 AM in Tokyo, Juanito. Close the show with some quick shout-outs.

JAGS (2:23:35) Shout out to Alex Perez-Palma, who's been a really good friend. Alex, formerly Workday, now Apple, just an excellent researcher and very good friend, who was among the many people who've just been shouting me down to actually take a trip on this vacation, which has been great. I've done nothing people have told me to do, but I am having a great time. Hopefully more excuses to come back.

I brought an extra empty suitcase just to shove full of shit here. I'm buying every toy I wanted as a kid and couldn't get. Retro video games. I got GoldenEye, Zelda, Ocarina of Time, Majora's Mask, all in the original Japanese, N64 shit. I'm just going wild on stupid shit.

Ryan Naraine (2:24:07) Did you find rice flour pancake mix? My hero. Shout out to Muji.

JAGS (2:24:33) In any case, shout out to our friends. And also, the Japanese conference scene seems to be getting interesting. I'm not here for any of that, but you've always had CODE BLUE, and now there's an OffensiveCon Tokyo, which sounds pretty interesting to me. Hopefully more reasons to come back to Japan, man.

Ryan Naraine (2:24:59) We should do an in-person episode, the three of us from Tokyo, before the end of this year. Let's plan it properly. Hurry back safely. We've got a LABScon to finish planning and delivering.

This is my quick shout-out. We mentioned the TeamPCP guys. Austin Larsen from Google is going to do a talk there on the behind-the-scenes things that those guys did at Google to track it. Not a pimping of the agenda, I don't know why, but I'm so excited about the LABScon agenda this year. You look at it from top to bottom, day one all through the second day, it's just some fascinating, amazing behind-the-scenes stories, never-before-seen research type things. So I'm excited about that. Costin, what do you have for us?

Costin Raiu (2:25:42) Shout-outs to our good friends in Japan. Shout out to our friend Suguru Ishimaru-san. Hello, and to Sato-sama. Sato-sama is an amazing person.

At some point there was an unofficial, I don't know how to call it, competition between Sato-sama and myself, at who was sinkholing more domains. Sato-sama was running his own sinkhole, I was running my sinkhole, and we were sinkholing the shit out of C2s. And I didn't know who he was, and he didn't know who I was, but everyone could see that sometimes a domain for some APT expired, and it was one or the other grabbing the domain. And they're like, who is that? Who is that?

I remember when I met him one day: that was you. Yeah, that was me. At some point I stopped doing the sinkholing, but he is absolutely the master of all sinkholing. So shout-outs to them.

Ryan Naraine (2:26:48) Shout out to our guy Misha-san as well, who was the spam tracker master.

Costin Raiu (2:26:50) Of course, shout-outs to our good friend Misha-san, who still rocks. Music composer, German guy living in Japan for his entire life, is now I think more Japanese than the Japanese, speaking German with a Japanese accent, and he was absolutely awesome.

Ryan Naraine (2:27:16) All right, guys, that's our show for this week. We'll catch you next week.

JAGS (2:27:21) Take care, guys.

Costin Raiu (2:27:22) Bye-bye.
