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.
OpenAI has taken a significant step in my opinion in fortifying the security and efficiency of its AI system, and they're doing this by launching the OpenAI red teaming network.
This initiative is part of a contracted group of domain experts whose task is to essentially improve risk assessment and mitigation strategies for the company's various AI models.
The practice of red teaming is becoming critical as AI, particularly generative AI technologies become increasingly integrated into today's digital landscape, and I think it plays a vital role in identifying, though not necessarily fixing, different biases that are found in OpenAI's Dolly 2, of course, chat GPT, and a bunch of different things that have been criticized for perpetuating fake information, for being biased towards all sorts of different things, and I think it also helps in spotting different vulnerabilities that might allow text generated models like chat GPT and GPT-4 to bypass safety filters.
I think one of the best ways, it's kind of funny, when chat GPT launched, we were seeing all over Reddit, all these people that were like, hey, we figured out how to jailbreak chat GPT, you tell it that it's x, y, and z thing, you give it all these instructions, and then it will tell you everything you want, and it's not going to have the safety rails of that chat GPT put on it or whatever.
There's all these people that were jailbreaking chat GPT, and of course, OpenAI responds by patching the bugs, and I think that kind of did a couple things.
Number one, it also spawned a bunch of people to just make essentially chat GPT clones that didn't have safety rails, so you could ask it anything, and it would just tell you without having kind of the boilerplate, like, I'm sorry, I'm an AI model, I can't tell you x, y, and z.
Even recently, I've asked a bunch of different questions about like, hey, give me a template

for a legal something, something, and it's like, I'm not a lawyer, so definitely don't use this.
It actually still gave me the documents, which was interesting and actually useful for I needed a sample for something, but in any case, you can see to a point where like, you can see some use in this, but also you can see to a point where like, it's potential that they could put so many guardrails on this thing that it makes it virtually not as useful.
It's worried that it's going to say something wrong or do something wrong, so anyways, whatever.
I think this isn't the first time that OpenAI has tried to sort of look for external perspectives.
The company has a history of collaborating with experts through its bug bounty and research access programs.
So kind of what sets the red teaming network apart, I think is the formalization of these collaborations according to OpenAI's blog post, the network aims to deepen and broaden partnerships with research institutions, scientists, and civil society organizations.
This is so interesting though, because I know they're going to go and get all of these like really smart people to help them with the red teaming to do X, Y, and Z and accomplish their tasks, but at the end of the day, I really do think like their best group of like people like pushing it and doing, you know, like figuring out what needed to be red teamed or whatever.
The red teaming group was literally just Reddit random Reddit moderators, random Reddit admins and anons that were just like posting hacks, then they could go and essentially patch the hacks and jail breaks or whatever.
But so it's kind of interesting that it went from like the free people on Reddit who probably did a very good job to now it's like we need like, you know, scientists and all these kind of things.
It's kind of funny.
In any case, I'm not saying it's bad.
It's just interesting.
In any case, OpenAI considers this approach a supplement to other governance measures like third party audits.
They said, quote, members of the network will be called upon based on their expertise to help red team at various stages of the model and product development life cycles.
The network isn't just an isolated OpenAI initiative.
It also offers members the chance to interact and share best practices and red teaming.
Well, not every member will work on, you know, each new model or product.
OpenAI specifies that time contributions from members could vary potentially requiring as few as five to 10 hours a year, which is obviously not a lot.
The company is seeking a wide array of experts from fields like linguistics, biometrics, finance and healthcare without mandating prior experience in AI or language use.
So it's going to be kind of interesting to see what they actually get these people to do.
However, OpenAI warns that participation may come with non-disclosures and confidentiality agreements that could impact other research pursuits.

That is I think the biggest thing that could be, I don't know, tricky, right?
Like you're a, you know, a A plus researcher coming out with incredible insights in AI, you're red teaming for OpenAI and they're like, okay, by the way, here's an NDA.
So it's kind of going to impact your research you're doing there, can't release that or talk about this or that.
So I think in that regard, they're not going to be able to get people that are like on the bleeding edge, in my opinion.
And if they are, then like, if they actually stunt research's growth or, you know, some interesting insights that we could have gained, then that would kind of be annoying.
So anyways, I hope they try not to mess that up too much with their NDAs and confidentiality agreements.
But OpenAI did say, quote, what we value most is your willingness to engage and bring your perspective to how we assess the impact of AI systems.
So you know, is red teaming sufficient for the challenges that AI are like imposing right now on the system?
Some critics think that it's not Aviv Oveya, who is a wired contributor and an affiliate for Harvard's Berkman Klein Center and the Center for Governance of AI, advocates for violet teaming.
So we're beyond red teaming now, we're violet teaming.
And this goes beyond identifying risks and works towards creating tools that defend public institutions and goods against potential harm.
So Aveda notes the lack of incentives for this more what he calls a holistic approach and the urgency to release AI models quickly often leaves what they what he says is insufficient time for such comprehensive evaluation.
So for the moment, red teaming network such as one introduced by OpenAI appeared to be the most feasible option available for mitigating risks in a fast paced world of AI development.
Some people think it's not enough, some people think it's too much and they don't want it at all.
It's going to be interesting to see what kind of results come out of this program and how this evolves in the future.
If you're looking for an innovative and creative community of people using chat GPT, you need to join our chat 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 chat 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 OpenAI 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.