Welcome to the OpenAI podcast, the podcast that opens up the world of AI in a quick and concise manner.
Tune in daily to hear the latest news and breakthroughs in the rapidly evolving world of artificial intelligence.
If you've been following the podcast for a while, you'll know that over the last six months I've been working on a stealth AI startup.
Of the hundreds of projects I've covered, this is the one that I believe has the greatest potential.
So today I'm excited to announce AIBOX.
AIBOX is a no-code AI app building platform paired with the App Store for AI that lets you monetize your AI tools.
The platform lets you build apps by linking together AI models like chatGPT, mid-journey and 11Labs, eventually will integrate with software like Gmail, Trello and Salesforce so you can use AI to automate every function in your organization.
To get notified when we launch and be one of the first to build on the platform, you can join the wait list at AIBOX.AI, the link is in the show notes.
We are currently raising a seed round of funding.
If you're an investor that is focused on disruptive tech, I'd love to tell you more about the platform.
You can reach out to me at jaden at AIBOX.AI, I'll leave that email in the show notes.
As AI systems become increasingly sophisticated, there is a need to handle them with care.
This is according to Anthropic, who is the AI safety company that's behind Claude's chatbot.
They've been positioning themselves as the safe or responsible AI from the very beginning.
I think this is great.
I think they saw a big boom when chatGPT came out and people noticed it had hallucinations.
It was not always perfect.
Then they're like, see, when you be responsible, when you be safe, it could lead to the end of the world.
I think right now they're making a big move to position themselves in this way again, which I think is really smart and great business move.
Kind of stepping into this, the company has unveiled a new policy that outlines its dedication to the responsible expansion of AI systems.
They're calling this the responsible scaling policy.
This framework is particularly tailored to address what Anthropic labels as quote unquote catastrophic risks.
Those type of risks represent situations where AI's action could directly instigate massive calamities, so imagining unsettling events leading to quote thousands of death or hundreds of billions of dollars in damage, end quote, right?
They're really looking at like the worst case scenario for AI and they're trying to build a framework to avoid those situations.
What is the distressing part?
These catastrophes would be unprecedented incidents that wouldn't have transpired without the AI's involvement.

That's kind of their framework for saying what they're really specifically trying to avoid.
In a new conversation that they recently had, Anthropics co-founder Sam McAldish talked a little bit about the motivation and intricacies of this new framework.
At its core, the policy introduces AI safety levels, which is a stratified risk system mirroring the US government's biosafety levels earmarked for biological research.
This four tiered ASL spectrum spans from ASL-O, and of course ASL is AI safety levels, but ASL-O indicating a low risk entity to ASL-3 marking a high risk system.
So McAldish pretty much said quote, there is always some level of arbitrariness in drawing boundaries but we wanted to roughly reflect different tiers of risks.
So he recognized that the AI models of today might not be immediate threats, so you know chatGPT isn't about to take over the world or release all of the world's nuclear missiles, but the future landscape might be very different.
So highlighting the policy's adaptability, he mentioned quote, it's not a static or comprehensive document, we envision it as a living entity ever evolving based on our experiences and the feedback we gather.
Anthropics kind of ambition is straightforward, but I think it's really interesting, essentially they want to kind of harness competitive dynamics to navigate crucial safety hurdles.
And this vision they're saying is going to ensure that the pursuit of safer AI paradigms doesn't just, is it not, you know, resulted in, like people are actually making these things safer.
So, and you know they're worried about like quote unquote aggressive scaling, which I'm not sure if that means their competitors going a lot faster than them, or they're saying you know there's some actual danger, but in any case, McCandlish was candid about the complexities in this whole mission saying quote, we can never be totally, totally sure we are catching everything, but we will certainly aim to.
This is really interesting, like they really are trying to position themselves as being like the responsible AI and the safe AI people.
So another feature in this whole RSPs, like framework they have, is its emphasis on independent oversight.
So no modifications to this policy can proceed without board approval.
And this might sound really cumbersome, but they believe that making a measure like that is really valuable, explaining quote, given our dual roles in rolling out models and also appraising them for safety, there exists a genuine concern, there's always a lurking temptation to perhaps be lenient on our tests, an outcome we ardently wish to sidestep.
This is interesting, right, they're putting up some internal governance and I do respect and appreciate that.
This move by Anthropic, a lot of people are saying couldn't be more timely.
So the AI domain obviously is, you know, has a hundred new companies rolling out very, very quickly and Anthropic with its roots kind of tracing back to some former members of OpenAI and of course, boosted by substantial investments from tech giants like Google.
This is, you know, when they talk about it, they're like, we have substantial investments from tech giants like Google, also substantial investments from Sam Bankman Freed, who famously

led FTX, which was a massive Ponzi scheme, but I mean, aside from that, it's not throwing shade at Anthropic, they just took money that they probably thought was good money at the time, but it's, you know, probably just a little blight on their history that they don't like to bring up when they talk about their, you know, substantial investors.
I think, you know, I think Sam Bankman Freed put in like over 500 million and Google put in like 300 million like much later.
So it's just funny that they only bring up substantial investments from Google.
In any case, Claude's chatbot, which is Anthropics, of course, first kind of ChatGPT competitor, I love it personally, it's great.
You can put like the context windows massive on it.
So if I have like, you know, 10 articles I need to throw in there and get like it to consolidate them all, it can do stuff like that, which ChatGPT obviously cannot.
So I do like Claude.
It actively deters harmful user prompts by looking for potential hazards.
I haven't run into this myself because I mostly just use it for summarizing long articles, complex topics, complex AI stuff to, you know, get good kind of bullet points for this podcast, essentially.
So I've used it for that.
I haven't had them kind of ban me from asking about a specific topic.
Like I haven't asked them about making, you know, disposing radioactive waste or something.
But in any case, the capability stem from Anthropics, quote unquote, constitutional AI strategy.
Now for those that don't know, I would look up constitutional AI, I do actually agree with that strategy.
Essentially what it's saying is a lot of times there's not a lot of transparency in AI models.
Like we don't know why open AI is, some people call it censoring.
Some people call it guardrails of different topics, right?
Like you ask it to do something and it's like, sorry, I can't do that.
I'm more sorry.
This is not a good topic or I don't know, Chajapiti has their own guardrails, right?
They're a trust and safety team.
Now those might not all be bad.
There's definitely people that will argue whether some of them are good or bad and some biases in it and whatnot.
But the problem I have with it is that they're not transparent.
So you don't know what the guardrails are specifically.
And so constitutional AI is a model that Claude has adapted, where essentially the AI model has a quote unquote constitution that every response has to look at to make sure it follows a set of guidelines or ideologies.
Now you might not agree with a specific AI model's constitution, but I think it's really important that it's transparent, right?
I'm thrilled to use an AI model that has a constitution.
It might not be exactly one that I love.
Maybe there's one or two points on there I don't like, but at least I know where it's

coming from and it's transparent.
So I really do appreciate that transparency from different models because I hate using something and feeling like maybe the response it gave me has some sort of bias, not just because of the data, but that the creators or the developers who probably live in San Francisco and have a bunch of ideologies that some of them I agree with and some of them I don't may have imposed or injected into this thing and its responses are reflective of that.
But I don't know, right?
Kind of algorithmic serving me up of information and me not having transparency.
I think everyone wants more transparency and everything.
So long story short, constitutional AI, good move by Anthropic.
Now I think a lot of these mythologies employ chain of thought reasoning which in turn amplifies the transparency and efficiency of AI's decision from a human perspective.
With fewer human labels, I think this also paves the way for modeling AI systems that are both ethical and secure.
I really do think it pushes it in that direction which is good.
So right now the emergence of this RSP coupled with research into constitutional AI, I think underpins Anthropic's move and commitment to AI safety, this is the direction that they're looking at.
Now I'm going to put a caveat on this whole thing and say some of the downsides to this, but I really do think they probably have good intentions here and by kind of trying to trim down risks while amplifying benefits, Anthropic is essentially trying to, you know, has a set of commendable benchmarks for AI's future trajectory.
That's their goal, right?
Now the one thing I will say that did strike me as being slightly, you know, a little bit of a red flag, a little bit of alarm bells, is when they talked about the fact that, you know, today's AI models might not be immediate threats, but that the future landslide might look very different.
For some reason, I just got this like my spidey senses were tingling and it kind of made me think of the fact that, you know, they're doing this internal governance right now.
Very cool.
They have like different tiers of like what risks they think a new model will come up with.
Very cool.
All of a sudden I kind of started thinking about the fact that governments will probably in all the regulations want to adopt a similar, probably a similar structure to this where your AI model has to get like evaluated and if they, and there's going to be governments all around the world, I don't think this is controversial to say that the Chinese government is going to approve different AI models than America.
I mean, this is already happening, but it's just, I don't know, it feels, sometimes it feels a little dystopian in America to think that in the future there's going to be these like bootlegged models with no safety rails on them that go one way or the other and, you know, the government decides that they're like highly dangerous.
It's going to be like interesting.

It's like, I wonder if you could get like in trouble for having a USB thumb drive with like some sort of bootleg AI model that's super dangerous on it.
Of course, there's like hackers that are going to have bad stuff that is probably bad doing bad things, but it'll be interesting, right?
Like a totally no, like the equivalent of chat GPT, it's going to be completely open source.
The government's going to try to shut it down for X, Y, and Z reason.
Maybe it's a Republican in office.
Maybe it's a Democrat in office.
I don't know, right?
Like people of course are going to have their own reasons why they think this thing is going to be bad.
It's going to be the bureaucracy of people that we are essentially have not elected as well, right?
So all of the different federal agencies are going to want to get involved.
These are bureaucrats that no one elected, but they're going to want to try to impose their will on AI.
I don't know.
It's just a, it's just a fun little dystopian future that I think inevitably we're going to have to face.
But I just wonder if there's going to be this like guy running around with a USB thumb drive and it's like the equivalent of having a, you know, like an automatic weapon today that's like, of course, if you have an automatic weapon in America, those are illegal without the licenses and you go to jail.
It's like you have this thumb drive with this highly dangerous AI model on it and they like the FBI busts into your house and like, you know, arrests you pulls you out because of because of that.
Anyways, it's going to be a fun future.
But yeah.
So that's the only thing that is my, that I'm a little bit skeptical about is like when talking about these regulations, oh, and the whole reason I bring this up is, is it really the government that feels threatened by maybe an AI model that has the truth about X, Y and Z topic that I want you to know, or is it that right now, you know, open AI is the number one lobbyist that's helping to build these laws.
Are they going to build laws that say, you know, different AI models are not safe for X, Y and Z reasons and they are.
And so they, you know, penalize their competitors.
That's really what I'd be concerned about.
I know I, I paint a fun dystopian future about where the government's going to chase us down for our thumb drives and AI models, but really in reality, it's probably going to be someone like Chad GBT and other big AI companies that are building regulations, building a regulatory moat right now.
That's really what their vision is with regulation.

I don't think they're worried about how they're using AI.
I think they're just worried that maybe their competitors can catch up and they can build a regulatory moat.
So in any case, it'll be interesting to see where this goes.
I do think the anthropic probably has, you know, pure intentions on this right now, but it'll be interesting to see how this plays out in the future.
If you are looking for an innovative and creative community of people using Chad GPT, you need to join our Chad GPT creators community.
I'll drop a link in the description to this podcast.
We'd love to see you there where we share tips and tricks of what is working in Chad GPT.
It's a lot easier than a podcast as you can see screenshots, you can share and comment on things that are currently working.
So if this sounds interesting to you, check out the link in the comment.
We'd love to have you in the community.
Thanks for joining me on the open AI podcast.
It would mean the world to me if you would rate this podcast wherever you listen to your podcasts and I'll see you tomorrow.