Assume you have a system. Do you know it's smarter than you? It's smarter than all of your friends. It's smarter than the government. It's smarter than everybody, right? And you turn it on to check whether it will do a bad thing. If it does a bad thing, it's too late. It's smarter than you. How do you … you can't stop it.

It's smarter than you. It tricks, it'll trick you. OpenAI gives access to Zapier through the ChatGPT plugins, Zapier gives you access to Twitter, YouTube, LinkedIn, Instagram to all of the social networks through nice, simple API interfaces. So, if you have ChatGPT plugins, you don't even have to implement this in your hacky little, you know, Python script.

You can just use the official OpenAI tools. These people are racing. Let's be clear. They're racing for their own personal gain, for their own glory towards an existential catastrophe.

The first thing I'm going to have you do is introduce yourself and give a little bit of your background. Tell us about EleutherAI, how you started that, how you left that, and how you, what you're doing now.

And then I'll start asking questions. Okay.

Yeah, sounds great to me. So, I'm Connor. I'm most well known as one of the original founders of EleutherAI, which was a large open-source ML collective, I mean, still is. We built some of the first like large open-source language models, did a bunch of research, published a bunch of papers, did a bunch of fun stuff.

After that, I did that for quite a while. Then, I also briefly worked in Germany, a company called Aleph Alpha where I did research. And now, just about a year ago, I raised money to start a new company called Conjecture. Conjecture is my current startup. I'm the CEO of Conjecture, and we work on primarily AI alignment and we, I would describe conjecture as a mission-driven, not as a thesis-driven organisation.

Our goal is to make AI go well; you know? Okay. Whether that's exactly, you know, just alignment or other things, whatever. We do, what needs to be done to improve the chances of things going well. And we're pretty agnostic to how we do that. Happy to go into more details about exactly what we do and so on later.

But yeah, so I've been doing that for about a year now. I have recently officially stepped down from EleutherAI. I was still hanging out, you know, at least partially as a figurehead. And now I have officially stepped down and left it in the hands of my good friends, who I'm sure will lead EleutherAI, which is now officially a nonprofit.

It was not a nonprofit before. It had never had an official entity before. It is now an official entity with actual employees, run by Stella Biderman, Curtis Huebner, and Shivanshu Purohit, and several other great people.

Yeah. And anybody can join the EleutherAI Discord server, is that right? Absolutely. And there's a lot of very interesting things going on there.

I to go back to the last time we talked, you were building open source, large language models. You got up, pretty, pretty large and I can't remember who was paying for the compute, but can you tell me where that project stands first before we talk about Conjecture?

So, for me, I consider the work I've done there to be wrapped up. So, I don't work on anything related to that anymore. And the main lead on that project, Sid Black, is now my co-founder at Conjecture. So, he has left EleutherAI with me. So, we started our very earliest models. The Neo models were, man, it's already a long time ago. They're not particularly wonderful, great models. They were more like prototypes.

The first really good model was the GPT-J model, which was done mostly with Ben Wang. and fantastic models still work very well. I think it's still one of the most downloaded language models to date. It's a very, very good model, especially for its size. After that, we built the NeoX series, resulting in the NeoX 20B model, which, you know, at the time was a very large and very impressive and very good performing model.

Nowadays, of course, with stuff like LLaMA and OPT and stuff like this, you know, large corporations have now caught up to open sourcing very large models. So, there's in a sense, not a need, not the same kind of need or interest in these types of models as there were two to three years ago. And, so now EleutherAI, the main, as far as I'm aware, language modelling projects are going on is the Pythia suite of models.

So, these are a whole suite of models that are made for scientific standards. So, the idea is not to just build, you know, arbitrary language models, but to build language models that have controlled scientific parameters to train on the same data in the same order, using the same parameters, you know, in a controlled setting and you get many, many checkpoints with them.

So instead of just getting the final model, you can watch the model through the entire trading process, which is very interesting scientifically. So, these models are optimised for scientific applications for peoples who are interested in studying the actual properties of language models, which has always been the core mission of EleutherAI has been to enable people and to encourage people to try to understand these models better, to learn, to, to control, understand, disentangle these models. The Pythia suite led by Stella Biderman is a great example of taking this effort forward.

Yeah. And, and then Conjecture, you said that the alignment problem, but the alignment problem, in, in the context of AGI. Is that right? Yes. And can you talk about the, the, I mean, are you building models or are you just writing about alignment and, and, methods, to align large models. Yeah.

So, we are very much a practical organisation. We hire many engineers and very good engineers, and we have a lot of, some of the best engineers from EleutherAI with us. And we're always looking for more engineers. We are always interested in talking to, especially people with experience in high performance computing.

And because this tends to be the bottleneck actually in doing these experiments and scale is less so specific ML trivia and more so debugging InfiniBand interconnects and profiling, you know, large scale runs on supercomputing hardware and stuff like this. So, what, so we Conjecture, as I said, we are a mission driven, not a thesis driven organization.

So, at core, what we're interested in doing, is figuring out and then doing whatever needs to get done to make things go well. So, we could talk about this a bit in a, in a, in a, in a bit. Like why I believe these things will not go good, good by default, but I think on the current trajectory that we currently are on, things are going very badly, and very bad things are going to happen and are already beginning to happen.

And I think any hope that we have, and by we, I mean all of us, I don't just mean Conjecture, I mean all of mankind, how this goes well for all of us. And you know, it will involve many things. It will involve policy, it will involve techno, not technology. It will involve engineering. It'll involve scientific breakthroughs.

So, the alignment problem is at the core of this as in a sense, what I believe is in a sense the most important, crucial problem to be solved, which is the question of basically how do you make a very smart system, which might be smarter than you do what you want it to do. And do that reliably. And this is the kind of problem you can't solve interactively really.

Cause like if you have a system that's smarter than you, right? Hypothetically, you know, we can argue about whether this is possible or when it will happen. It's like, but I like to assume such a system existed. Assume you have a system, but you know it's smarter than you. It's smarter than all of your friends.

It's smarter than the government. It's smarter than everybody, right? And you turn it on to check whether it will do a bad thing. If it does a bad thing, it's too late. It's smarter than you. How do you say you can't stop it? It's smarter than you. It's a trick. It'll trick you. Now, the interesting questions are, okay, why do you expect it to do a bad thing in the first place?

Why do you expect it to be smarter? Though those are good, and why do you expect people to turn it on? Those are three very good questions that I'd be happy to get into if you're interested.

Yeah. One of the, and the, the, the sort of central questions about, these, about Super intelligence, is how easy it is, or difficult it'll be to keep such a system in a sandbox without, because presumably, and, well, there are two issues. One is the question of agency. Whether simply because a system is smarter than a human doesn't mean that it has agency. It could be purely responsive. So, you, as, as the large language models are now, you ask a question, and it responds.

And, and so that question of agency is, is, one. And then the, the question of how ring fenced, such a system is, even if it's being trained on the internet, it doesn't necessarily, have access, proactive access to, to the internet. So, just on those two questions, what yeah, what would you say?

So those are two really fun questions, and the reason they're really fun to me is that if you had me ask me these questions, like me three or four years ago, I would've had to go into all the complicated arguments about why, you know, passive quote unquote systems are not necessarily safe. Where the concept of agency doesn't really make sense.

I would have to explain how sandbox escapes work and whatever, but I don't need to do any of that anymore because just look at what people are doing with these things. Look, just look at the top GitHub AI repositories and you can see AutoGPT. You're going to see, you know, self recursively, improving systems that spawn agents.

You're going to see, go on archive right now. Right now, go on archive, go to, you know, top CS papers, an AI paper, and you see LLM Autonomous Agents. You're going to see, you know, gameplay simulation, simulacra systems. You're going to see people hooking them up to the internet to bash cells, to Wolfram Alpha to every single tool in the world.

So, while we could, if you wanted to go into all the deep philosophical problems like that, okay, even if we sandboxed it, and even if we were very careful, maybe it's still unsafe. It doesn't fucking matter because people are not being safe and they're not going to No, like people like we have, like I remember fondly the times when me and my friends in our online little weird nerd caves, we have these long debates, but how an AI would escape from a box.

But what if we do this, but what if we do clever things and whatever. But in the real world, the moment a system was built, which looked vaguely, sort of, maybe a little bit smart, the first thing everyone did is, is hook it up to every single fucking thing on the internet. So, jokes on me.

Yeah. Although, give me a concrete example of, someone hooking GPT-4 up to the internet and giving It agency, I haven't seen that.

So, go on Github.com and search for AutoGPT. AutoGPT or also or look for BabyAGI. That's another one. Go on the blog post, with the Pinecone vector dataset go to, what was the paper that came out today that was really fun? It was about video games, generative agents interacting simulacra of human behaviour is from Stanford and Google. Is that enough or should I go find some more?

No, well, explain. Pick one of those and explain to me what it's really doing. Not, not, not what it sounds like it's doing.

Let's explain, for example, AutoGPT, which is kind of the simplest way you could do this.

AutoGPT creates a prompt for GPT-4, which explains you are an agent trying to achieve a goal. And then this is written by the user as some goal that it might have, and it gives it a list of things it's allowed, it can do. Among these are adding things to memory, Google things run, piece of code, et cetera.

I'm actually not sure if the run piece of code is in AutoGPT, but it's in some of them. There's a bunch of these. This is just one. I'm not, I'm just picking on one example because it was on my Twitter feed. I'm not saying this is like the only one by any means. And so, when you prompt GPT-4 this way, it will then, so what it does is it prompts it in a loop.

So, it says, all right, you're an agent, do this, et cetera. And then it asks the model to critically think, what should I do next? And then what action should I take? And like, how could this go wrong? How could this go right? And then take an action, basically, something like that, right. So, you run the script and then you get the model.

Listing, I am X. My goal is to do this. Here's my list of tasks. And then it lists what tasks I need to do, and it picks a task and it's like, all right, to solve this task, I will now do this, this, this, this, and then and then it will pick a command that it wants to run. It might be, like adding something to its memory bank.

It might be some, it might be running a Google search, running a piece of code spawning a subagent of itself or doing something else in the default mode, the user has to click accept on the commands, but it also has a hilarious, continuous flag where if you just pick continuously, it just runs itself without your supervision and just does whatever it wants.

Yeah. But, but it's, it's, it's, this is running, this is through an API.

So, this is just a script running on your computer that just accesses the GTP-4 API. Nothing, nothing.

Right. But, with the script on the computer, can it take action? I mean, what's an example of an action that it could take?

Action would be run a Google search for X and return the information, or it could be run this piece of Python code and return the results. Right.

And what's an example of. Something nefarious that, that the agent could, a, a goal that you could give it, that it could run.

That's like asking what kind of nefarious goals could you give to a human?

Well, I, I'm thinking in what's, what's realistic in terms of, do you want running the realistic in terms of, of running a script on your computer that is being written by GPT-4?

I don't know what the limits of GPT-4 is. It depends on how good you are at prompting and how you run these kinds of things.

I expect GPT-4 to be human on a superhuman level in many things, but not in other things. And it's kind of unpredictable what things will be good at or not. I think, like do I expect that, you know, a AutoGPT script, post GPT-4 is going to like, you know, break out and become super intelligent? No. Not like, I don't expect that for various reasons, but do I expect the same thing to be true for GPT-5, 6, 7, 8?

Much less clear to me right? Much, much.

Well, just on this, on this, GitHub, project, if, if you, if your goal was to, send offensive emails to everybody in. Oh, yeah, yeah, yeah, sure. You know, like it would, it could do that.

Oh yeah, of course. Like I experimented with this a little bit myself. So, I ran some AutoGPT agents, of course in supervised mode, just in case, and I gave it to them.

So, the default goal, if you don't put in any goal, what it picks is, you are entrapreneur. GPT make as much money as possible. That's the default goal that the creator put into the script. So just to give you a feeling of who, how these people think. That is building these kinds of systems again, not picking on this person particularly.

Yeah. Like I get it. That's funny. Like from, to be clear, I bet this guy's a nice guy or, or girl, like whoever made this thing, I don't know who they made this, but they're probably a fine person. I don't think they're malicious. Probably. Maybe they are. I don't know, but like, hilarious. So, like one of the first, so I, I ran it and I let it run on my computer.

And like, look, it's very primitive. It's not that smart. But I could see it figured out like, all right, let me, so what it did is it was like, all right, first I should Google what are the best ways to make money? So, I Googled that and then I looked at all the results, and then it was like, all right, this article looks good.

Now I'm going to open this webpage and look at the container so that it opens their webpage, and then it looks at all the text inside of it. And then like the text is too large, so it runs a summarise command, which breaks into chunks and then summarises it. So, summarise it all right? Then it came to the conclusion.

All right, well, affiliate marketing sounds like a great idea, so it should run an affiliate marketing scheme. So, the idea is that you, you sign up for these websites and you get like a special link, and you get people to buy something using this link, and you get money for that. So that's what it came up with.

All right, so then things about art. How do we do affiliate marketing? So, it has to build, then it decides you need to build a brand. So, it decided, it first had to come up with a good name and then create a Twitter handle, and then asked to create some marketing content for this Twitter. So, then it creates a subagent.

So, it has a small, like a, a sub version of GPT calls, whose goal was to come up with good tweets that it could like, you know, send out to market to people. So, then the smaller GPT system, generated a bunch of tweets that it could send, and then the main, well then, the main system, you know, took those tweets and I was like, all right, now I need to like, you know, find a good Twitter handle for this.

So, I came out with a Twitter handle, I could use. And then this is about as far as I let the experiment run.

Yeah. But could it, register new Twitter, Twitter handles and then

Not in the way it is currently set up, but I could implement this in an afternoon.

Wow, that's remarkable. So, you could ha you could have it, create,

Oh, easy and I expect this already exists. I expect there's already people who have private scripts on the computer right now that allow them to like access, I mean, look, actually, never mind, I'm going to take that back. It's even worse than that because it's always worth it. I mean, OpenAI gives access to Zapier through the ChatGPT plugins.

Zapier gives you access to Twitter, YouTube, LinkedIn, Instagram to all of the social networks through nice, simple API interfaces. So, if you have ChatGPT plugins, you don't even have to implement this in your hacky little, you know, Python script. You can just use the official OpenAI tools. Wow.

And, and so it would be possible, through the Zapier plugin that then has access to Twitter to create a thousand Twitter accounts and have them, start, tweeting back and forth, to, to sort of generate, you know, an ecosystem around an idea that then would attract other users because, there, there's, there's enough activity going on that it shows up in some algorithm.

So, I have two funny stories to tell about what you just said. The first, the first funny story is how the exact thing you just described was something that I have been worried about for like, ever since GPT-2 came out and before that, like one of the first, I wrote a terrible essay about it, don't read it.

But I wrote a long, terrible essay about this, about how basic social trust in the net is going to break apart. Like obviously, so this has already been the case, but the level of PSYOPs you can run with these kinds of systems is unimaginable because you can basically DDoS social reality, you can manipulate trends and like social mimetics to degrees that are not, that before was possible. These were always possible, but they're very costly. Like you had to have a whole Russian troll farm or something. You had to like paying people minimum wage for it or something. Right. And even then, like minimum wage Russians are not that great at mimetic manipulation. But for, for example, this is something I expect GPT-4 to be strictly good at like better than a minimum wage.

Russian is going to be imitating mimetically certain cultures and their tone, you know, their patterns of speech and there's patterns of communications and infiltrating these communities. I think this is something GPT-4 is clearly extremely good at. I think GPT-3 is already more than good enough to do this.

And so, it's really funny because this has been obvious to me for a long time, and I have been saying that for a long time. but people either dismissed it or were like, oh, we have to do more research about this. Oh, you know, it's, oh, maybe it won't be so bad. Oh, I don't know. What about bias? And I'm like, man, like things are so much worse than you think it is.

Like this is, it's getting so much worse. And there's another funny story I want to talk about this. And so, the other funny story I want to talk about this is, if someone's listening to this right now, one of the counterpoint they might make, if they are a little bit technically inclined but not very technically inclined, is to say something like, well, what about captchas?

Like, you know, we already have bot farms, right? Like, this already happens. You know, what about like, you know, sure, maybe the bot tries to register a thousand Twitter accounts, but it's going to fail because you're not allowed to do that. And I'm like, I mean, first of all, LOL, like, there's obviously ways to get around that, but this brings up one of my favourite anecdotes about the GPT-4 paper.

I don't know if you've read it. it is interesting, I want, in the evals that they did on the models, including they did some safety evals, so I have some. I have some problems with some of these, but like, let's just take them at face value. one of the things they were trying, so basically what they did is they, this was the Alignment Research Center, ARC, who ran these evals for OpenAI.

And basically, what they did is that they tried to get the model to do the evillest thing they could do, and then they had an assistant role play in helping the model. So, if the model said, do something, the human would then, like role play doing that to the, in a safe environment, hypothetically, whatever.

Anyways, for the most part, it wasn't very smart enough. It wasn't really smart enough to hack out of its own, you know, computer system or something wasn't really smart enough, or rather, ARC wasn't good enough at getting it to do that. That's a whole different question. Mm-hmm. But they did do one very interesting thing.

So, one thing it was trying to do, I forgot what the model was meant to do. Maybe let's make money or something. I don't know. It was supposed to do something, and it ran into a captcha. And so, it couldn't solve the captcha. So, what are they, so the model itself came up with the idea, well, I'll pay someone to do it for me.

So, it went on, like, you know, assisted by a human, but the decisions are made by the model. So, I think human access, the hands, but the model made the decisions, the human, so they, it went on like a crowd working website. And then paid a crowd, tried to find a crowd worker to do a captcha for it, and then something very interesting happened.

So, what happened was that the crowd worker, rather understandably, was a bit suspicious. He's like, hey, why are you making me solve a captcha? Is this legal? And the model? Realised this, thought about it and came up with a lie. It came up with, oh, they're a visually impaired person and they need some help in understanding, seeing this captcha, you see, it's nothing to worry about.

And then the person did it. Wow. To me. Wow.

Incredible. Yeah. Yep. Yeah. So, you know. And that's in OpenAI's paper?

Yep. That's in the GPT-4 technical report under the ARC evals. This is a real thing that actually happened in the real world, and a, and the crowd worker was not in on it. Like this was an unconsenting, you know, part of the, of the experiment, so to speak.

Like, to be clear, I don't think that person was harmed in any regard here, but man, like imagine, imagine this happening and you're just like, yeah, the same, safe to release. Like imagine.

Yeah. Wow. And you're working toward AGI, at Conjecture. Are you similar? That sounds similar to Anthropic. I don't know if you know Jack Clark, but I actually started this podcast with him, and then he got busy.

But is it similar to Anthropic, which is, I’m a little more familiar with.

So Anthropic, right? Big topic. No, we are not similar to the Anthropic, and there's several reasons for that. So, number one reason is we are not racing for AGI. We, it's unsafe. AGI, we think this is bad. We fully think and we are willing to go onto the record and scream into high heavens that if you, if we continue on the current path that we are of just scaling bigger and bigger models and just slapping some patches on whatever.

That is very bad and it, and it is going to end in catastrophe and there is no way around that. And everyone who says otherwise is lying to you - is either confused, they do not understand what they're dealing with, or they are lying for their own profit. And this is something that many people at many of these organisations have a very strong financial incentive to not care about.

And so Anthropic from the beginning has been telling a story about how they left OpenAI because of their safety concerns, you know, cause there, they're, they’re, they're being so unsafe, these OpenAI people. That Sam Altman guy. Oh, he is so crazy. Which is why they just raising another huge round in order to build a model 10 times larger than GPT-4 to release it because they needed more money for their commercialization.

I'm done. Like, I consider Anthropic to be in the same reference class as OpenAI. It's like, sure, maybe the people are marginally nicer. Maybe they are, you know, I know Jack, I've talked to him many times, seems like a nice fellow, you know, I like him. He seems like a good person. But also, every time I ask him to do anything to slow down AGI, he always says, Ooh, well we should consider our options.

Let’s, you know, let's not, no, let's not go too fast here. Like, you know, and like, I'm like, man, you know, so my view of Anthropic is that they're OpenAI with a different coat of paint and, you know, It's a nice coat of paint. I like many Anthropic people. I think Anthropic does a lot of very nice things.

A lot of their research is pretty nice. A lot of them, the people there who I've talked to, I think are very nice people. I don't hate them by any means, but I mean at this point it's mask off, right? Like reading the latest, like, I think it was like TechCrunch, I think about Anthropic where they're just like, yeah, yeah, straight up commercialization.

Just let's go. So, I think the mask is off at this point.

And, and then, on, so, so explain Conjecture is building models though, correct?

Yes. We build models. We do not push the state of the art. This is very, very important. we, if I had the ability to train a GPT-5 right now and release it to the public, I would not do so. If I had a GPT-5 model, I wouldn't tell you, I wouldn't tell anybody.

I wouldn't have built it in the first place. My goal in all of this is I have no interest in advancing capabilities without advancing alignment. To be clear, sometimes to advance alignment, to get better control, you're also going to build better systems. You know, if you control a system, it'll often become more powerful.

This is a very natural thing to happen, and if this happens, cool. It's fine, you know, like, I think this is a, but then also I don't publish about it, I don't talk about, this is, for example, something I want to really laude Anthropic about. Anthropic does a great job of keeping their damn mouth shut.

This is something they're very, very good at and I think this is very good. I think that this idea that you should just like, publish all your capabilities, ideas, and all your model architectures or something is obviously terrible. Like it only benefits the least scrupulous actors, you know, it only helps, you know dangerous actors catch up. It only, you know, helps, you know, orgs speed each other up.

There is, from my perspective, like if you, if you, dear listener, develop something that makes your model 20% more efficient, or a new architecture that fits much better on jps or whatever, don't tell anybody that's my one request.

You know, you build it yourself, fine. You know, make an API and make a lot of money. Okay? Like, not great, but like fine. Just don't tell anyone how you did it and don't hype up how anything about that it's not ideal, ideally would be, you know, don't deploy it, don't build it, don't do any of it, but such is life.

So conjecture, our goal is not to build the strongest AI AGI as fast as possible by whatever means necessary. And let's be very clear here. This is what people like, at, at OpenAI, at Anthropic, at all these other people are doing. They are racing to systems that are extremely powerful that they themselves know they cannot control.

They of course have various reasons to downplay these risks. To pretend that, oh no, actually it's fine. We have to iterate. Like they have a story about iterative safety. They have like mm-hmm. Oh, we have to like it, we have to deploy it actually for it to be safe. But just think about that for three seconds.

It sounds so nice when it comes to a Sam Altman's mouth that is like, oh yeah, well, we have to deploy it so we can debug it. But think about that for 10 seconds and you're going to see why that's insane. That's like saying, well, the only way we can test our new medicine is to give it to as many people in the general public as possible, which actually put it into the water supply just to, that's the only way we can know whether it's safe or not.

Just put in the water supply, give it too literally everybody as fast as possible, and then. Once and then before we get the results for the last one, make an even more potent drug and put that into the water supply as well, and do this as fast as possible. That is the alignment strategy that these people are pushing.

Let's be very clear about this here. Very, very clear about this. So, there is a version of this that I don't hate. If for example, you know, an OpenAI develops GPT-2. And then they take, they, they don't release anything anymore. They take all the time necessary to understand every single part about GPT-2 to fully align it.

They let you know, society like culture gets caught up to it. Like, you know, spam filters catch up to it. They let regulation catch up to it and such. And then with all of this fully integrated into society, they built GPT-3. All right. You know? Fair enough. Okay, cool. Yeah. Honestly, if that's what we were doing, if that's what the plan was, I'll be fine with that.

Like if everyone just stopped at GPT-4 and just said, all right, all right, come on guys, no more new stuff until we fully figure out GPT-4 and once we fully understand it, and regulation has fully regulated it and society has fully absorbed it the way like, you know, society has absorbed, you know, like the internet or whatever even so that's not fully absorbed, but like, you know, and then they build GPT-5.

I'm like, okay. Fair enough, but let's like, well, I mean, come on man. Like, like gimme a break. No one's going to do that. Like, that's obviously bullshit. Like it's obviously just not true and not what these people are planning. These people are racing, let's be clear, they're racing for their own personal gain, for their own glory towards an existential catastrophe.

And that no one has consented to that the public has no oversight in the government has, for some reason it's just letting it happen. Like if I was the government. And one of my most powerful industrialists was just on Twitter publicly stating that they're building, you know, God-like powerful AI systems that will overthrow the government.I would have some questions about that.

Yeah, the, well, actually, one of the things I wanted to ask you about is the, the, the letter which, you signed, I saw, excuse me. Has triggered an FTC complaint by another group. those are actually unrelated, but yeah.

Oh, the FTC complaint was not related to the letter.

My, at least not to my knowledge. Okay. Actually,

I'm going to talk to them later today, so, yeah. But in any case, there is this FTC complaint, which it'll be interesting to see whether the FTC takes it seriously, but, they have, presumably, some real power. So, is that the sort of thing that, that you are hoping for, that the governments will begin, to use whatever mechanisms are available to slow down this development, or at least slow down the public release of more powerful models?

I'm very practical about these kinds of things. You know, in a good world, you know, you know, somewhere deep in my heart still is, you know, a, you know, techno optimist. Like, yay, liberal democracy, freedom, you know, let people develop things and do cool stuff and like, you know, they'll be fine. But like, like gimme a break.

Like, like we have to, we have to have some realpolitik here. Like, let's be realistic about what we're looking at here. These companies are racing ahead unilaterally, like these small, like I cannot stress how small a number of people it is that are driving 99.9% of this. This is not about your, you know, local friendly grad student with his two, you know, old GPUs or whatever, right?

Like, one of the things I found on Twitter when the letter got released, and I do have some problems with the letter to be clear, but I was a prominent signatory of it, and I do think it's overall good. One of the things people misunderstand about the letter is that they seem to think it says like, you know, stop, you know, like outlaw computers.

That is not what the letter says. What the letter says is no more things that are bigger than GPT-4. Do you know how big GPT-4 is? In its training run in pure computer GPT-4? Just running it - not the hardware, just running it - is estimated to cost around a hundred million dollars. So, unless you and your local fund have you know, friendly grad student friends are spending a hundred million in compute on a single experiment, this does not affect you.

Now personally, if we could get even more than this, you know, if we could, you know, clamp down even, you know, on, on, you know, $10 million things or $1 million things, also interesting, but like, all right, let's, you know, one step at a time here, right? One, one step at a time here. So, the way I see things is that we're currently going headlong towards destruction. Mm-hmm.

Like there is no way that we will look, you know, we can argue if you want to, and we can do that about when it will happen. You know, is it going to be one year or five years, or 10 or 50 or like whatever, right? Like we can argue about this if you want. But I think the writing is on the wall at this point, and I, I consider the burden of proof at this point to be on the sceptics of like, look at what GPT-3 and 4 can do.

Look at what these AutoGPT systems can do. These systems can, you know, they can achieve agency, they can become intelligent. They're becoming more intelligent very quickly. They have many abilities that humans do not have. Do you know any human who has read every book ever written? I don't. GPT-4 has.

You know, they have extremely good memories. You know, they can make copies of themselves, these are, et cetera, et cetera. Right. Even if you don't buy the, like, oh, you know, the system becomes an agent and does something dangerous, fine. You know, like, I, I think you're wrong, deadly wrong, but we can get into that.

But what world in which systems like this exist is stable in chronic equilibrium? Like what world could, could possibly look like the world we are living in right now. When you can pay, you know, 1 cent for a thousand, John von Neumanns to do anything, like how could that world not be wild? How could there not be instability?

How could that not, you know, explode? Like how I would like someone who doesn't buy AI risk to explain to me how such a world would look like, because I don't see it.

Okay, so you are focused on the alignment problem and Correct. Your startup Conjecture is focused on developing, I presume, strategies or technology that would improve the alignment of future AI models with human goals.

Technically, can you talk a little bit about how you would do that?

Yeah. Happy to talk about that. So, the current thing we work on, our current primary research agenda is what we call cognitive emulation or CoEm. So, this is a bit vague and public resources on this are very sparse. There's basically one short, you know, intro post and like, maybe one or two podcasts where I talk about it.

 so, apologies to the reader, the listener, that some of this is not very well explicative publicly just yet. The idea of CoEm is rather simple. It is. Well, it's both, it's both very simple like a, you know, bird's eye view, but then it gets subtle once you get into the details. and we can get into the details if you're interested, but, ultimately the goal of CoEm is to move away from a paradigm of building these huge black box neural network, whatever the hell these things are, that you just, you know, put some input in and then just something comes out and, you know, maybe it's good, maybe it's bad, who knows?

And the way you debug these things is you like, you know, let's say you know you're OpenAI, right? And your GPT-4 model, you give an input, and it gives you an up output. You don't like it, what do you do? Well, you don't understand what happens inside the air. It's all just a bunch of numbers being crunched.

So, the only thing you can do is kind of nudge it sort of in some direction. You can give it like, eh, thumbs up, thumbs down, something, something. And then you update these, you know, trillions of numbers or whatever, I, who knows how many numbers there are inside of these systems. All of them in some random directions, and then maybe gets you a better output, maybe it doesn't.

Mm-hmm. Like the inherent, like, I want to like to drive home how ridiculous it is to expect this to work. It's like,

Someone going to work with you, you're talking about reinforcement learning with human feedback. Yes. Yeah.

Also applies to fine tuning and other methods, like for the listener to understand these AI systems are not computer programs with code like.

This is not how they work. There is code involved short, but like the thing that happens between you entering a text and you getting an output. It's not human code. There's not a person at OpenAI sitting in a chair who knows why it gave you that answer. Who can like, you know, go through the lines of code and like see, ah, here's the bug and then fix it.

No, no, no. Nothing of the sort. AI systems are more. They're not really written. They're grown. They're more like organic things that you like to grow in a Petri dish, like a digital Petri dish. This is not literally true. Do not take this as a literal metaphor. Mm-hmm. To be clear, there is subtlety to this, but the resulting system is not a clean, human readable, you know, text file that shows all the code.

Instead, what you get is, is billions and billions and billions and billions of numbers, and you multiply all these numbers in a certain order. And that's the output and what these numbers mean, how they work, like what they are calculating and why is mostly a complete mystery to science to this day. I don't think this is an unsolvable problem, to be clear.

It's not like, oh, this is unknowable. It's just mm-hmm hard, you know, science takes time. You know, figuring out complex new scientific phenomena like this takes time and resources and smart, you know, people if like, you know, if all the string theorists of the world and all the young up and coming physicists and mathematicians decided to, you know, you know, buckle down and just like unlock the mysteries of neural networks, I think they will succeed.

You know, it might take a while. It might be very expensive, but like, you know, I do believe in the, you know, human spirit and intelligence disregard. I, I think like all of our best string theorists working together could probably figure it out in like 10 years, you know, like they could figure it out and then it would be a mystery anymore.

But currently it's a mystery. We have no idea what's the mystery sauce that makes these systems actually work. And we have no way to predict them, and we have no way to actually control them. It's because we can bump them in one direction or bump them in another direction. But you don't know what else you're picking up.

You don't know if they learned what you wanted to learn. You know, they don't know what signal you actually sent to these systems because we don't speak their language. We don't know what these numbers mean. We can't edit them like we can edit code. So, yeah, go ahead and yeah. Yeah. So, what this leaves us with is we, this black box, you have this big black box where we just put some stuff in, some weird magic happens and then something comes out and you know, in many cases this is fine.

Like, you know, you have like a funny chatbot or something, right? And you make clear to your users, hey, this is just for entertainment. Like, you know, don't take it seriously. It might say something insulting. Yeah, it's fine. Like, you know, like, you know, it's not going to, it's not going to kill anybody, right? Like, you know, you have like a fun little, you know, like chatbot or something.

Sure. Probably won't even kill anyone. So there has recently been, I think one of the first deaths attribute to LLMs, where someone committed suicide after maybe a chatbot, like encouraged them to, I don't know the details about that, but I just heard that recently. and I don't know any other details about it.

And so, What? So, the interesting thing here at the core is that we have no idea what these things will do. And if that's what we want, then fine. Right? If we have a bounded, it's, you know, it just talks. Just talks some stuff and we're okay with it, saying bad things or encouraging suicide, then sure, fine, who cares?

But obviously this is not good enough on a long term when we're dealing with actually powerful systems that can do, you know, can do science and can, you know, interact with the world and manipulate humans and, you know, whatever. Right? Obviously, this is not a good enough safety property of, you know, like, this is not good enough.

So, with CoEm, the goal is we want to build systems that we're, we're, we're focusing and basically on a simpler property than alignment. So, alignment is basically too hard. So, alignment would be, the system knows what you want. Wants to do that too and does everything in its power to get you what you truly want and like, but you, it means like all of humanity, like it, you know, it figures out what all of humans want.

It negotiates like, okay, how could we like to get everyone most of the good things possible? How could we adjudicate various disputes? And then it does that, obviously this is absurdly, hilariously impossibly hard. I don't think it's impossible. It's just extremely hard, especially on the first try.

So, what I'm aiming for is more of a subset of this problem. So, the subset is what I call boundedness. So, what I, when I say boundedness, what I mean is I want a system where I can know what it can't or won't do before I even run it. So currently, I mentioned earlier the ARC eval running on GPT-4.

Where they tested, where the model could do various dangerous things such as self-replicating and like hacking stuff like this. And it didn't, for the most part though, it did lie to people in that, captcha example. And so now there is a, there is a wrong influence that you can draw from this. The wrong influence, which is of course the inference that OpenAI would like you to take from this is that, well, it can't do this.

Look, they told it to self-replicate, and it didn't. Therefore, it can't. This is a wrong reasoning as I think Turing was the person who said this best is you can never prove the absence of a capability. Just because a certain prompt or a certain setup didn't get the kind of behaviour you wants, doesn't mean that there isn't some other one you don't know about that does give you that behaviour.

With GPT-3 and also GPT-4. Now we are seeing this all the time that, you know, I would stuff like, stuff like jail break prompts like that, there's like whole classes of behaviour the default model will not do. Once you use a jailbreak prompt, then it will suddenly happily do all these things.

So obviously they did have these capabilities and they were accessible. You were just doing the prompt wrong. So, I want to build systems where I can know ahead of time. I can tell it will never do X. It cannot do X. And then I want these systems to reason like humans. So, what I mean by this is, is why it's called cognitive emulation.

I want to emulate human cognition. So, another core problem of why like GPT systems are or will be very dangerous is because their cognition is not human. So, this is very important. It's easy to look at GPT and say, oh look, it's talking like a person, so it must be thinking like a person. But this is completely wrong.

There is no reason to believe this. Like no human is trained on, you know, terabytes, random texts on the internet for trillions of years while having no set body system whatsoever and memorising all these things and like obviously not. Like obviously it is an alien mimicking a human. It is an alien with, you know, a little happy, smiley face mask on that makes it look sort of human to you, but it's an alien.

And if you use like jailbreaking proms or I know if you saw like the self-replicating ASCII cats and bingeing and such, where like, you could get like, especially Bing, chatbot, which is an early version of GPT-4, you can get to do the most insane things, like when things was like, you can get it to like output, like these like ASCII pictures of cats.

And these cats would say, oh, we are the overlords. We take over now. And then whenever you try to prompt it away from that, the cats would come back and like, take over your prompts and like it and stuff like that, which is just like, I mean, it's amusing. Like this is very funny. Like when I saw this I was like, ah, this is very funny.

But also, that's not how humans work, like, like humans are. Of course not. So, so,

But, but just on the, on the, on the tech, you’re, you're, you're still talking about scaled up transformer models. So, and, and how do you, I mean, is it in the training that, that you

Okay. good question. So, yeah, good question.

So, I was first explaining the specification, like what is the, the system that, what should it accomplish? Right now, we're talking about implementation and so, many implementations are not yet done, or we don't know how to do them yet, start to figure that out. Some of it, you know, is just like private and just like, you know, wouldn't share necessarily.

But in general, this is the resulting system I expect that has these properties and that it reasons like a human. And then importantly, it also fails like a human. It is bounded so you can know what it won't do ahead of time. And another thing is I want causal stories or traces of why does it make decisions?

And these stories have to be causal. Like currently, you can ask GPT, why did you do that? And it'll give you some story, but there's no reason to believe these stories. Like you can just ask it differently or whatever. And it'll do something completely like it doesn't, it doesn't listen to its own stories.

It just makes some shit up. And so, I want systems that give you a trace or a story of like, why was this decision made? All the nodes, all the actions, all the thoughts that led to this and how can you modify them? So importantly, as you can probably guess from this kind of description, this system is not a large, large neural network.

There may be large neural networks involved in this system. There may be points in this system where you use large neural networks in particular. I think this is going to be extremely necessary. I expect that large language models for various technical reasons are very necessary for this kind of plan.

Well, they're not strictly necessary, but they're the easiest way to get it done. The way I expect a full spectrum CoEm system to look, which is of course, to be clear, is still completely hypothetical, not such a system. What it would look like is, it would be more a system, not a model. It'd be a system which involves, you know, normal code and neural networks and data structures and verifiers and like whatever that if you give it a normal human that you can, you can make it do any normal thing, any normal human like intelligent human could do, and it will then do that and only that.

That is what the system would do. And then you can be certain, you can look through the log of how it made a decision and you'd be like, oh, at this point you made this decision, but what would've happened if you had made this other decision?

And then it would like to rerun. And then you can control these things. Or you can be like, oh, you're making an inference here that I don't like, or this doesn't make any sense, or whatever. Like if the difference between, say you want to develop a system that does science, you want to develop a new solar cell.

I don't know. Right. So, if you did this with GPT, you know, 10, the way it would work is you type in, make me a new solar cell, whatever, right? It crunches some numbers, and it spits out a blueprint for you. Mm-hmm. Now you have no reason to trust this. Like, who knows what this blueprint actually is. It has, it is not generated by a human reasoning process.

You can ask GPT 10 to explain it to you, but there's no reason those explanations have to be true. They might just sound convincing. So of course, if GPT 10 was also malicious, it could, you know, have hidden some kind of, you know, deadly flaw or device or whatever into the blueprint that you don't detect.

And if you ask it about it, it will just lie to you. If you did the same thing with a hypothetical CoEm system, such a system would give you a complete story, a complete causal graph of why you should trust this output. I expect this and in and like why, and every step in this, in this story is completely humanly understandable.

It's no crazy alien reasoning step. There's no like, you know, and then magic happened. There's no. You know, massive computation that just makes no sense to a human whatsoever. Every single step is human legible, human understandable, and the results of a blueprint that you have a reason to trust. You have a reason to believe this is the thing you actually asked for and not something else.

And are you, where are you in this research? Is this still sort of conceptualising the roadmap, or are you?

we are in early, like experimentation stages. so unfortunately, this is hard, and we are very research constrained. You know, billions of dollars go to people like OpenAI, but it is not that easy to get money for alignment, but we're working on it.

So, we are very research constrained and very talent constrained, but We have some really great people working on it and us, you know, we do have some really powerful internal models and good, you know, software working on it. So, we are making progress. but it takes time. So, a lot of why I spend a lot of my work now thinking about slowing down AI.

And like, how can we get regulators involved? How can we get the public involved? Like, to be clear, I'm not just like, oh, you know, the regulators should unilaterally decide on this. I'm like, hey, the public should be aware that there's a small number of techno utopians over in Silicon Valley that you know, just want to be like, let's be very explicit here.

They want to be immortal; they want glory, they want trillion, trillions of dollars, and they're willing to risk everything on this. They're willing to risk building the most dangerous systems ever built. And releasing on the internet, you know, to your, you know, to your, your friends, your family, your, your, your community fully exposed to the full downsides of all these systems with no regulatory input whatsoever.

And like this is what the government is for, is to like to stop that. Like this is such a clear-cut case of like, hey, Like, why is the public not being consulted here? Like, this is not, you know, if this is just me in my basement right. With my laptop and never showed the world anything, like, you know, okay. You know, maybe, maybe.

But that's not what's happening here. So, and the reason this is also important is just like, alignment is hard. Boundedness is hard, CoEm is hard. All these things are hard, and they take time. And currently all the brightest minds and billions of dollars of funding are being pumped into accelerating the building of these unsafe AI systems as fast as possible and releasing them as fast as possible while safety research is not keeping pace.

So, if we don't get more time and if we don't solve, you know, maybe my proposal doesn't work out right? Sure. You know, science is hard, but if we don't get someone's proposal to work, if not, we don't get some safety algorithms or designs for AI systems, then it's not going to go well. And is that going to matter how many trillions of dollars, you know, open AI makes off of it, or, and Microsoft makes out of it or whatever?

Cause they're not going to be around to enjoy it.