

MacroVoices Episode 549

Feature Interview with Matt Barrie

Recorded September 10th, 2026 · Hosts: Erik Townsend & Patrick Ceresna · Feature guest: *Matt Barrie*

Sponsored by Big Picture Trading. Full episode and downloads at Macrovoices.com

Matt: And if you think about the scale of \$1.65 trillion in debt in five years, subprime peaked at \$1.3 trillion in 2007, and that, that was, had 55 million mortgages behind it while this AI explosion has two customers, OpenAI and Anthropic

Erik: That was Freelancer.com CEO and artificial intelligence expert Matt Barrie.

I'm Erik Townsend, and this is Macro Voices, the free weekly podcast targeting professional finance and sophisticated private investors. Episode 549 was produced on September 10th, 2026. The artificial intelligence arms race is really taking off, and the pace is accelerating, and the stakes have never been higher.

Freelancer.com CEO Matt Barrie needs little introduction to our longtime Macro Voices listeners, and this might be his best AI interview yet. So make some popcorn, sit back, and get ready. I intentionally let this one run long because the subject matter is so important. Matt also produced another of his famous missives on the subject of AI, and you'll find that linked in your research roundup email.

The difference this time around versus previous interviews is that in the past, we covered the actual content of the writing in the interview. This time, they're two separate pieces, giving you a double dose of Matt Barrie on AI. So listen in here for the audio, then download and read his latest piece, which is linked in your research roundup.

Patrick: And I'm Patrick Ceresna. Let's dive into this interview.

Erik: Joining me now is Freelancer.com CEO and founder Matt Barrie, very well known to our audience as Mr. Artificial Intelligence. Matt, I want to start with just an update. So much is changing in the world of AI. Since the last time that I spoke to you and interviewed you here on Macro Voices, I'd say probably the, the biggest thing is agentic AI.

I think really understanding what that means takes some understanding. So put it in perspective for us.

Matt: Thanks, Erik. Thanks for having me. I've been totally red-pilled in the last month or so from what you can do now with AI and agents. What I mean by this, I think AI has reached the point where now you can quite reliably get it to do workflows, which means that you really now have an ability to take something that someone in your company might be doing in a, as part of their daily job and automate that in a very reliable way at greater than human capability in many circumstances and very quickly and very efficiently.

So to give you a concrete example of the sorts of things that I've been doing, and I currently have about 40 agents, 44, I think last night, agents in my fleet that I personally put together to do various bits and pieces. I could look at a team, and quite a number of teams in large companies have workflows.

So one might be queue processing. Every day, projects come into Freelancer.com posted from the world, and it's free to post a project. We do have a statistical classifier that kind of runs over those projects to determine whether they're good or bad. But sometimes the statistical classifier doesn't really know the answer to whether the project's good or bad, and we'll escalate it to a human to review.

And so there's a team of people that twenty-four by seven will go into that queue, look over projects, look over their provenance, look through a whole bunch of metadata related to the user, Etc. And so forth, and make a decision on whether to approve that project or not. Now, to get twenty-four by seven operation I need about 11 people in that team, and they don't just process the projects, they also deal with violations that are reported on projects that are posted, and contests, and a range of other things.

And in the space of about two sessions of half an hour to an hour, you can now completely automate that queue processing. So what I mean by that is you can have an AI go in there, pick up things that the, the human would normally process in the queue. The human might take, I don't know, five... three, five minutes to process that item.

AI can do that a lot quicker, reliably, 24 by seven, and all of a sudden the queue is completely automated. And I had the other day a, a performance marketer quit on me. That- that's a fairly expensive role. It's a six-figure role in the order of, I don't

know, \$150,000, \$200,000 I guess if you're out in the market trying to replace that role.

And what do they do every day? They have a very repeatable role, which is they look into Google Ads every day, they look through the dashboard graphs performance on the site, they literally look through all your different ad campaigns, and they figure out what should they tune today. Should they shift a little bit of budget from one portfolio of ads to another portfolio of ads?

Etc.. And really it's a fairly repetitive role in many circumstances, tuning knobs and dials. And as that person left, I thought to myself, for the first time, I should really learn how AdWords works, because we spend in the order of half a million dollars a month on this. So I thought if I'm gonna replace the role, I really should understand how it works.

So I jumped in, and I was using Claude to tell me which reports to download, and explain the interface inside Google Ads and so forth, and cutting and pasting bits and bobs backwards and forwards to Claude, and uploading CSVs and so forth, and getting a handle for how it worked. And then I had a, a, a breakthrough moment where I was like, "You know what?

I could probably get an agent to do this." And every day before I get into work, look through everything, look through our Google Analytics, look through our Google Ads account, look through our database, look through our financial performance in the last 24 hours, look through the code, look through what's been shipped, look through Fabricator, which is the ticketing system, and basically provide a comprehensive report in my email every morning reliably when I come in that will tell me how I should modify the ads account.

And it does it at a superhuman level because it has access to all that data and reviews all that data, and no human, no matter how much I looked into the market to find someone great, there's no way that they would be getting up every morning at 4:00 a.m. and producing a report by 4:10 for landing in my inbox to make sure when I come to work in the morning, it's there.

I don't want to replace that role, a technical specialist in performance marketing. I'd rather hire someone maybe a little bit more marketing-orientated or think down funnel in the whole flow of the customer journey And in the last couple of weeks

in particular, I have been going through team by team looking at which teams are basically queue processors, i.e.,

they come in the morning and there's a whole bunch of things queued up for them to do, whether it's analyze some data or process a report or whatever it may be, and sit down with the team leaders and basically automate that flow. And the speed at which you can do it is absolutely phenomenal. For 17 years, I've tried to have my team produce a daily report in the morning by 9:30 AM in their time zone, and for 17 years, I don't get it at 9:30.

I get it maybe at 11 o'clock in the day, 2:00 PM, or maybe they forget to do it. But now I've got AI producing that report, and reports for basically every team about what they did in the last time period, how that's tracking, if they ship something to production, Etc. And so forth. And this is really just lifting the productivity across the board, and it's achieving insights into my company that I've never been able to achieve before reliably at a superhuman level.

Erik: Now, all of that I'm sure was covered in just a single \$20 per month Anthropic subscription to Claude.

Matt: No. So that, therein lies the rub. So I went and looked at my OpenRouter token usage for that day, and it came in at, on, on one particular day last week, and it came in at four billion tokens in one day, and that's four billion with a B.

Now, I looked at that and I thought to myself, "Jesus." that cost me about \$1,300-what's that in dollars,

Erik: roughly?

Matt: That cost me about 1,300 US dollars. Okay ... now, I was using western models. \$900 of it was Sonnet, Anthropic Sonnet model, which is their cheaper, faster model. There was a little bit of Fable in there for doing things like writing effective tickets to feed into our AI software engineering team, 'cause we've got a thing called Diffity which will process tickets now and actually do the engineering.

There's a bit of Groq in there, Etc., but it was at \$1,300. And I thought to myself, "Jesus." This, I-- Th-this is adding huge amounts of value to my company, but I'm only just getting started, and at 1,300 US dollars, I could imagine going from 40

agents to 400 agents in no time, and all of a sudden that becomes a, a \$13,000 bill a day, and I could probably even add a zero to the end of that in terms of the usage.

So the first thing that, that kinda starts coming towards my mind is how can I cut the cost?

Erik: H-Hang on. Let me pause you just for one second there because I do wanna understand this in perspective. I heard you say to me off the air the other day, "Oh my God, \$1,300 in tokens in one day."

We're used to getting free stuff on the internet, but you also told me in this conversation so far that y-you had just one of the employees who used to do this stuff for you. It's a \$200,000 a year position. So that's \$800 a day right there. You're spending a half a million a month on AdWords. If somebody just optimizes that to save 10% of that, that's worth by my calculations, about 16, \$1,700 a day.

So you're up to \$2,500 a day just right there before doing anything else. This is much bigger than getting free stuff on the internet, but it sounds like it's kinda worth it.

Matt: It certainly is, but then you start thinking to yourself I'm gonna 10X this," because that, that's all the work I did last month.

What's gonna happen next month? And then you're also thinking about how's pricing gonna head in the future, and are they gonna escalate prices? And we can talk about that a little bit later in the show. But the first thing you're thinking of that was, that's a month's worth of work getting that fleet up and running, and yes, there's some optimizations you can do.

And it turns out that four billion of token burn in a day wasn't really four billion in a sense, in that these models today now have what's called prompt caching. When you do agentic work, what you're doing is, a lot of these agents are just going in loops. So for every hour, they might be checking a queue, or every hour they might be, scanning something or checking the database or what have you.

So there's a lot of tokens that kind of just get reused in the models. And so it turned out with my four billion of tokens I spent in that day, three billion of them were actually in the prompt caching. If that was not the case, and I was using an Opus level model, it would've been probably \$80,000 worth of spend.

But, I was using a low-grade model, and there's a lot of caching going on. And it also turns out my code wasn't very efficient, right? So the next thing that came along was I kinda looked at it and thought, "Okay, it's not a few of these things I can that I'm doing around the polling and so forth, I can actually push into software."

The neat thing is with AI is you can write a prompt saying, "Please optimize yourself for cost and take a look at your token usage, Etc. And try and reduce it." So it turned out that by just typing in that one single prompt, I managed to optimize the token usage down by another 85%. Still, if you think about it intuitively, that was \$1,300 of spend, and the first thought I had was which other models could I potentially switch this to?"

It turns out that if I went to the Chinese models, and there's been huge breakthroughs, I think as we talked about last show with the companies like Z.ai, which produces GLM, DeepSeek, Qwen from Alibaba, and so forth, that if I wanted to run this on the Chinese models, that \$1,300 of spend would've been closer to \$150 of spend.

And so that's \$80,000 on the Opus, 1,300 the way I did it- \$150 on G- GLM 5.3, which is a 500X spread just switching between models. And the crazy thing is, as, as we've talked before, is there really is no switching costs with the- with these models. Yes, at the high end the Fable class models from Anthropic, you can c- you know, clearly see that when you use those models that there's a high level of intelligence and thoroughness that comes through, but you're paying a, a, an astronomically high price for that.

But the Chinese are hot on the heels. GLM 5.3 is almost Fable class or just getting there, and it's open source. You can download it, you can run it on your own hardware, or you can run it on these hosted environments, and it's pretty damn neat. In fact, when they launched it as Ox Alpha, and they actually made it for free and gave away 100 trillion tokens.

And for a whole week, you could switch your models and all your agents across to run on this, this free resource and get pretty damn good performance. Now, the problem with all of this is they're all training on it. So if you're wondering why, why it's free, it's because you're probably the product.

And as we observed over the course of that week, there was a lot of comments on the internet saying that the Ox Alpha model was improving quite dramatically, and that's because I think 42 trillion tokens were shoved into it when it was available for free and it used it as training data. Now, I don't doubt for a second that Anthropic's doing the same thing in some very tricky ways, and I don't doubt that OpenAI is doing the same thing a- again.

But so you start getting into this situation where you're thinking to yourself, "Okay, I'm burning \$1,300 a day. I could turn it down to \$130 a day probably moving to the Chinese models, but then all my data goes to China." And then you're thinking to yourself what are your other options?

'Cause I know I'm gonna go 10X, I know I'm gonna go 100X from here in terms of what I can do. Where do I go?" And that's where I started thinking about buying my own hardware. So I called up my local computer shop and I said "Can I get a, a pair of DGX Sparks?" And this is the NVIDIA little AI supercomputer.

It's a little modular form factor box. It's about 10 by 10 square centimeters. They were about 4,000 US dollars at the time where I bought... I bought two pairs, so I bought four of these boxes and I set them all up. It's pretty quick and pretty easy to set up. And what's amazing is you can actually load up something like the Pi coding harness, which is Pi, and you can just basically tell it, "Please set up the box and, download DeepSeek or download GLM 5.3 and configure yourself."

And you sit back and you watch it for 15, 20 minutes and it's up and running for you. So the AI's really helping you get these things up and running. But-

Erik: Okay. H- hang on a second. Before we go on, I just wanna qualify. You spent about \$4,000 for each one of these boxes. How much of the 13... You were spending 1,300 bucks before.

Does one box like that replace the whole 1,300 bucks or do you need more than one box? And also- Yeah ... you said you, you gotta use the Chinese models because they've gotta run on your own hardware, which Anthropic and OpenAI won't allow you to do with their models. Yeah. I can't believe that DeepSeek is just as good as Fable 5 and Claude.

So what percentage would you say? You're accepting a little bit of capability downgrade in order to put it onto your own hardware. How would you characterize

or describe how much of a hit you're taking in terms of the capability in order to get that benefit of your own hardware and therefore unlimited use?

Matt: Yeah. So DeepSeek version four Flash, you can download that and run it on a pair of DGXs. So there's a cable that plugs the, the... These NVIDIA DGXs are modular, so you can stack them and you can chain them together. And a pair which shares the memory between the two on a, a 200 gigabit link, you can load in a 300 billion parameter, which is quite a big model, DeepSeek V4 Flash, and run it on your own hardware, and you get around somewhere between 45 and 60 tokens per second.

So that gets me to about 5 million tokens a day, and my 4 billion token day was about 46,000 tokens a second. So by comparison, it's about a thousand X difference. But effectively, a pair of DGXs can run DeepSeek V4. You get kind of Opus level quality. It's pretty damn good. It's fast enough for you to program in a couple of windows because you can get up to 100, 120 tokens a second running multiple streams, like six streams at the same time.

So you might have one window open doing a bit of programming, another window open doing something else, Etc.. Yeah, so it's about a 9,000 US dollar spend, or at least it was when I bought them, where you've got, it can power a programmer, a single programmer doing a bunch of things. I obviously haven't got all the bandwidth I've got with all my agents running because the agents are burning tokens while you sleep not while you're driving it interactively.

And if you do the math on the DGXs to run my kind of workload with the code optimizations and so forth, I would've needed about 16 of them running, right? Now, the interesting thing about that is that, th-that would've cost me about 65,000 US dollars in terms of hardware to do that. It draws collectively less power than a kettle to boil 'cause they're about 100 watts each.

Lit-literally one of those supercomputers runs at, 100 watts, so it's about a strong light bulb. And then I can actually run that workload even if the internet turns off. And the other great thing is that when these models come out in the open source world, whether it's DeepSeek or whether it's GLM or whether it's Qwen from Alibaba, is you give it a couple of days or give it a week and the open source community does a thing called obliterating.

They obliterate the models. So what they do is they sit there and fine-tune the models that are, that have been publicly released to uncensor them. So all the refusals that you get, with Anthropic which is incredibly annoying and I guarantee you is hurting their adoption in a big way, where you ask it to review a legal document and it says, "Sorry, I can't do that, tool policy error."

And it's I'm g- it's my own terms of service. Why are you giving me a policy error? I'm trying to get you to help me improve one of my own documents. You don't get that 'cause what the open source community does is it gets these models and then basically fine-tunes all the refusals out of the model.

And all this AI safety stuff that's going on, it's all pretty pointless with open source because every model out there has been obliterated, and you can download it and it won't refuse any, any request. So that's an advantage in that you don't have these models sit there refusing things all the time, and the Anthropic models are pretty bad.

You can't even just get ordinary work done in the Fable model without it refusing all the time. So that cluster would be at 16 Sparks. It'll cost me about 60, \$65,000. If you amortize that over two years, that's it's about 100 bucks a day of hardware and then the power usage. It's about a dollar a day to power two Sparks, so 16 Sparks is probably somewhere around eight bucks a day.

So you start thinking about that sort of economics, and so you can see the Chinese models, when you go out there and you use the Chinese models that are hosted and you rent them, they're a little bit more expensive, but they do offer frequently discounts. If you go to OpenRouter or LLM Gateway is another OpenRouter style site, which is, really just a, like a, a meter or a place where you can select all the different models, Etc..

You've got different hosting providers offering discounts all the time. The problem is, because everyone's training on the data that you're sending it, not my AI, not my data. And I really do think, and we've talked about this in previous episodes, that this whole thought train of, wow Gee, where's everything going in terms of pricing?

I did note by the time I went back to the computer shop and bought four more DGX Sparks, I got eight now, I've got four at work, four at home, that the price had

already gone up, and also the distributors were saying they're out of stock, and in fact, there's a n- nice little switch by Mikrotik that will chain four of them together.

So you can unify the memory across all four, and you get 512 gigabytes of memory. And so you could load a much bigger model in much more easily. They're completely out of stock, not just near me. They were out of stock in the US, they're out of stock around the world. The last Mikrotik switches that were available, I think it's the 8112 model, were in Sweden, and by the time I hit reload on the page, they're out of stock.

So I think everyone's got the same idea now, which is to thinking about buying their own hardware. And I certainly think, and I think we've talked about this extensively, there's gonna be an emperor-has-no-clothes moment which is coming very soon, where enterprise is gonna be really fed up that everything they put into G Suite, everything that they put on the public-facing internet, is scraped by the AI scrapers.

The traffic to my website, Freelancer, has gone up 1300% in the last 12 months which is ex- incredibly annoying in terms of getting the content delivery network bill and having to reconfigure the network or, you gotta be careful what you block on the AI scrapers because some of those AI scrapers also index you for the search results.

So if you block the AI scrapers, you can damage the, the customer acquisition coming in. So you gotta be careful about that. But I think there's gonna be a moment coming real soon, like now, where a lot of these large enterprises are like, "You know what? I don't want these big foundational models reading all my email.

I don't want them reading all my data if they host it for me on AWS or what have you. I don't want them knowing what my weaknesses are where my valuable customers are, my strategies, Etc.." Because everything that's out there on the internet is being sucked into the AI, and it's used for training.

And I think there's certainly a trend, and you can certainly tell it from when you go to the computer store and you try and buy a DGX Spark, or you try and buy the RTX graphics cards. RTX 6000s are pretty good for running very high performance. Two of those cards cost you about \$50,000, and you can run something like GLM at 1,000 tokens a second and run a whole engineering team off them.

There's a big scramble for this hardware, and I'm seeing it also in one of my other businesses, which is escrow.com, because what will very soon be our second biggest category is GPUs and servers, as we're, there's a very active broker market out there that are buying and selling, data center hardware.

Some of the stuff is coming off lease now because the Hopper series is starting to be retired. As you got, and then you've got the Blackwell series, and then you've got now Vera Rubin and so forth coming. Some of that stuff is, that old kit is starting to get rotated, so there's a very active sort of secondary market that is booming.

And I think there's just gonna be a scramble for your own hardware.

Erik: Okay, let me understand this in a little more detail 'cause I think it's gonna interest a lot of our listeners. It's not just the model that you can buy, but it's also the degree of integration with all of this new agentic stuff. 'Cause as I hear you talking, like you're, you're running a public company, I'm retired, I'm a t-totally different set of circumstances.

I'd be delighted to buy as many of these DGX boxes as it takes if it would allow me to do just one simple task, which is automate going through my email every morning and filtering out all the spam and organizing and prioritizing it without sharing any of my content with Anthropic or OpenAI or anybody else, because I don't trust anybody any more than you do.

Is that degree of integration where this DGX box and a Chinese model has that agentic ability to go talk to my Microsoft Outlook and command it to delete and move things around into folders? Does it do all that? And if not, how long until it does?

Matt: It's trivial. There, there's a very active community around the Qwen 27B model, which you can run off a laptop

Erik: Okay.

But if I wanna buy one of these DGX boxes to run a, a Chinese model or two of them or whatever the requirement- Yeah ... \$10,000 worth of hardware, I can not just run AI to answer fascinating math questions for me, but I can run AI that talks to my computer, controls applications, does stuff for me, but it's all ring-fenced so

that nobody is sharing my data with anybody else on the internet, and I've got really solid confidence that's not happening behind the scenes.

Matt: Your problem is that your email's probably in Gmail and Google has it, so you gotta get your data out of that first.

Erik: It's actually coming from several different POP3 and IMAP servers, and it's going- ... into Microsoft Outlook, which means something needs to be able to either control Microsoft Outlook or deal with the Microsoft Outlook file format or something.

My inbox consolidates a whole bunch of mailboxes into Microsoft Outlook.

Matt: You can guarantee all those big cloud companies are peaking at the data. I-- you know that because when you log into Gmail, and occasionally there are ads in Gmail, so they're obviously training for their a- their ads to be context-specific.

So the problem is with all the SaaS stuff is that this data is sitting in, inside these big hyperscalers, Google Microsoft, Amazon, Etc., and the temptation to peak on it or train on it is too much. There's all these little tricky things they do, updating terms of service and by default flicking certain things on, and particularly if you're using the free versions of anything, usually the free means the product is you and it's your data they're training on.

And I do think there's a trillion-dollar industry available right now around confidentiality and privacy, and I think Apple's trying to make headway in some ways, into that privacy space and so forth, but I don't know if I completely trust them either. Certainly as they head into the next version of Siri and the handset, that's too much of an invaluable device for training data.

'cause you gotta think about, where are the contemporary data sets coming from, right? I do think enterprises can be fed up. I think particularly at the big end of town, you'd hate to be a McKinsey or a Deloitte at this point in time in many ways, 'cause it seems that's potentially where the direction that Anthropic and might head is limiting access to its API for high-end, its high-end models because the Chinese just come in and distill it by getting a whole bunch of subscriptions and sucking the knowledge out of the model.

The big guys do it to the Chinese as well, Etc., through open models. And the big guys at the end of town might start becoming more like a consulting firm trying to solve these, big problems for pharma or for defense or whatever it may be and You have this sort of stratification in the market where the commodity tokens come from the Chinese on open source and open models.

But I do think enterprises are gonna get really fed up having their data out there and having it trained on and having it in the cloud.

Erik: So if that's the issue and more people are gonna get more concerned about it, I could see it being resolved in either of two ways. Either everybody does what you're already doing, so there's this huge new industry of hardware that allows everybody to run their own AI in a ring-fenced environment off the internet, or the market responds to need and, somebody comes along and says, "Look, we're going to be the leader in actually respect your privacy AI, and it costs more.

And guess what? It is somehow delivering some technologically assured guarantee that you can really believe in that it's not basically... You're not the product. It's actually doing what it's supposed to do for you, and it's not stealing your stuff." You could do that with trusted public SaaS services if there were trustworthy people selling them.

That, that might be a harder problem to solve. What do you think?

Matt: I think you've gotta step back and look a little bit at the economics of the space and how it's playing out at the moment. You've obviously-- One thing that's become very clear in my mind is that the amount of tokens that on a per person basis or per enterprise basis that are going to be consumed are gonna go through the absolute roof, right?

If just by myself, and somewhat naively, I can burn 4 billion tokens in a day, and that's only really after a, a month or so of sitting down and getting s- a bit of an agentic framework up and running. You know- The token usage is gonna go through the roof. Now, it's gonna go through the roof, but I'm not gonna pay the prices that what the, the frontier models wanna charge.

So while token usage is going through the roof and you've got this big build-out in data center capacity, and I will note that this is all fueled by debt now. If you look at Ed Zitron, I think I've talked about him before, he writes a, quite a a negative

blog on the whole AI space. He says, "Data center special purpose vehicles are this bubble CDOs, collateralized debt obligations, and AI data centers are at subprime mortgages," is how he looks at it.

And he says, "What's happening is these hyperscalers are loading all the risk into a special... To build a data center through these special purpose vehicles which own the chips and the debt and the risk, Etc., and takes it off the balance sheet." So Meta, for example, has \$46 billion of exposure to Hyperion in its filings, but the balance sheet shows none of it.

So there's these huge debt-fueled explosion in data centers and data center build-out, where in the last five years alone, the hyperscalers have loaded on \$1.65 billion of debt, \$500 billion from data centers and \$200 billion from pr- from private credit. And that's before all the off-balance sheet stuff.

And if you think about the scale of \$1.65 trillion in debt in five years, subprime peaked at \$1.3 trillion in 2007, and that, that was, had 55 million mortgages behind it while this AI explosion has two customers, OpenAI and Anthropic. That's literally true. OpenAI and Anthropic are 73% of Amazon's AI revenue and 74% of Microsoft's, soft's, and over half of all Google Cloud by 2027.

So it's kind of interesting dynamics in that you've got this big data center boom, OpenAI and the hyperscalers are effectively financing it in, in, in very financially engineered ways. You've got the Chinese coming in undercutting them with models that are, DeepSeek is Opus class and GLM-5 three is almost fable class.

And the token usage may go up dramatically, but it's not gonna be at the pricing that, and OpenAI wanna charge. And so it's interesting to see where this goes. I think you talked about this before, we're on still on the teaser rate for token pricing, and I think that adjustable rate token pricing might run out when the lenders start doing adjustable rates on their lending.

And I think what's interesting is if you kinda watch what Jensen's doing, four of his customers now are producing their own chips. So I think OpenAI just announced the Jalapeno chip, which is supposed to be twice as efficient in terms of what's per inference, and that's a chip designed for inference rather than training.

When four of the biggest customers are building their own chips and the rental market looks like a debt bomb, he's now starting to sell the box to the end user and

let the, all the SPVs fight over the leases. So he, he's got the DGX Spark, which I can buy at, it was 4,000, it's now I think about \$5,000 a box.

You've got a prosumer model, which is the, the, the DGX station, which is about 100 grand, which kind of steps it up. It's got about 748 gigabytes of memory, and I'm sure he's gonna come out with more products directly to the enterprise. If you think about my, my 12,000 tokens per second usage and my four billion a day for agent sales deploying it, you can buy the generation one, the, so the H100s, which are the Hopper series, eight GPUs in a box, which does about five to 10,000 tokens a second, which is what they would deploy in the first gen of these data centers, and they cost about 150 grand to 200 grand if you buy them secondhand.

And the new Blackwell, the newer Blackwell B200s, they're three to four times as fast, could do about 20 to 40,000 tokens a second. They're about half a million bucks each. So if you amortize those boxes, they're somewhere between, 400 to 600 bucks a day to get your own eight-way Hopper box or, \$700 to \$900 a day to get your own, B200.

So I think there's gonna be, NVIDIA seems to be sidestepping this kind of, I think growing catastrophe between OpenAI and Anthropic, and both of them are trying to get out on the public markets as fast as they can, and go directly to the consumer, directly to enterprise, directly to prosumer, and make a big move into the open models, 'cause I think the open models are gonna win here.

It's, I think it's quite clear that they're least they're gonna win the bulk of the market. They're gonna win the bulk of the token market. That's why I think Jensen bought Hugging Face just recently which was a, what? \$12.9 billion acquisition. And Hugging Face is started by some French guys, and it's basically the place where all the models are, open models are hosted.

So if you-- when a new version of DeepSeek comes out, you go to Hugging Face and you download it, and then all the enthusiasts there will upload various derivatives. They will quantize it, which means they will downscale it so it can fit on smaller hardware. They will obliterate it, so they'll tune it to remove the censoring.

They will make various trade-offs and engineering decisions around getting it efficiently working on, two DGXs, four DGXs. And I've even seen now some enthusiasts looking at mixing the hardware up and using the Mac, the new Macs

that are coming out, the studios and so forth and mixing them with DGXs to try and get certain, engineering trade-offs.

So NVIDIA seems to be, I think, going, "Okay, there's a bit of a catastrophe in the data center market." That- that's where they make most of their revenue at the moment from all the stuff they sell from their Mellanox acquisition, and I think they're now going, "Okay, there might be a catastrophe there.

Token usage is gonna go through the roof. The, the open models seem to be winning," which is kind of the Chinese models, which is the Chinese strategy, open source everything. "Let's start selling the hardware directly." And the hardware they're selling directly as well y- you gotta remember that NVIDIA, in a way, is a bit like a software company in that they take the high bandwidth memory that's produced by three companies, SK Hynix Micron, and Samsung, and they package it up around a GPU, and they mark it up, with a 70% margin, they sell it to the market.

So they're still dependent on the high bandwidth memory, which is a memory architecture that's very important in the data centers to run a- at scale. And so with this direct hardware that they're producing for you and I to try and, not my AI, not my data avoid that sort of model and also to have some sort of hedging against what might be potentially highly escalating token costs as these subsidized programs come off, as the lenders start getting a bit shaky about this sort of data center debt.

The models are good enough. It's expensive, but cheap enough, and I think you're gonna see an explosion of this stuff coming out to the edge, and NVIDIA's gonna go direct. That's what they seem to be doing, and that's why they bought Hugging Face, and that's why they're producing these, these boxes. So if Anthropic and OpenAI particularly Open, OpenAI has a, you know, blows up with all the circular arrangements and so forth, NVIDIA's still out there in the market selling direct.

Erik: Matt, I wanna put some perspective around how quickly this agentic AI revolution that you're describing here could actually happen. A- and I wanna share a story with you and our listeners, whi- which I really think is instructive, and I'd like to you to comment on it. When you and I agreed to do this interview, it was a little over a month ago, you gave me a, a little bit of context on the phone about what we would talk about today when we booked it.

And at the time I was at my summer home, and I was thinking to myself after that conversation, Matt's vision, I'm totally bought in. But Matt's basically a genius. He's on the leading edge of this stuff. The, the people who are consultants who know how to do the kind of stuff that Matt's talking about are super expensive.

Smaller businesses can't afford them. It's gonna be a long time before humans and businesses are really ready to embrace things on the level that Matt's embracing them. So this is probably farther off than Matt thinks. I swear I'm not making this up and I'm not exaggerating. I'm having that exact conversation with myself late morning one day.

I decide it's time for lunch. I go hop on the elevator to go downstairs to go grab a sandwich. I run into the maintenance guy in my building. His name is Pete. Super nice guy. Pete's the guy who fixes the plumbing and changes the lights when the light bulb burns out and that stuff. Blue-collar guy, super nice.

Known him for a few years. Usually, when I talk to him, my conversation was about his Jeep. He really liked his Jeep, and I noticed he sold his Jeep and he seemed to have a new truck, so I thought, "Hey, Pete, whatever, what you up to? H-how's your week, winter going?" He says, "Boy, I'll tell you, I am absolutely loving Claude Code.

All these years I've been using the internet, I thought it was just websites. I didn't know about APIs. I didn't know you could write your own code to actually script APIs on different servers to do different things and develop your own dashboards. It's really cool. I'm really enjoying it." It's this guy's hobby.

He's a maintenance guy, and he figured it all out because someone who is a above average, smarter than the average blue-collar handyman who's curious about how things work and willing to spend a little bit of time investigating and self-teaching with no help is able to figure out how to write AI agents that talk to APIs Whoa.

That says to me that this whole thing can happen way faster with far less specialized human skill than I ever imagined.

Matt: The AI writes the code now, right? So we went from machine code to a higher level language, assembly language, to compiled languages like C, to interpreted languages, and now we've basically moved to English.

I don't write code anymore. I use my phone and in the chat, the agentic chat harness, I use PyRemote, but there's plenty other that you could use. I literally just write English. And so someone like the maintenance man can literally say, "I wanna create a dashboard. I want it to show the following things."

Put it together and the AI might say, "Okay, I need an API to this, an API to that." And really what you do is you just log into these sites. You get the key, which is like a password, which is the AP- to access the API. You provide it to the AI, and all of a sudden the dashboard's made.

And you really see a progression of this sort of agentic work when I see people get red pill'd in my company that they start off, they start making a dashboard so they can make their job more efficient. For example, my ops team in Loadshift has a... Which is, moving heavy haulage freight.

They've got a dashboard that says, "Okay, you gotta call this guy in the next two minutes." "This is a particular customer that's blowing up at the moment. Here's your next activity to do," etc. "Here's how you're tracking towards your commissions, your hurdles your accelerator for your commissions," and so on.

So everything they need is in that dashboard. And in fact, the way my- the team leader wrote it, is they can give feedback to the dashboard to improve it. So they can... It has a little chat, little box in the side. They can just type in a bit of feedback, and bang, the dashboard update itself. So you can make dashboards very easy.

You can make reports very easy. So as I mentioned before, every morning I have a report around my ads program and what I should do and what knobs and dials I should process. But then you move to the next level, which is queue processing. That report says, "Here's the next few action items to take."

Ticking through those action items, you can get AI to do that. Many businesses have queues already. They might have a support ticket queue or a support chat queue or what have you. AI now does 100% of our tier one interactions with customers. And with the advantage there is that you get instant response, 24 by seven, any language you like.

The answers are four times longer, four times better, and 10 times more empathetic at a level of empathy that a, n- no support person in the world could ever sustain

over a 24-hour period. And yes, a lot of stuff gets escalated to humans, and then the human comes in and ha- has to do what have you. But increasingly, that's gonna be less likely the case.

And, ultimately, I think there's gonna be a massive transformation. If you think about jobs in big companies, right? We've talked about before how anywhere there's a large number of people doing the same thing over and over again, you might have, 100 lawyers, junior lawyers doing drafting in a law firm.

You, in the future you might not need 100, you might need 13. 10,000 people in a call center, you might need 100 or 200, what have you. This is now taking that to the next level because the, the AI now is getting quite reliable for what's called agentic use, which means that they can, have long loops and get, do quite complicated things over a period of time, and so forth, without drifting.

Admittedly, some of the Chinese models do have a bit of a struggle sometimes with, things like tool use and so on. DeepSeek sometimes go into a little bit of a l- a loop, but that's more of an engineering problem around the harness. That's why you pay a little bit more for the Anthropic models 'cause they do have an advantage in this agentic work for now, but that, that's being eroded quite rapidly.

But what it basically means is, what business wouldn't wanna have something that logs into all your databases or your web traffic or what have you, looks through your customer interactions and just tells you what's going on in your business every day? And tells you where the opportunities are, and tells you where the weaknesses are, and tells all the humans what they should be doing next in the most optimal order that they do the next thing.

And free you up from a lot of this work to focus on the creative, the, the really high value-adding Etc., work. And frankly, no human can compete anymore with some of these tasks. Like the ability to issue a re- a, a report from an AI agent, it's far superior to any human. It's superhuman already because no human would ever get up in the morning at 4:00, 5:00, 6:00 AM, review what needed to be reviewed in the time period that w- the, to issue the report, and do that so with the level of consistency day in, day out, every single day of the year.

It's impossible to do that. I think there's going to be tremendous dislocation in the corporate world. I wouldn't be, wanna be long commercial real estate at the moment. It's not because of the work from home, it's because it's, it, the AI's gonna

chew through a lot of these jobs. And it's gonna chew through a lot of the, every, every business.

If you're a banker, one of our big banks in Australia's got 50,000 employees and 700 branches, and you kinda think to yourself, "What are those people doing?" You don't need 50,000 people to run 700 branches. What are the workflows inside a bank? You've gotta ap- approve mortgage applications. There's probably a queue of them.

You've gotta do, credit checks and AM- AML checks. What else have you gotta do? Put a few faces on at the tellers to make people think you're still a friendly bank. But you could probably run that entire bank not with 50,000 people. You could run it with certainly 10,000 people, and certainly probably a lot less than that.

You g- you're gonna get a lot of dislocation now coming through. Where does it head? Just thinking more about the product design, the architecture, and so forth, so they're moving up the stack And perhaps a lot of the illustrators now are not really pushing pixels around the screen. They're thinking more like a creative director.

Where do we end? I think we end in a, an explosion of entrepreneurship. I think we end in a hyper-competitive, very entrepreneurial society, and I think, too, where's one of the most competitive markets in the world for this sort of thing? And you think about India. You go to India, and there's a lot of people who are not in traditional em- employment, who are...

have to fend for themselves and what do you end up? Everyone's got a hustle. One guy's making food at a stall, one guy's repairing motorcycle seats, one guy's being a tour guide, Etc.. Maybe they're doing a few different things, Etc., and I think that's gonna happen in the Western world.

You're seeing it already with people having hustles doing, UGC content, or maybe they're playing poker online, or, they've got a drop shipping business or, selling courses or whatever it may be, and I just think playing with prediction markets, whatever it may be vibe coding apps and so on, and I think that's where we head is we end up in this, this world where you're gonna have, just everybody's gonna have a hustle.

Everybody's gonna run a startup or a business, Etc.. We've had these points in human history where you've had technology get to a point where it's created such productivity breakthroughs that you have dislocation, so you've got the mechanization of agriculture. That day that someone drives the tractor onto the farm and there's 1,000 people on the farm and you don't need 1,000 people anymore, you need 20.

You've got the mechanization of the factory. You've got the computerization of white collar jobs with the transistor and the computer and then the internet and mobile. And now you've got AI, which is really going to do it to white collar jobs, and, there's gonna be... in the future, there'll be jobs thought of that we never might today be fanciful if we imagined it.

I can imagine that 100 years ago you tried to describe someone that produced memes or UGC content. They'll probably think this is a ridiculous sort of job, but, those jobs exist today. But we are gonna go through a tremendous dislocation now in white collar jobs.

Erik: Let's go a little deeper on that, that white collar job- dislocation in terms of how it plays out over time because if you talk about the long term, this seems really clear to me. People whining about losing their jobs, look, AI is an incredible productivity tool. If you take the, the bl- blue collar physical labor equivalent, before the Industrial Revolution, it took hundreds of men with picks and shovels to clear an acre.

Now one guy on a big Caterpillar earthmoving machine can do it in a half an hour because we have much more productivity thanks to automation. AI is the same thing for white collar. It's much better that we have a world where the blue collar guy can drive a machine instead of having to break his back with a pickax.

That's better But wait a minute. If you were one of the thousand guys that got let go on the day that first tractor got delivered, and not the one guy that got, you know, kept to drive it, you're not feeling good about that. Maybe your kids are gonna love the, the fact that they're growing up in a, a new, better world, but it's gonna suck for you for a while.

Seems to me like AI is an incredible productivity tool for white collar, except for the people that kind of are hanging around when it shows up What do we do about that?

Matt: Yeah. I think about it from a slightly different perspective at the same time. Someone asked me, "What do you tell your kids?"

Should they go to uni? What skills do you need to survive in this world if you are laid off because you work at a bank or an insurance company and all of a sudden they don't need 50,000 staff, they need 5,000, and you're laid off? Who's gonna swim? Who's gonna sink? What happens in that world?" And I when I see what's happening in my company, people who survive in that world and do really well are people with initiative. You need to be flexible. You need to be dynamic. I think there's a whole bunch of people in my company who are completely red-pilled. You're seeing them, watching what they're doing, and they're writing agents to automate, 'cause this whole technology explosion is essentially about automation, whether it's automation of the farm, automation of the factory, automation of knowledge jobs, Etc..

So they get excited, and they come up with all these different roles in automation. They all have teams underneath them. One thing that's i- interesting is I find now that where I've had a functional team before, for example, processing a queue, where there might be a team leader and there might be, five people underneath them, moving forward it'll be just the team leader and the AI.

You don't need the five people underneath them. It'll be the team leader controlling the AI, and the AI will be the team effectively. And the team leader's there coming up with new ideas and directing it and so forth. It's the people with initiative. It's the people who have, who are very creative.

It's the ideas people. 'Cause the ability now for, for you to execute an idea is very simple. Over the weekend I vibe coded a, a whole app for my girlfriend for her birthday coming up around fitness, and it was... It's just amazing just how quickly you c- you can put some of these things together.

So if you just got the idea and you got the initiative, now the execution's there, right? The, it really, in terms of productivity, the productivity gains are so high, you've got really good execution. Yeah, I think, y- you really have to be highly flexible. You've got to be highly adaptive.

I think the people who really like a lot of structure in their job and the 9:00 to 5:00 and know what they do ahead of time, they will struggle. But the flip side is if you want to learn anything now, everything is out there on the internet. I have learnt

through AI how to use a whole range of desktop CNC machines, desktop printers, laser cutters.

I've learnt about things that would just take me forever to... I've done electrical engineering, but I, as, after I got a 3D printer, I printed out a robotic arm. There's Parallel 6. It's quite a sophisticated robotic arm. And I managed to, with AI, figure out how to get it all working, wire up all the electronics and get a robotic arm working, which I- would've been a, a incredibly long project.

Anyone can do this. You could load Claude up, and you could just say to itself, "I want to learn a particular topic. Teach me maximally, in, in, in bite-sized chunks how to do that, and test me and to maximally increase my knowledge of a particular topic." And, as a tutor, your ability to learn anything is unparalleled, so you really, ironically- As a human, what do you need to do?

You need to learn agency. You need to have initiative a- and go out there and u- generate the ideas and, 'cause your ability to execute now is just so simple.

Erik: Okay. If you have initiative and a really strong current frontier AI model and a subscription to it and you're 19 years old. What is the reason to go to university other than the social and networking experience?

Matt: I've done quite a number of degrees. Not all degrees are equal, I think as we've found out. Some degrees like engineering, you go there to learn how to analytically think and how to trade. And, you learn basically around learn technology and this, that and the other.

So if you do electrical engineering you get very deep understanding of that. But there are other degrees like an MBA where I think anyone would tell you that half of that is the network you're forming, and that network after university life can become quite valuable in terms of your sphere of influence.

And also it's around the collegiate life and some people, that particular age just need to be around people and wanna live on campus and have that kind of lifestyle. I was an adjunct professor for many years. I do think the university curriculum is very static and very rigid. It's always been for the-- with the exception of places like Stanford, where sometimes the stuff you learn in the university ha- hasn't made it into the lab yet, and you're, you're really at the cutting edge.

Most universities you're studying things that happened 10, 20 years ago. And in many courses, the lecturer puts together the course 10 years ago and then every year does a little update here or there, and the courses become quite stale quite quickly. I do think what it does do is it does through rigor and through discipline.

I think it would be hard in a way for someone to sit down and actually go, "Okay, I've got to learn signals and systems. I've got to learn electrodynamics and all these really boring and very difficult topics, and I'm just gonna do self-direct on my phone without the peer group pressure of my university lecturer and my class, and the fact I'm doing a degree."

So I think there's some element there that won't stick. But I think for people who have got to a certain level of foundation, particularly people who have come out of university, who have got a degree, who have-- who can, have got some level of foundational knowledge, it is very good for postgraduate education and continuing education and professional education and so forth.

So I don't know. I think the nature of education has to change some, somewhat. At the very least, they've got to probably move to oral exams away from, homework in many ways. But if you've got a passion to learn something, it's all out there and it's all free. So the most important thing you need to develop is as a human is agency.

Erik: I wanna move on now to some of the big things that are happening right now in the AI industry. You've talked a lot in the past about this idea of switching cost being almost zero. When I first started talking to you about AI, you were just every day in and out, you were Mr. OpenAI and ChatGPT. I think last time I chatted with you, you hadn't used ChatGPT in months and you're Mr.

Anthropic. Now, the big thing, at least if you believe the marketing according to OpenAI, is okay, it's here it is. We're gonna make the next thing. Everybody's gonna come running back to us because Astra, their new model, they say is going to be remembered as the model that achieved artificial general intelligence.

So first, let's define what is artificial general intelligence? How's that different? What's that milestone? Does this new thing from OpenAI called Astra really actually deliver that, or is that just marketing? And are we about to see a big wave of switching back from Anthropic to OpenAI because of this new model?

Matt: There's a lot of marketing obviously going on there.

Erik: Quite a lot.

Matt: Yeah.

Erik: I was setting you up.

Matt: Yeah. Look, I think one interesting thing about the competitive dynamics of AI is that every historical moat that you would talk about in business models i- is dead in AI. Sometimes venture capitalists talk about what's your intellectual property?

It turns out that that basically the science around AI is just published in 40 technical papers is the most of it, and it's all out there, and a lot of the models are open, Etc. And so forth. While there might be a little bit of edge in some intellectual property that Anthropic has and OpenAI to a little bit of extent, essentially it's all out there, and that's why these labs just come out of nowhere like DeepSeek, and all of a sudden they're state of the art.

And then next month there's Z.AI. It came out of nowhere and suddenly state of the art, and next month there's another one. It just keeps-- You got this kind of round robin where it bounces through all the different AI labs as they rush out a model that kind of is, tuned for certain benchmarks and so forth.

There, there's certainly no customer lock-in. To switch between models is simple. In a, a coding harness like Pi, it's just simple, you type /model and then just change the model, and it just loads in the context window and loads in the model and it-you're done. There is literally nothing stopping you switching between models at all.

Zero switching costs. The agents have no loyalty. I literally-- It's quite funny. I did-- My first agents were inside Claude Code, and then I just said to the-- I thought when I went to Pi, 'cause I could get access to OpenRouter and some of the other models that are out there, I literally just wrote one line into Claude Code saying, "I'm going to move you to Pi now.

Please provide a file which I can load into Pi, which will transfer you completely across And it just dumped out a file, loaded into Pi and bang the AI's in a different

model. Or you could be a little more sophisticated and just say, "Okay, I'm gonna transfer you across just," and it just S- SSH a- across to a different machine or whatever it may be.

And then if you look at all the other things around scale, DeepSeek didn't, w- apparently with 2,000 GPUs and a side project, what these big guys are doing with, hundreds of millions of dollars, Etc.. So if you want competition, AI foundational models are like opening a Thai restaurant in a row full of other Thai restaurants in Thailand on steroids.

It is the most brutal competitive market in the world, which is interesting. A- Astra's come out. I had a play with it. There's certainly... Every time one of these new models comes out, there's a lot of fanfare and there's a lot of influencers that seem to paid off on social media that kind of sit there and spew a whole bunch of stuff.

There's a whole bunch of games that people have shown clips for on social media around Astra, which I don't know if those games last more than 60 seconds. Would you ever-- You'd ever be able to get any commercial production maybe a, a trivial little game for an iPhone app or what have you.

And then Blender, it seems to do pretty well at 3D modeling and be able to do that fairly efficiently. But, by the same token, there's a lot of people complaining that that Astra really mucked up the code base, that they asked it to do a little change, and it decided to rewrite their entire authentication system and completely stuffed it.

I tried Astra, and I had a pretty simple thing. I wanted to draw a border around a graphical design that I made ar- for this fitness app And despite six attempts and waiting a couple of minutes on each attempt, it completely stuffed up every single time. I don't think Astra is there yet.

I would think it would be incredibly challenging to be a foundational model company at the very cutting edge, specifically Anthropic and OpenAI, and have that model be great at everything. And every time you release a new latest, greatest ultimate frontier level model, it's great at computer security and it's great at Blender and it's great at coding and it's great at, drawing lines around a thing.

I think you're gonna have all sorts of artifacts, particularly when you mix in all these refusals and policy and so forth. I'm sure that makes the model worse and the ethics and so forth, I'm sure that interferes i- in a strange little way with the model.

Erik: I wanna make sure we cover the definitions though.

Artificial general intelligence, what exactly is that milestone about?

Matt: Jensen's proclaimed that we're probably there with Astra, which is, AI basically reaching human-level intelligence. And then if you follow the exponential curve that Ray Kurzweil has when he talks about the singularity, that, very soon it's not just AI is as good as a human in general tasks, not specific tasks, but instead that AI is as good as all humans on the planet, and then multiples of humans on the planet, Etc..

And, and you got this big lift off event, right? They talk about the singularity, which is really a point in time where technology's improving so rapidly, it's faster than our ability as a human to comprehend the improvements. We just get this massive lift off. And this is where man and machine starts to merge.

It's probably why Elon's off there with Neuralink and thinking about if the interface has just gone from programming language to English language, what's the next level of interface and what's the next level of agency that these AIs could have? Maybe it's a neural implant or a neural link or some sort of way of scanning your brain, so it will be able to thoughts will be able to control the AI, and I know there's been some leaps with that.

And I also know there's been some AI models they've trained on brainwave scans to be able to decode them And then maybe the next level from that is intent. Maybe the future version of DoorDash is a nice cold can of Coca-Cola arrives at your doorstep because it's known in advance that, it looks at your blood sugar levels and your brainwave activity and you, gee, you'd really love to have a Coke right now, and the Coke just appears.

And you go, "Oh, thanks. How'd you know I wanted a Coke?" Or a cold beer or whatever it may be, and so forth. But anyway, Ray Kurzweil really popularized von Neumann's singularity, where you have this liftoff event, and effectively it's where machines take off, the mind and machine potentially merge, and technology

innovation is faster than our ability to comprehend it, and we really have this explosion.

And they think it's not too far away, maybe 2032 or so, or 2030 or what have you might have this liftoff. And I think in the last few weeks you've had Jensen proclaim that AGI is here and you've had I think it was Elon said that singularity has been entered already.

Erik: I think what Elon said is that singularity is a process, not an event, and the process has clearly begun.

Yeah. Paraphrased, that's about what he said, and I agree with him. As I understand it, the really big milestone on this whole singularity thing is when AI is able to improve AI without human participation. Because at that point, all it is a question of who throws the most money at it, and it's actually the most energy and GPUs, but that translates to money.

Whoever gives it the most resource, AI doesn't need human AI engineers anymore. AI does its own improvement of itself. That's what invokes the analogy to the, the word singularity in mathematics means when an equation suddenly blows up. Like the classic example is one over X. Well, as soon as X is zero, you're dividing by zero, it's impossible, you can't solve it anymore.

At that moment, the equation blows up. There's a moment where how fast is all this AI stuff accelerating? The equation blows up when it can accelerate itself without human participation. It also invites the Terminator scenarios, and Elon also said around the same time, he says he sees an 80% base case that AI ushers in more than an industrial revolution size change, and we have an era of abundance and human prosperity that is beyond almost anyone's imagination.

In the very next breath, he said, not jokingly, "And there's also, unfortunately, an outlier case of 10, maybe even 20% probability of outcome, which is that it's an extinction event for the human race because it goes out of control." That's a pretty darn broad range of outcomes, Matt. Is that realistic, or is that Elon trying to foment some kind of emotional reaction on people?

Matt: ChatGPT make a better version of ChatGPT is already happening. I think there's been some research that's been put out this week about just how much of the

code inside these frontier models is actually written by other AI, and that's accelerating, so we're heading in that direction.

But I don't think we're gonna enter into an age of abundance as my token burn of four billion tokens in a day has proven to me that is not going to be free, and that has a real cost behind it. I did the math on if every person in the world burnt four billion tokens in a day how much energy we would need, and it came out at 30 terawatts, which by comparison with the world today is what?

Something like, is it 10X the world's energy production? Something like that. It's some... It's many orders of magnitude more than what is being built out at the moment for data centers. I think at the moment there's plans for like somewhere between 200 and 300 gigawatts of extra energy to power this gigantic data center explosion, which is going exponential.

But still, even if that gets built out, it's nowhere near what would be needed in terms of energy. So energy is the final boss that had the need to be solved, I think, for the singularity to occur, and I think we're a long way away from that. The other thing he's doing that we haven't really talked about is he's building this robot army.

So you've got this explosion now, the next big thing that venture capitalists are all financing, and you've had, Unitree's IPO just recently. Unitree, the really the leader in humanoid robotics, is they've actually managed to solve locomotion in humanoid robot form. Back in the day, what was it?

Stanford, I did a robotics course, and it's all about calculating which angles each of the servo motors need to be, or actuators need to be moved to in order to get from A to B reliably. It turns out they've just had explosions in advancements in that field, and not just can the robots walk now, but they can do kung fu they can carry a gun and raid a structure as part of a, an advance team in an army.

We just had the robot games. They can run faster than a human. They can jump higher than a human, etcetera. Superhuman performance. And these things are selling starting at \$6,000. I think Unitree is, what, doing nine billion a year of revenue and profitable already, and you're just gonna see this explosion in humanoid robots, in drones, which is the Ukraine war's been fueling, etcetera.

And I think in a couple of years, not too far away, maybe two years from now, these things will be pretty common out there in the streets Doing your Uber Eats

delivery or whatever it may be. But yeah, Elon's building his own robot army with his Optimus series. And who knows? May- maybe this whole Terminator vision where the robots are up a- and rising and running over the hill, Etc., towards you, they're all Unitary robots.

Erik: Matt, I could easily do an hour plus on robotics and AI and particularly industrial versus humanoid robotics and where all of that is headed, and so on and so forth. If there's listener interest, we'll have to do that next time because we've already gone well beyond our usual time budget. So of course as always, I love these conversations.

Before I let you go, tell us a little bit more about what you do at Freelancer, and particularly, I know that you guys have started to move not just to using AI yourselves, but organizing your AI-oriented freelancers. You've got project managers that will help you put together a team that will do AI for your company.

If I think this stuff is all interesting, and I think it's cool that you're doing all these things, but, look, I'm running a business where I don't have the expertise that knows as much as you do about AI. Can you guys help, and if so, how?

Matt: So at Freelancer, we have the world's largest collection of humans.

We've got 90 million people in the marketplace that can help you with any job you could possibly think of. These humans are effectively 20-watt inference av-engines with free pre-training. So in any skill set you can possibly think of, we have AAI, artificial intelligence, i.e., humans, who have access to all the AI in the world, so they've got access to every model you could possibly think of who can get any job done for you.

A big thing that's happening right now is going into companies and automating them with AI. So going in and plugging up, plugging all these databases and so forth and getting these agents to write reports for you on a regular basis or hooking up your telephone system to answer phone calls or take bookings or whatever it may be.

So they're in there doing all of that. Basically, we do \$10 jobs to \$10 million jobs. We work at the high end with large corporations and organizations. With NASA, we solve problems such as crack propagation in satellites or gene editing in the

central nervous system of humans, Etc., by effectively crowdsourcing our 90 million 20-watt inference engines at scale.

So yeah, any job you can possibly think of, you can get it done on Freelancer. We have all the humans and all the AI, so you get the best of both worlds. And the advantage as well of going to Freelancer is if you-- everyone's had that experience. You get in there with the AI and you sometimes can go in circles because AI sometimes can't get the job done properly and can't hill climb out of it effectively.

So in a way, and I wrote an essay on this, it's pay to pray. It's a slot machine. You pull the handle, and sometimes AI does a great job and you're amazed by it, and sometimes you pull the handle and you go in circles, and you're sitting there as a bit of a degenerate pulling a slot machine handle.

With Freelancer, you pay per outcome, not pay per token. So you put a price in, a budget, the freelancers will bid on it, and when you agree to the price, you pay that and you get your outcome with AI it's a bit non-deterministic. So come to Freelancer. We've got all the people with all the AI, and you get any job done you can think of, and the big thing right now is AI automation, AI agents.

Just go to freelancer.com and just type into the click on post a project, type in whatever you want. It's free to post a project, and no matter what you wanna get done, small or large, we'll do it for you.

Erik: I've been a customer several times myself, and I definitely vouch for the value of the service.

And one final point before we close, Matt, in the past, our listeners have become accustomed to you writing an entire missive, which we basically used as the outline for our past interviews. We had so much to talk about this time that we've actually done the missive separate from the interview. So in addition to everything we've discussed, what else can people find if they read your latest piece on Medium, which is titled...

Tell us both the title, where to find it, it's in your Research Roundup email is where you can find it, of course, and what's it about?

Matt: All this and more, it's called Agent Provocateur, and it covers the agentic explosion and the dynamics in the space and everything we've talked about today

and more. So just get it through your Research Roundup or follow me on Twitter or on Medium.

Erik: I can't thank you enough, Matt. And of course, boy, this means it's time for Patrick and Massil to come in, take this amazing, so much stuff to talk about, translate that to how we actually put trades on in the market. Patrick, you got a lot of work from here. Where's the trade?

Patrick: Thanks, Erik. Now, coming out of Matt Berry's interview, it's pretty obvious that this week's trade has to live somewhere inside the AI complex.

His entire thesis is about accelerating AI adoption. But having a strong structural thesis doesn't automatically mean it is a perfect tactical entry right now, and that's why I wanted to express the view through Nvidia options rather than the shares themselves. January two thousand and twenty-seven option implied volatilities have collapsed towards the lowest levels in the past year, which means the cost of owning optionality is about as cheap as we've seen over that period.

So rather than taking full delta one exposure in the stock, I'd rather use that relatively inexpensive volatility to buy defined risk upside convexity. With Nvidia trading around two hundred and twenty-four dollars and thirty-one cents at the time of this recording, I'm looking at the January 15th, two thousand and twenty-seven hundred and twenty-five dollar strike call, which is essentially at the money and gives us a hundred and twenty-eight days till the expiration.

The call is trading around twenty-one dollars and sixty cents, or just under ten percent of the value of the stock, with an implied volatility around thirty-eight percent. The attraction to expressing the trade this way is flexibility. If Nvidia corrects into the midterms, the call gives us the defined risk, downside delta compression, and positive vega if volatility spikes, potentially cushioning any drawdown versus owning the shares outright.

That preserves capital and most importantly, gives us the ability to tactically reposition at lower prices. Alternatively, if Nvidia breaks to new highs and the semiconductor bull market re-accelerates, we own open-ended upside convexity purchased when volatility was near its cheapest levels in the past year.

So the trade is simple. Cheap, long-dated optionality, defined downside, positive vega, and open-ended upside participation. That's where's the trade.

Erik: Patrick analyzes and trades the markets every day over at Big Picture Trading. Macro Voices listeners can sign up for a free two-week trial at bigpicturetrading.com.

Now back to Patrick and Massil

Massil: All right, Patrick, let's start with the big picture here. What's been driving markets over the past week, and how do you see them developing over the next few weeks?

Patrick: All right, so the simplest way to understand these markets is that oil has just been the catalyst, and the bond market now is the transmission mechanism, and now equities are absorbing all of this damage.

Now, oil just pushed towards a hundred dollars on WTI. We are seeing Brent clearing a hundred dollars on the upside, reviving inflation fears and forcing markets now to reconsider whether central banks around the world will have to continue to tighten. The front end is repricing all of the central bank policy.

You're seeing that right across the world, whether it's the US, whether it's Canada, or whether it's Europe. We've seen substantial declines in sulfur futures as they're now pricing much more hawkish central bank action. But the long end faces a much broader problem. There's heavy treasury supply, fiscal concerns, and rising term premium.

And of course, the big story everyone's talking about is the enormous corporate borrowing competing for that capital. Now, Bessent just announced a six billion dollar buyback, but that disappointed the market. Many were expecting ten billion dollars on that. And so we see pressure on these yields. The ten-year is now at four hundred and eighty-five basis points, and the thirty-year is back above five hundred and thirty basis points.

That risk-free hurdle rate is rising, pressuring these equity valuations. I wouldn't call this yet a, a broad credit stress. It hasn't translated into credit spreads in other areas like that. But treasuries are under pressure, and it is causing a lot of stress in asset prices. Now, equities are therefore caught in this tug-of-war.

Their earnings have been resilient so far. AI boom is actually finding some short-term support while energy stocks are rising, but these higher yields are

compressing valuations and causing the breadth to deteriorate. The next catalyst is that CPI on Friday, followed by the FOMC decision next week.

So there's lots of things that could either stabilize the conditions or make things more volatile going into next week. This is certainly a fragile moment in these markets.

Massil: All right, I wanna take a step back here and break things down one by one, but I wanna talk about energy markets because of the big moves we've seen in crude oil over the last few weeks.

We have been calling for higher prices on this show as positioning remained short for quite some time now. Just two months ago, back in July, gross short positioning in WTI reached a five-year extreme near two hundred and forty-two thousand contract. And despite some traders covering those contracts, significant short exposure still remains in this market today.

Now, with oil surging again this month, how much do you see this move being driven by the fundamentals, and how much is a crowded short trade being squeezed?

Patrick: The squeeze is underway. Literally as we're recording, we just printed a hundred on WTI. The geopolitical premium has been a physical supply story again.

This is not merely just a fear of a disruption, but the disruption's occurring. The global inventory buffer is dangerously thin. The EIA estimated global oil inventories have fallen roughly four hundred million barrels, and oil is ripping. But the fascinating part is large speculators, that hedge fund community, has not been rebuilding their speculative positioning, and shorts have stubbornly held onto their positions.

So the flows have clearly fuel to drive this. With this kind of a rip, this is certainly a disruptor in the markets, and it'll be very interesting to see whether oil sustains above a hundred or whether this fades quickly or into next week.

Massil: Now, when you put all the stuff you were talking about together, so rising oil prices, surging yields, and even renewed inflation risk, how vulnerable does that leave the equity market?

Are we setting up for a meaningful breakdown, or are there still enough bullish forces to keep the market together?

Patrick: This is such a fascinating topic because on the surface, the markets are absorbing the initial oil and rates pressures. But under the hood, there is a clear structural deterioration.

I'd like to e-think of it like a, an analogy of a castle built on sand, and the sand is shifting under the castle, but the castle hasn't yet moved. The puzzle to solve, is there a strong enough foundation to keep this market elevated, or is the deteriorating market breadth going to inevitably crack this market?

Now, the first thing to highlight is there's a huge deterioration in that breadth. A month ago, we were at seventy percent of S&P five hundred stocks bull trending above their fifty-day moving average. Since then, over the last month, we've seen breadth collapse to near thirty-five percent. That means fifty percent of those trending stocks all broke down So why no response in the cash index?

Because this was accompanied by a pop in a few of the mega cap heavyweights in the Mag Seven and semiconductor basket. Now sector-wise, financials have weakened to testing their fifty-day moving average, while clear breakdowns have occurred in consumer discretionary space, consumer staples, retailers, industrials, defense contractors, home builders, utilities, transports, REITs, the list goes on.

That's counterweighted so far against miners holding up pretty well. The energy stocks continue to trend strongly, and the biotechs, at least for now, are holding up. But the vulnerability really comes here in the next hundred S&P points. Looking at a lot of the quants, like reports from Charlie McElligott and others, that the CTA sell triggers are all around the seventy-five hundred to seventy-five fifty area, which is just fifty to a hundred S&P points below where we are.

Right now, the selling pressure is coming in a little bit, but the question becomes, do we trigger systematic selling flows that can often lead to a hundred billion-plus in, in flows pivoting? We've already seen a substantial pivot in those flows in the Russell, which has already rolled over. And so will the S&P join that party?

And more importantly, will we see things like the semiconductors be able to hold things up? If they suddenly start to fall back below their fifty-day moving averages, things could get pretty spicy pretty quickly.

Massil: Now let's hone in onto one of those sectors that were the winning one here over the last few weeks.

'Cause when you look at gold here and when you look at the metal, we've been on a two-year bull trend, but the last six months has been struggling to make further gains. Now, when looking at the COP report, we can see speculative long positioning is rebuilt, but it has not become heavily crowded. Do you see this just as a healthy consolidation within a broader uptrend, or will higher yields keep putting pressure on gold's price?

Patrick: Listen, let's step back and look at gold in the big picture here. First of all, we went through a two-year roaring bull market that ended in January 2026 with a blow-off up at near fifty-six hundred. Since then, over the last six months, we have seen a twenty-five percent correction in bullion and fulfilling what a typical market correction would look like.

Now, we had a substantial breakout here in the month of August, and n- now we're backfilling that. Now, from a headwind perspective, obviously we're seeing real rates rising, but the tailwind is obviously the dollar has weakened, and it's a, a situation where there is some positives, some negatives, and we're in middle of a consolidation.

Overall, I think gold has genuinely turned the corner. We haven't seen speculators really lean into this in a big way just yet will-- it will be very interesting to see whether this gets bought on dip. Overall, it could be a scenario where gold does some consolidation for even a few more months as we maybe grind towards the midterms.

But at some stage I'm leaning that the correction is over and that this bull-- gold correction should be bought on dip. And one thing I wanna move on to here though, Mass, is I wanted to just briefly touch on uranium, because what's interesting is the U308 is actually near ninety dollars now, continues to trend higher.

You're seeing general accumulation in uranium and... But what's interesting is that uranium miners broke out and seem to continue to correlate with gold miners. Maybe there's an equity basket that includes them together. Not sure w- how to size that up, but it's clearly been there. Overall, be very interesting to see whether

that correlation continues and whether uranium has also turned the corner here and started something more bullish.

Massil: Yeah, for sure. We'll keep watching. But let's turn to the dollar, and more importantly, I wanna talk about the yen because we've seen the pair break sharply higher over the past week from both administrations intervening. But the latest COT data still shows large speculators deeply net short this market.

So is this simply an intervention-driven squeeze, or could we be seeing the beginning of a much larger yen carry trade unwind going into year-end?

Patrick: Listen, the price action substantially has pivoted over the last month and a half. Now, first of all, it started with that intervention, but the entire correction that happened throughout the remainder of August simply retested the fifty-day moving average and was a typical fifty percent retracement, and that decisively broke out.

What's really interesting is in the latter part of August, as we saw in those COT reports, actually short sellers came roaring back in. And so when we get that COT report tomorrow, it'll be very interesting to see how many of them, once again, got squeezed out of this. What's significant for me is that we've now broken on that yen above their fifty-week moving average.

Why this is significant is that the yen basically for almost a year and a half has been in a decisive downtrend, and when we usually get a key breakout like this, it tends to follow through. So this is certainly high on my watch list to see whether quick dips in the yen are being bought and whether we sustain above moving averages and whether a new trend is being established.

The fact that the US dollar is generally trending down and weakening is a further support to the idea that the yen can be the c- the strongest currency in this basket.

Massil: All right, Patrick, let's turn to this week's positioning pulse 'cause I was looking at the whole COT report over the last few weeks here, and I saw that the most extreme positioning was in corn.

Large speculators in this market were sitting at the hundredth percentile on both the one-year and the three-year score, and that just means that we've not seen positioning more bullish than this over the last three years. So is this setup

dangerous crowding, or is the market simply pricing in a deeper tightening supply story?

Patrick: Listen This isn't just corn. Whether you take wheat, soybeans, sugar, bean oil, or bean meal, we have seen us reach the hundredth percentile on one-year and three-year positioning right across the board. When looking at corn specifically on these COT reports, what's interesting is going back a few months, the gross short positioning in corn was almost at a five-year extreme.

And in the last two months, almost all of those corn shorts have been squeezed. At the same time the longs have finally come in. We're at a five-year high in gross long positioning and net long positioning. And so this is genuinely demonstrating crowding. But crowding is not a contrarian trade right off the bat because there is genuine fundamentals behind this, right?

The agricultural complex is repricing supply risk. Y- there's a U- Ukraine logistics story in there all sorts of issues with weather, like the El Nino. The bottom line is that there's a genuine fundamental reason that these grains are ripping this way. So I'm not ready to short this crowded trade, but this is a very well-known and already very well-positioned long position in this space.

It'll be very interesting to see whether we continue to see the pattern of dips being bought and the bulls keep being rewarded for this long positioning. All

Massil: right. Thank you, Patrick, for this. And for any listeners that want to dig into these positioning charts, you can find it all for free at cotsignal.com.

That's cotsignal.com.

Patrick: All right. That does it for this week's Market Desk. Thanks for joining us, everybody.

Massil: Thank you, and see you next week.

Erik: And a reminder, as a Macro Voices listener, you're entitled to a two-week free trial of Big Picture Trading, where you can watch Patrick analyze and trade the markets live every single day at bigpicturetrading.com.

No credit card is required to sign up, and there's nothing to cancel. I'm Erik Townsend, and this is Macro Voices. We'll see you next week.

Announcer: Macro Voices is presented for informational and entertainment purposes only. The information presented on Macro Voices should not be construed as investment advice.

Always consult a licensed investment professional before making investment decisions. The views and opinions expressed on Macro Voices are those of the participants and do not necessarily reflect those of the show's hosts or sponsors. Macro Voices, its producers, sponsors, and hosts shall not be liable for losses resulting from investment decisions based on information or viewpoints presented on Macro Voices.

Monthly Membership Includes:

- Daily “Macro Outlook” video
- Watch Patrick analyze and demonstrate trades LIVE during “Where’s the Trade”
- Daily LIVE Q&A
- Weekly position summary
- Hours of FREE educational videos and bootcamps
- Access to members only LIVE webinars

SIGN UP NOW!

