



SuperDataScience

SDS PODCAST

EPISODE 1034:

**SDS 1024: IN CASE YOU
MISSED IT IN
SEPTEMBER 2026**



Jon Krohn:	00:00	This is episode number 1034, our In Case You Missed It in September episode. Welcome back to the Super Data Science Podcast. I'm your host, Jon Krohn. This is an In Case You Missed It episode that highlights the best parts of conversations we had on the show over the past month. My first clip is from episode number 1023 where I speak with the wildly popular data scientist and entrepreneur, Ishwaria Srinivasan. Ash has over a million followers who clamor to read her posts about shipping agentic AI in production. With the cost of writing code collapsing for all of us, the bottleneck in shipping has moved from execution to judgment. So I ask Ash, which failures a person ought to live through themselves before anyone should trust them to deploy an agent? Let's get into the nitty-gritty of some of the stuff that you teach on agentic AI at the Gen Academy.
	00:57	I think that that's probably one of the most useful ways for our audience to spend their time learning about the most cutting edge things. So you argue that AI shifts the bottleneck from execution to judgment. So if the first generation of AI education taught people how to make models execute, agentic AI may require teaching them how to judge when systems are not working. So what does a curriculum centered on failure literacy look like? And which kinds of failures should our listeners be most looking out for, should they maybe even encounter deliberately before they're trusted to deploy an agent?
Aishwarya S.:	01:38	Very much. And this is something that I've also shared in the past in one of my previous sessions at open source conference. And I was talking about this, that the cost of building code has become so cheap and that is one of the reasons people misinterpret the fact that software engineering is going to become obsolete or people are not going to need software engineers anymore. I think the entire role is shifting. The amount of time that you used to spend on writing import statements for your code or fixing the intendation of your code is not the same. So the



time that you invest in doing different parts of your job for a software engineer is not going to look the same. It's obvious. It's dead obvious that nobody's writing code by hand anymore. Everybody is doing AKA wipe coding. I don't know how I feel about that terminology.

02:32 It's good and bad at the same time because I feel it's great because it lowers down the floor of what you can do and going from an ideation to building something. But at the same time, I feel if it's not interpreted correctly, it gives away a feeling that building code and running code in production is as easy as thinking about an application, which it's not. So I feel having an understanding of where you cannot let go of not just engineering fundamentals, but also knowing what exactly are you even trying to build, the optimization part of it, the architecture part of it, the decision making part of it. If you don't know what an IM is, you'll not know what an IAM is. So that's why having that literacy of that software engineering fundamentals is very crucial if you really call yourself an AI builder. If you want to really not be using Codex to build these cute little applications run within your Codex browser, and then as soon as you deploy it on Vercel app and you give it out to the first thousand customers, it's going to start breaking.

03:46 So I think that is a huge difference. So as it reduces the floor, as it reduces the barrier to entry, for more people to produce code, it reduces the barrier to entry for the cost of producing code, people are also generating spaghetti code. It is a shit ton of spaghetti going all the way and that's the AI slop even for code that's happening everywhere. And people don't know how to really make use of it. So what used to take you a few weeks to build something which was more mindful is now taking you 10 minutes to build, but is mindless. So it's only accurate to say that while the execution has become so much cheaper, if you're not intentful about what you want to build, then you definitely don't have a mode because what



you can build in 10 minutes, somebody else can build in 10 minutes too.

04:45 So if you're not intentful about what you're building, why you're building, who's it for, how is it going to be used, how is it going to improve over time, how is it going to compete in the market, that's still classic business, that's still classic product, that's still classic engineering. So that's something which definitely hasn't changed.

Jon Krohn: 05:02 Yeah. And I think anyone who's interested in getting involved in a business or in a product, they're going to want to know that somebody has been mindful that a lot of thought has gone into this, that there's a moat, that it's going to be secure, that it's going to be compliant, that it's going to scale. Like you said, if you just all of a sudden have a thousand users and the system's not set up for that, then you're going to run into a lot of trouble. And yeah, I think it is so easy today to create not just software, but just about anything. I think I actually recently told this story online recently, but I think it aligns with what you say. A friend of mine sent a pitch deck for a new business that he's creating and it was obvious that the whole deck was AI generated.

05:42 And I didn't want to read it. I wrote back, I was like, I feel like I'm wasting my time when I'm sent a document where I don't know if you spent more than two minutes on this. So I would prefer to have a terrible looking Google slide that's just like 10 slides with a white background, but some diagram, a few diagrams and a few bullets that explain what your business idea is. And I know that you thought through it. That would be better to me than this 20 page, amazing formatted, all this detail. I don't know.

Aishwarya S.: 06:17 I mean, it's a trend. I feel the unfiltered, raw things are more authentic now and authenticity is being credited for and authenticity is what people are resonating with because they've had enough of the slop. There are enough



beautiful looking presentations which is meaningless in text. So I think people are recognizing that.

- Jon Krohn: 06:39 Yeah. Anyway, you gave a great answer, but there was to the question that I asked you now a few minutes ago, but there was one part of it that I'm not sure if we got to it. So I asked a very long question, but the end of it was which failures should people encounter deliberately before they're trusted to deploy an agent specifically? So not just general apps, but when you're thinking about training people up to create and deploy agents, what are the kinds of experiences that they need to have that they need to see go wrong in order to be trusted in real production?
- Aishwarya S.: 07:18 So I would say that's one of the areas where AI evaluations are such an important topic and that's also a huge, huge area. Because of the non-determinism nature of the models that compounds with the tool use that you give it access to, that compounds with more loops that you're creating with it, more drafts that you're creating with it, the agent harness that you're creating along with it, that has a huge compounding effect of uncertainty in that entire system. That's the reason understanding how these models work in different scenarios is what is AI evaluations, which is you're not just evaluating the model's output. It's not that you're giving a model input through a LLM API and you get a response and you're seeing how the text response looks like. You're actually deploying and evaluating it end to end, which is right from when a user puts in an input all the way to all the tools being called, all the failure modes being addressed.
- 08:17 If there is an API fail, if there is a tool login issue, all of that being considered till the end of it, when you're actually getting the final response, all of those breaking points need to be evaluated. And that is a crucial aspect and it is not something which is one size fits all. And that's what's very important for people to understand that a lot of times people are looking for prescriptive



approaches. While these prescriptive approaches or the frameworks can help you give a certain direction, which is just a generic direction, or even these metrics can give you a general sense of how your agent is performing, every single attachment that you do to your model, every single of those joints can introduce incredible amounts of challenges, incredible amounts of failure modes. And that is only something that you will be able to assess if you are mindful of making those joints and making those connections, knowing how many tool access does it need to have access to?

09:17 How are you going to manage the role-based access to different users? What kind of database does it need access to? When should it be able to go right back into a database or not? So having a deep understanding on what are you really giving the control for a specific agent is very important. And that's one part of AI evaluations. Surrounding that is the traditional software engineering thing that I'm coming back to, which is not something which is just new to AI agents. It is something that has existed even with traditional softwares. At the end of the day, if you have evaluated your AI agents inside that box where you have all possible different combinations of things that can go wrong, at the end of the day, that becomes a software. That is the new software. Any of the new softwares which are being built in 2026 are not non-AI powered.

10:07 Everything has that non-determinism introduced to it. Now that's the new software, that's the definition of a software now. So when you think about that as a software and you're deploying it, you come up with the traditional AI engineering constraints of what happens when you have a DDoS attack? What happens if an unauthorized user is trying to log in? What happens during a prompt injection? What happens if a certain user who's unauthorized is able to get access to a database and so on and so forth. So that comes back to traditional



security and safety, which is not entirely new to AI agents, but knowing that is also equally important as much as knowing AI or agent AI

- Jon Krohn: 10:51 Stuff. Ash keeps coming back to fundamentals, knowing what is actually happening underneath the thing you are building. My next guest took that instinct about as far as it will go. In episode number 1025, Dr. Luis Serrano, founder of Serrano Academy and author of the bestselling book, Grokking Machine Learning, tells me that none of the standard explanations of attention ever clicked for him, not the query and key search table, not the formula. So he went off and built his own picture of what's happening inside a transformer. Then a physicist sitting in the front row of one of his talks came up afterwards and told him it was something far bigger. Something else that you published is actually a new paper. So you published last November with, I'm going to try not to butcher their names, Ricardo DiCipio and Jairo Diaz Rodriguez.
- Luis S.: 11:41 Yes. Very good. Yeah.
- Jon Krohn: 11:45 I really put a lot of effort and thought into that. You guys wrote a paper together about the curved space time of transformer architectures.
- Luis S.: 11:55 Yes.
- Jon Krohn: 11:56 And that is pretty mind blowing. I think we're going to spend a bunch of time on that right now because we'll learn about transformer architectures in a way, but I think we're also going to learn about space time and relativity and these kinds of concepts.
- Luis S.: 12:10 Yes. This was definitely very exciting, definitely very exciting to work on. And yeah, definitely for the physicists listening, we use the word relativity, but in a very loose way. It's basically a spacetime curvature analogy, weak



analogy of what's happening inside a transformer. But I think it opens the door to what's happening underneath and the fact that physics related things start appearing, I found it mind blowing. The story of that is that in order to understand attention, it never clicked to me with the... People say it's like a search table with a query and a key never made sense to me. Then they said, "Oh, the words pay attention to other words never made sense to me." Then they gave me the formula, made even less sense. Soft max of KQ divided by squared root of DK times V . So I started looking at videos and looking at other things and looking at just writing and watching videos.

13:18 Somewhere in the process, somebody, which I bless his soul, I don't remember which channel was this. I think it was kind of an underrated channel that didn't have very described, but this person just kind of made a... This is a beautiful description where at some point they did a linear combination of the words and they said this word becomes more like that one. And the linear combination added to one, the coefficients added to one because there's a soft max. So you turn your word apple into 70% of apple and 30% of orange, if you said the two words consecutively, say orange, apple, you know what I mean? So the word becomes a percentage of itself and the rest of percentage of another word. And to me, that's moving in a line. If I have two points and then I take a percentage of the position of one point and the other, I'm moving in the line between them.

14:12 And so I thought maybe words are moving in a line. And I started rewriting all the equations as in words are in a position in space because embedding words are in a position in space and then attention just moves them in a line toward each other. And I immediately thought, oh my God, that's gravity or magnetism, right? Words pull each other. And I thought that makes a lot of sense because if I'm saying the quintessential example is the river bank, bank is a bank in the financial sector of the embedding



around stocks and bonds. And then you say river bank and the bank just becomes a nature thing. So the word river just pulled it towards itself into the nature region of the embedding because in the nature region of the embedding lives a river and tree and stream and sea and all that stuff.

15:03 And so it just pulled it. It infused itself with it. It infused some properties of it. For example, the nature property. It didn't infuse all the properties. There are some that don't, but the nature property, it moved in. So the moving actually works in different directions. It's not towards it because of the value matrix. But anyway, the fact is words, I started calling it word gravity. And as I said, any physics words that I say is a very loose analogy, but I started calling it word gravity, word gravity, word gravity. And then one day I gave a talk in the Toronto Machine Learning Summit and Ricardo de Cipro, physicist, I also, by the way, I hope I'm pronouncing it well because I don't know Italian. And Ricardo's a physicist who was sitting in the first row and then he came to me and said, "Hey, what you have is actually when Newton was talking about gravitation and the Einstein came, which is a force between objects.

16:01 He died and he had no idea why this force happened. And then Einstein came hundreds of years later and he said, "It's not a force. The masses are bending the space. The reason you fall towards the earth or the earth falls towards the sun is not because the sun exerts a force, it's because the sun bends space. And all of a sudden the line in which the earth should be flying in a straight line, it's curved because of the sun and it happens to be curved around the sun and that's why we're there. "So he said," I think it's the same concept. You're talking about word gravity as a force between words. I think it's more like it's probably a space-time curvature thing. Probably words are bending space in a way that they just pull towards



each other. And so we started working on that and he worked out the math a lot.

16:47 He knows the physics a lot more than me. So he actually worked out the geodesics and very much like the matrices that appear in the geodesics appear in are the key query and value matrices. And then we started working with another friend, hire the professor at York University in data science and statistics to run a lot of experiments. So these two guys are wonderful. They actually know a lot more about that than me and worked out both the physics and a bunch of experiments that really study this analogy. And so we're very excited actually of this. I mean, it provides an analogy. I think it's more of a visual work. We meant it as a visual work. We've got notices of labs that are working with it for something else. So I'd love to see applications of it. But as of right now, we thought of it as a fun analogy, like physics analogy of what's happening inside ChatGPT or inside the brain of these models.

Jon Krohn: 17:45 That's the view from inside the model. My next clip is about what it feels like to work alongside these systems all day. In episode number 1029, Dr. Katie Malone, host of the very popular Linear Digressions Podcast, a show she stopped making for five years and is only now brought back with the help of AI automations. In this clip, Katie tells me which part of her career turned out to be the best possible preparation for working with AI agents. Her answer is reassuring for some of us and rather less so for the engineers who are the strongest individual contributors. An interesting part of your podcast journey is that you started in 2015, so 11 years ago, and you ran it for five and a half years, almost 300 episodes. And then the pandemic summer, July 2020, you stopped and you stopped for five and a half years.

18:39 You did it for five and a half, stopped for five and a half. And the reason you gave publicly at the time, it was that



there's no particular reason, you just couldn't do it forever. And we have some quotes from you at that time that if you felt that the field was moving away from you, the show had started when people didn't even know what the realm of the possible was in data science. And by 2020, data science had become as much about management, responsibility and scales about the algorithms. And you said the content kind of stopped pulling at you. Then nearly six years later, you named two causes you hadn't said before, a pandemic burnout, a grind of production. And you kind of alluded to that now here today where that grind has been alleviated so much by tools like, you just said it before we started recording, what's the name of the tool for -

Katie Malone:	19:29	Descript.
Jon Krohn:	19:30	Descript, exactly. Yeah. I can't believe I didn't have that right in my brain. But yeah, an amazing tool for allowing people to edit episodes very quickly. Yeah, I don't know. I find that journey interesting and I wonder how many people do that, but we're so delighted to have you back on air.
Katie Malone:	19:48	Well, thank you. I'm delighted to be here. And I think this is interesting for me because I haven't gone in and excavated what I was saying or thinking in 2020. That resonates, that tracks. That sounds like something I would've said. And something that I've been thinking about a lot lately, and I'm really interested to hear the seeds of it and what I'm saying. So something I've been thinking about a lot lately, I mentioned in 2020 that data science was becoming, for me anyway, partly because of just where I was professionally, a lot more about management than necessarily hands on keyboard. And so struggling a little bit with coming up with new content that was faithful to what I thought my audience came to us for when my day-to-day job was managing people. I'm not writing algorithms anymore. I'm going to meetings.



And in the time since then, I've stayed in data science management broadly.

20:46 But I think that with the advent of AI, there's a very interesting synthesis maybe between person management as a soft skill that you might learn because you have to do it for your job and the technical management skills that you need to be an effective user of, especially agentic AI. So the idea that my job now is context switching between different work streams that are each being carried out independently. It's about defining the task to be done and the acceptance criteria for when it's going to be complete, that there's a fuzziness or there's a lot of different ways that what I say can be misinterpreted or done incompletely or not in the way that I intended. And so I have to be checking for that and kind of a trust but verify type model. Those are all management concepts that transfer very, very elegantly to being an effective user of contemporary AI tools.

21:48 This is just an idea that I've been developing a lot because I think a lot of people are maybe non-technical, but they have been managing people or projects or whatever for a while. That set of skills might be one that they have very developed. They're potentially being confronted with the possibility of needing to manage this new type of entity, like an AI agent for the first time and maybe feeling a bit out of their depth. And I would say to them, you might actually already know more than you realize. And I think to some of the very technical people, especially software engineers who've been effective ICs in the past and are now struggling as being agent managers and they're saying, I hate my job now. I don't like reviewing other people's content. I can't get into flow. 100% true. And I don't have answers to all of those problems or all of those questions as a manager.

22:39 I struggle with flow. I struggle with context switching. I struggle to articulate what I want sometimes. But there's



a lot of other people that have figured out ways to deal with that and maybe there's some cross pollination in the other direction. So anyway, maybe more than you were thinking when you asked the question, but I'm interested now that we see this, me back in 2020 saying, well, I don't know if AI is really what I do anymore because I kind of do all this management and I'm like, oh, those are the same things just in different clothing.

- Jon Krohn: 23:11 That's a really interesting answer and I'm so glad that you got into the Agentic stuff right away because I was starting to think after I had posed this question, I was like, have we been going on about podcasting too much? Is this just my interest? Is the audience going to be as interested in this as I am? And I don't know, lots of advice on hosting a podcast is that you should be getting into whatever interests you, but I was still like, maybe we should be getting into the technical aspect of this. And then you did anyway. So perfect. Yes, this new world that we're in where we're doing agentic management, it's only been the past year that this is something that people are doing. You have worked at a business with tens of thousands of people where you built the agentic AI platform, and this includes deployment, enterprise adoption, responsible AI governance, and there's a lot there to get right.
- 24:04 I don't know to what extent you can tell us about what it's like building, being the person responsible for managing a team of humans and agents to build a agentic AI platform.
- Katie Malone: 24:20 Yeah, that's an interesting question. I mean, it's hard and I think it's one of the things that's very challenging right now, and I think this resonates maybe with everyone to some extent, is as much as you can build something that's compelling and maybe a little bit future-proof and sets us up for some long-term growth and value and whatever, when the goalposts are just moving as quickly



as they are, it's really difficult. And big companies I think have it extra hard because they're kind of like aircraft carriers. They're just hard to turn. Once they're going in a certain direction, they can go very, very far, but they tend to not... It's just not as nimble to get 10,000 people going in a particular direction. I think I do kind of wonder, some of this is just reflecting where we are as a society right now. This might look very different in five years as people have had a chance to acclimate a little bit to some of the AI tools.

25:18 People are maybe a little more fluent with it. Some of the norms that I think we're figuring out now might have settled in a little bit. It might be you feel like it's not okay to send AI slop to your coworkers or something right now.

Jon Krohn: 25:32 I hope it doesn't. Please stop. If you're that one guy, just

Katie Malone: 25:36 Stop. Well, it's not one guy though. Yeah, that's the thing. It's like my AI slop is talking to your AI slop.

25:43 I talked to Tom Davenport a few weeks ago for my podcast. He's great. He's wonderful. And for anyone who doesn't know Tom, he's been writing, especially enterprise data science and analytics for decades. And he has coined the term process slop. And I think it's an idea whose time is rapidly approaching of... I'm a job applicant and you are the hiring manager on the other side, and it's just like our AI slop going back and forth. I have AI generate my resume, you have your AI that reads it, that automatically sends me some kind of reply, whatever. Anyway, so I think that those are challenges for us in general and in particular in large organizations where you might not have direct personal relationships with the folks that you work with, you're kind of relying on the machinery of the organization and some of the processes to get things to where they need to go.



26:41 Then injecting AI into that all of a sudden is not necessarily fitting in exactly with how these things are working. And so there's also, I think, a really important part of what you might call change management or something. Just how do you turn that aircraft carrier? And I don't know. I don't know how much I have to say here that's deeply insightful or specific and insightful besides it's just really hard work. And I think it's interesting to see in some ways as there's new companies that are popping up. They're obviously approaching how to build businesses in sometimes very fundamentally different ways. Established companies are retrofitting their operations and their technologies to varying degrees of success or maturity at this point. So it's an interesting, I guess, period of high flux for us all to be in.

Jon Krohn: 27:39 Managing agents is one problem. Paying for them is quite another. In episode number 1031, Dell Technologies distinguished engineers, Ish Shah and Tyler Cox explain why agentic AI burns through tokens at a rate that is catching companies out. Ish gets there by way of the anniversary present. He built for his wife a Pokemon style video game starring their dogs, which consumed billions of tokens over a single weekend. Why does token consumption go up so much with agentic AI? Ish, how did you burn through billions of tokens on the weekend on a side project? And can you tell us what it is?

Ish Shah: 28:16 I can. You may have to bleep out a word if I commit some sort of IP issue. Okay. So it's my one year anniversary this Sunday, and as my wedding gift to my wife last year, what I did was I took... Everybody played Pokemon as a kid. Pokemon's making a comeback. It's cool again. Everything old is new again. I basically built a fan game in the art of Pokemon where the map is my area that we live in Atlanta, where my wife and I met, where we got engaged, and I have these little pixel art maps and I had her caricature done as pixel art and I replaced the Pokemon with my dogs. That's the project. Every year,



every major life event that we have, I build a chapter into the game and that's my get out of jail free card on the present part of things.

29:14 And so what I've been working on is these models and their capabilities over the last couple months have shot through the roof. The artwork has gotten considerably better. The game mechanics and how much I need to supervise my little buddies as they go off and work, I can go have a cup of coffee and when I come back, the chapter is built. The reason the burn was so high is because what these agents are doing in order to achieve the task, just like humans, they're divvying up the work and they're spawning subagents. So now you've got an agent in charge of a bunch of other agents, right? And yes, the pie of work is finite. You have your finite pie of work. But because you've got all these sub-agents in action, are the sub-agents doing things to the Nth level of token efficiency that a single agent would've done or a single...

30:09 It's the same thing anthropologically as when you think about humans in a workplace. If one person says, "Everybody get out of my way, I'm going to own this task single-handedly. I'm going to do it as efficiently as possible, but I'm one person." This is like queuing theory. How much throughput do you have? Multiple sub-agents means that you go faster, means the work gets divvied up, but the pie of work might get a little bit bigger because those sub-agents are at liberty to do certain things. The point of this is best probably articulated by something that has almost nothing to do with what we've talked about, although I'm sure it'll come up. It's this organization called METR, M-E-T-R, Model Evaluation Threat Research. John, you're nodding, so I'm not sure if they've been on the pod or...

Jon Krohn: 30:56 I talk about Meter probably more than any other single thing on the podcast. And then almost every talk that I've



given for a year or two now, near the beginning, I show Meter charts.

- Ish Shah: 31:08 Ah, so our presentation and your presentation are basically starting the same way, and then yours continue to be smart and mine kind of plateau. METER, Model Evaluation Threat Research, and SDS listeners are going to be familiar with this at this point, has a chart, which when you land on their website, maybe we can put it in the show notes here, it shows on one dimension time, like 2021 until now. And then on the other dimension, it shows the ability of a model to operate unsupervised to achieve a certain goal at a certain fidelity of accuracy compared to a human given the same task. Now Meter, the reason they have this big scary name, which says threat research inside of it, their whole point was like, "Hey, at what point is AI going to cause harm to human beings? We should probably be tracking that." And the heuristic they came up to track that with is this chart.
- 32:07 How much can it do by itself? And that chart is just like, not only is it up and to the right, it's gone vertical. And at a certain point they just kind of said, "I don't know, it just keeps going up."
- Jon Krohn: 32:18 Since the release of Methos, they can't really track. It has gone off of the meter charts because in order to be able to benchmark the performance of models effectively on one of these charts, you have to have had humans doing these tasks and know how long it takes humans to do these tasks as supervised. And that was easy. 2021, when you're looking at GPT-3 level capability and the tasks are only seconds long or then minutes long with GPT-4 on average, it's very easy to come up with tasks that you can give humans to do, and it's not that expensive to pay them to do it and figure out how long it actually takes them on hours to do it. But now that Methos is doing or Fable or Astra, GPT-6 from OpenAI,



that class of models is now doing dozens of hours of work, work that would take a human dozens of hours.

33:11 It could take the AI model 30 minutes or whatever to do something that takes a human 16 hours or 24 hours or 36 hours. We don't know how long those tasks... We don't know how capable these models are because we don't have any human benchmarks. It's hard to even think of write a book chapter, write a book.

Ish Shah: 33:29 It is quite literally off the charts, quite literally off the charts. And they accidentally invented a chart for one purpose is now the best visual we have for capabilities of models over time. But as these capabilities go up, it's Jevin's paradox here. Even if token costs get cheaper over time, the base is going to move on you because people are going to realize they could do things like... It took Nintendo how many years to develop a Pokemon game? They'd come out every two or three years when we were kids. Now it's like in a weekend, someone can sit down and build a video game to the same level of fidelity. The token consumption is growing and it kind of doesn't matter how cheap you make the individual token if the order of magnitude of usage is just constantly chain reacting on itself to get bigger and bigger and bigger.

Jon Krohn: 34:28 And we're rounding up a great month with episode number 1027 in which Dr. Dilani Kahawala, a Harvard physics PhD who spent a decade in product leadership at Etsy, Facebook and Atlassian, tells me she has had to throw almost all of that experience away. Delani is now co-founder and CEO of Anna, an always on AI assistant for families, and she walks me through the three problems that turn out to be hardest when your product has no user interface and never stops running. So with your extensive product management background, and so to go through this, after doing a Harvard PhD in physics, you then went to McKinsey as an associate, Etsy as a senior... Well, product manager, then senior product



manager at Etsy, lead product manager at Facebook, and then group product manager, head of product, head of product management at Atlassian. So a decade of experience in senior product leadership positions, and now you're full-time creating Anna.

35:29 You're the CEO of the business, but you're surely also the head of product.

Dilani K.: 35:34 That's kind of all we do. The funny thing is I've had to pretty much throw away a decade of how we think products should be built for people.

Jon Krohn: 35:50 Oh, really?

Dilani K.: 35:51 And that's been really fascinating. So we've had to think very first principles from when you don't have the crutch of a user interface, that's one problem to solve. The other problem is the average person doesn't really yet regularly interact with AI agents. So I'm on called code every day, all day, and my mode is I just ask and it will have an answer for everything and the mode is you just ask and it gives, but I don't think the average person is yet familiar with deterministic set of options that you can tap and drop downs and things like that. You just asking an agent to do things for you is still a different mental model. And so the second challenge of how do you get someone to that operating model where you just ask? And then I think the third thing is when you work with cloud code, it's not inaccurate, but you need to correct it.

37:09 It will give you an answer, but you'll have to cross check it and be like, "What did you think about this? What do you think about that?" And these models are optimized for coding. They're not trained on household data. So it doesn't inherently know what to do with which calendar does this go in? It doesn't have that understanding. But consumers don't have that much patience for your assistant getting something wrong. It doesn't constantly



be correct. You don't want to be constantly correcting it. So those are the three things that I think we've had to really think about how to... Reliability, the interface, and just teaching users how to work with an agent have been the three biggest challenges.

- Jon Krohn: 38:03 I guess something that's quite different about a agentic interface like Claude Code and what you're building is that in ClawdCode, it is still a turn-based conversation where yes, it goes off. It's agentic because it figures out how to tackle a task, spins up sub-agents as it needs to, but ultimately when it's done what it's doing, it just stops. It gives you an output and then waits forever. And if you never come back to that chat, nothing ever happens again in that chat. It seems to me like with Anna, there will be times where Anna needs to reach out, where maybe Anna has sent the last message and needs to send another one before you've responded because something has changed with your kids' football practice or an important email has come through or a reminder of an upcoming appointment or something like that. So it seems like it's more discursive, more back and forth.
- 39:10 It's not as linear or just turn-based back and forth conversation.
- Dilani K.: 39:14 Yeah. And this is, I think, the biggest change. So I think when... Anna is a long running agent, meaning that it doesn't kind of stop and wait. It is constantly working every second, every minute, working for you behind the scenes. So when the open calls of the world, the Herman's agents, and now we see maybe Grockbot, all these agents are trying to tackle the same problem of how do you just continuously work with someone, but with not that much success because most people set up an open call and then they kind of give up on it after two weeks. Because what we have to have happening behind the scenes, Anna's constantly working for you on a set of things, whether it's checking your email or figuring out if a piece



of information is noise or if I've handled this before, is it already on your calendar?

40:14 Have you already tackled this task? Which kid is this relevant for? It's constantly working in the background. And you're at the same time having conversation with it where Claude, we will have a team of domain specific experts, agents who are going and doing a bunch of things for you, but Anna will be like, "Oh, I picked up that your meeting changed and it's going to clash with your school pickup. I need to interject and give you that message while you might be asking Anna to book a restaurant reservation." So we've had to figure out how to handle that so I queue things up in the correct way. There's a whole layer of ops that Claude doesn't have to deal with yet.

Jon Krohn: 41:00 Yeah, that's a really interesting use case there that I hadn't even talked about in the way that I was like, "Oh, this must be more complex, not just having back and forth." But it is also interesting that you could be having a conversation, you're in your car talking to your car, your car phone, but your car phone is Anna on the other side and you're having a conversation about scheduling some upcoming event and then it has to actually interject and say, "We're going to have to take a pause in the conversation that we're having because this important thing has come up." That is a really interesting... And yeah, I have never experienced anything like that in any conversation with a non-human to date.

Dilani K.: 41:43 So this is actually really obvious in voice mode. So when you put voice mode on, you could be... If you ask Anna to do something complex, like go find me a dentist, she has to go do some research and she has to look up where you are and who's best reviewed. That task takes sometimes like a minute or two because that's a complex task. In the meantime, you might, and voice goes pretty fast, you might have fired four or five things at her. And so we



fanned out a bunch of agents who are doing multiple things for you, but it has to be then queued up in the way that the conversation piece of it is understanding, okay, you asked me this first, then you asked me this thing, this thing is finished. Okay, now I'm going to finish what I'm saying to you and then get back to you.

42:36 That was a fascinating challenge. I don't think what's interesting is when we started, and it was only a few months ago, the voice models then were not good enough to do that, to even handle that upfront conversation. And it's only three months ago that Gemini live changed substantially. It could handle the conversation piece while we have the agentic brain behind the scenes doing all the fanning out and...

Jon Krohn: 43:04 All right, that's it for today's in case you missed it episode. To be sure not to miss any of our exciting upcoming episodes. Subscribe to this podcast if you aren't already, but most importantly, I hope you'll just keep on listening. Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperData Science podcast with you very soon.