Casey, we talk a lot on this show about multimodal AI, but you know what else is going multimodal?
What's that?
Our podcast. Because we're starting a YouTube show.
A YouTube show!
Next week, we will not only release the audio podcast version of this show,
but we will also have a version that goes up on YouTube where you will be able to see
both of our faces.
A dramatic face reveal that the YouTube community hasn't seen since Dream revealed his face
as part of a Minecraft event?
I'm very excited to start a YouTube show, but I'm also a little nervous because I feel like
we're just going to sort of speed run the evolution of a YouTube show.
Like we're going to start off making like, you know, harmless video game reviews,
and then we're going to say something horrible and get canceled,
and we'll have to make one of those tearful apology videos.
All our brand sponsors will abandon us.
Well, look, I'm thrilled because every day I wake up asking myself one question,
which is, what do total strangers think about my physical appearance?
So Kevin, wait, if people want to get this show, how do they do it?
Can it be done?
It can be done.
In fact, we are launching a new channel.
It will be called Hard Fork, and you can find it on YouTube.
And as they say on YouTube, smash that like button, hit that subscribe button, ding the bell.
Yes.
I'm Kevin Roots, a tech columnist at The New York Times.
I'm Casey Newton from Platformer.
And you're listening to Hard Fork.
This week on the show, three ways researchers are making progress on understanding how AI
works.
Then the hot new manifesto that has Silicon Valley asking,
is Mark Andreessen OK?
And finally, decoding an ancient scroll using AI.
So Casey, we talked a lot on the show about one of the big problems with AI,
which is that it is pretty opaque.
Yeah, I would say this is maybe the biggest problem in AI in a lot of ways,
is that the people who are building it, even as it succeeds across many dimensions,
cannot really explain how it works.
Right. And we also don't know a lot about how these models are built and trained,
what kind of data they use.
The companies have not disclosed a lot of that information.
Basically, we have these sort of mysterious and powerful AI tools.
And people have been really clamoring for more information,
not just about how they're built, but actually how they work.

And this week, we got some news on a couple different fronts about efforts to sort of make AI a little less opaque to demystify some of these large language models.
Yeah, that's right. And by the way, and if you hear that and you think, well, come on, Kevin, in case that doesn't sound like it is a crisis, why is this the most important thing of the week?
Well, this is our effort to understand something that could avert a crisis in the future.
Right. It's like, if we want to have a good outcome from all of this stuff, these actually are the questions we need to be asking and we need to be asking them right now.
Totally. I mean, I was talking with a researcher, an AI researcher the other day, and he sort of said like, you know, if aliens landed on Earth, like if we just sort of found like the carcass of a dead alien in the desert somewhere, every biologist on Earth would drop whatever they were doing to try to get involved in understanding what the heck this thing was, to pull it apart, to dissect it, to study it, to figure out how it works, how it's different than us.
And here we have this new kind of sort of alien technology, these large language models.
And we don't know basically anything about how they work.
And even the people who are building them have a lot of questions about how they work.
So this is really important, not just for sort of scientific understanding, but also for people like regulators who want to understand how do we regulate this stuff, what kind of laws could rein them in. And also for users of these products who after all want to know what the heck they're using.
That's right. And so as of right now, this podcast is Area 51. We're bringing the alien in and we're seeing what we can understand about it.
All right. So I'm getting out my scalpel. The first project I want to talk about is something that came out of Anthropoc, the AI lab that makes the Claude chatbot.
They released this week an experiment called Collective Constitutional AI.
I wrote about this. I talked to some of their researchers.
One of the things that they have been trying to do is to try to invite members of the public, people who do not work at Anthropoc, to weigh in on what rules a chatbot should follow, how an AI model should behave. Yeah. And this is really exciting, right?
Because when you think about the past 20 years of tech, we haven't really had a meaningful vote on how Facebook operates or how Google operates. But now Anthropoc, which is building one of the most important chatbot models, is at least dabbling with the idea of asking average people, hey, how should one of these things work? And by the way, not just how should it work, but what sort of values should be encoded inside of it?
Totally. So the way that Claude, Anthropoc's chatbot that we've talked about in this show before, works is through this thing called constitutional AI, where you basically give the AI a list of rules, which they call a constitution, which might include things like, choose the answer that is the least harmful or the least likely to inspire someone to do something dangerous. So for this collective constitutional AI experiment, what they did was they enlisted about 1,000 people, just normal people who don't work at the company. And they asked them to vote on a set of values or to write their own values. Basically, how should this chatbot behave? They did this in conjunction with something

called the Collective Intelligence Project. This was a panel of roughly 1,000 American adults. And they then took those panel suggestions and used them to train a small version of Claude and compared that model with the Claude that was trained on the Constitution that Anthropoc itself had put together. Yeah. So tell us a little bit about what they found and what differences, if any, there were between the people's chatbot and the corporate chatbot.
So there were some differences. The people's chatbot had more principles in it around sort of neutrality and avoiding bias. It also had some things that Anthropoc hadn't explicitly said in their Constitution for Claude. For example, one of the principles that the panel of public participants came up with was AI should be adaptable, accessible, and flexible to people with disabilities. So that was maybe something that didn't make it into Claude's Constitution but was part of this experiment. So this list of suggestions got whittled down into 75 principles. Anthropoc called that the public Constitution. And when they compared a version of Claude that had been trained on the public Constitution to one that was trained on the regular Constitution, they found that they basically performed roughly as well as one another and that the public Constitution was slightly less biased than the original. Which is interesting. Although I was noting, Kevin, in the methodology that while they tried to get a sample of people that was representative of America in terms of age and some other demographics, race was apparently not one of them. Did you notice that in their methodology? No, I didn't. I mean, I did notice some interesting other quirks in their methodology. Like they tried to have it just be sort of a representative cross sample of everyone. But then they found that some people would just give these like answers or suggestions that were totally inscrutable or off topic. So they had to actually narrow it down to just people who were sort of interested in AI. So already they have had to do some sampling to make this project work. But I think more so than the sort of results of this experiment, I'm interested in kind of this collective governance process, right, inviting people to take part in writing the rules of a chatbot. And it's interesting to me what sort of problems that might solve and actually what problems it could create. Yeah, I think the general idea of give the people a say in the values of a chatbot is basically a good thing. Maybe we don't want it to be the only input, but I definitely think it should be an input.
I think the thing that I worry about is that to me, the ideal chatbot is one that is quite personalized to me, right? There are people in this country who have values I do not share, and I do not want them embedded in my chatbot. And in fact, there are already countries that are doing this. If you use a chatbot in China and you ask it about Tiananmen Square, it is not going to have a lot to say, right? Because the values of the ruling party of China have been embedded into that chatbot. Again, thinking about the kind of chatbot I want, I might want my future chatbot to be gay as hell, right? And the majority of Americans, they're probably not going to think about wanting that to be a big feature of their chatbot. So by all means, take the people's will into an account. It's a massive step forward from where we are today. But over time, I think the best versions of these things are very personalized. Yeah, I agree. And I think, you know, especially when AI
chatbots are going to be used for education in schools, like you really don't want what has happened with things like textbooks where you have different states teaching different versions of history because they're controlled at the state level by different parties. You really don't want like your Texas chatbot that'll teach you one thing about critical race theory and your Florida

chatbot that'll teach you another thing and your California chatbot that'll teach you a third thing. That just seems like a very messy future. But also, it seems like our most likely future because if you believe that these AIs are going to become tutors and teachers to our students of the future in at least some ways, different states have different curricula, right? And there are going to be some chatbots that believe in evolution and there are going to be some chatbots that absolutely do not. And it'll be interesting to see whether students wind up using VPNs just to get a chatbot that'll tell them the truth about the history of some awful part of our country's history. Yeah. So that is crowdsourced AI constitution writing. But there was another story that I wrote this week about something that came out of a group of researchers at Stanford at their Institute for Human-Centered AI. They released something this week called the Foundation Model Transparency Index. This is basically a project that scores companies that make big AI models, open AI, Anthropic, Google, Meta, et cetera, based on how transparent they are. How much do they disclose about where their data comes from? What labor goes into training and fine-tuning them? What hardware they use, as well as how people actually end up using their AI models
in the world. Yeah. So let's not build it up too much, Kevin. How did these companies do on their big transparency scores? Well, they didn't do great. Meta actually had the highest score for their model, Lama 2, but the highest score was unfortunately only a 54%. Now, I spent a while since I've been in school. Is that a passing grade? No, that is not a passing grade. So they failed, but they failed less than other models. So GPT-4 and Palm 2, which is the language model that powers Bard, both received 40%. So pretty abysmal scores on this test. So what the Stanford team is basically trying to do is to start kind of a competition among AI model developers for who can be the most transparent and to reward companies for being transparent by giving them higher scores, which will hopefully encourage other companies to take note and do some of the same kinds of disclosures. It's like the US News and World Report for college rankings, but for AI models. Exactly. So the researchers who set this index up, they told me that part of the reason they did this is because as AI systems have gotten more powerful, they have actually gotten more opaque. We now know less about models. Five years ago, if you built an AI language model, you might talk about the data you used. You might say more about the hardware that you're training the model on. But for various reasons, some of which have to do with the threat of lawsuits and the fact that a lot of these companies see themselves in sort of a competitive race with one another. They don't disclose much of that information at all anymore. And so you really do have this scenario where AI is getting more powerful and at the same time, we're learning less about how it works, what kind of labor and practices go into it, and how it's used in the world. Yeah. Now, of course, Kevin, there is a tension between sort of giving it all away and being able to stay alive as a business. And there's probably, I would imagine, some risk in sharing absolutely everything that goes into how these models are trained. So in some ways, these companies may have a point when they say, we have good reason to be more opaque. Absolutely. I mean, if you just look at the just the lawsuit angle alone, a lot of the suits that have been filed against these AI companies by authors and artists and media organizations that accuse them of sort of using copyrighted works to train their models, those lawsuits have mostly targeted projects that divulged a lot of information about where they got their data, right? There's sort of like a penalty

right now in the legal system. If you do disclose where you got your data, you're more likely to be sued for it because, you know, most if not all of these developers are using copyrighted works to train their models. Yes, there is unfortunately a penalty in our legal system for breaking the law. And so there's probably a lesson in there for future AI developers.
Right. But do you think transparency on the whole is a good thing for AI models?
I think so. Transparency winds up becoming one of the best ways that we can regulate tech companies, particularly in the United States, because our First Amendment just prevents the government in a lot of cases from adopting regulations that you might see in Europe in other places, right? You can't tell corporations how to speak, how to produce speech, but what you might be able to do is say, well, okay, you at least have to tell us some information about what is in this thing, right? We're not going to compel speech from these models, but we are going to make you tell us a little bit about what you've put into these things. Yeah, I think this is really important. I don't know if there needs to be some kind of like, you know, safe harbor established, like you can share this information and you, you know, will sort of shield you from some kinds of lawsuits. Some researchers I've talked to have thought that that could lead to more transparency if there wasn't this kind of threat of litigation. But ultimately, I think this will just need to be sort of pressure that is applied from regulators, from the public to sort of increase the transparency of these models.
Yeah. Good thing. It's a good thing. We like it.
It's a good thing. We love transparency. Yeah.
All right. So story number three in the world of bringing more transparency and understanding to AI is basically some research that has come out in the past few weeks about interpretability. Now, this is the subject that we've talked about on the show before. There is now a field of AI research that is concerned with trying to answer the question of like, why AI models behave the way they do? How they make decisions? Why certain prompts produce certain responses? Because this
is one of the big unanswered questions when it comes to these AI language models.
Yeah. And by the way, if you have ever used a channel, I don't know about you, Kevin, but when I use these things, I have one thought every single time, which is, how is this thing doing that? Right? Like, it's impossible to use one of these things and not have this exact question at the top of your mind. Right. So there were two very similar studies that came out over the past couple weeks, one from a team at Anthropic and one from a team of independent researchers, both trying to make progress toward understanding this issue of interpretability. And it gets pretty complicated, but basically, these AI models, they are composed of these little things called neurons. Neurons are little mathematical boxes that you put an input into and you get an output out of. And if you ask a chatbot, write me a poem about baseball, all these different neurons in the model get activated. But researchers didn't actually understand why these neurons got activated, because some of these same neurons could also end up getting activated by something totally different, like a prompt about geopolitics or something in Korean. So for a while now, researchers have been trying to kind of group these neurons or sort of figure out which ones are linked to which concepts. This is very hard for reasons that are kind of technical. They involve this thing called superposition, which I won't even try to explain, because I only sort of understand it myself. But basically, this turns out to be... Can I just say, Kevin, I'm in a superposition

right now because you're having to explain all these hard things. So that to me, that's what a superposition is, but go on. So anyway, basically, they ran some experiments, these researchers, where they took a very small model, like a much smaller model than any chatbot would use. And they figured out that you could use something called a sparse autoencoder to identify something called features, which are basically groups of individual neurons that fire together that end up sort of correlating to different things that we humans would recognize as a coherent concept or an idea. So one feature might be JavaScript code, another one might be DNA sequences, another one might be Romanian language text. All of these sort of features would go into producing the response that the chatbot gives, and they were actually able to map the relationships between individual neurons and these so-called features. So very dense, very technical, wouldn't recommend reading these papers unless you have a PhD or are willing to sit there for many hours trying to make heads or tails of it. But basically the researchers that I've talked to about this say, this is a huge win for the field of interpretability, because finally we are able to say, at least on these very small models, that we now understand why certain models activate and do the things that they do. And so the result of that is that if this work continues successfully, when you go ask a chat model to write a poem about baseball, we will have some understanding of why a poem about baseball is delivered to you. We will be able to see maybe where poem is in the model and where baseball is in the model. And if you're saying, well, that doesn't seem very useful, you might think about it if it's like, well, someone just used one of these things to devise a novel bio weapon. In that case, it'll be good to know where novel bio weapons are in the model, which we currently do not. Exactly. And the researchers that I talked to about this, they also said that one way this could really help us is figuring out if and when AI models become deceptive, right? If you ask an AI model something, do you know how to build a bomb? And the AI model has been sort of trained or fine tuned to answer, no, I don't know how to build a bomb. But the AI in its model actually does know how to build the bomb. It's just lying. Researchers believe that they might actually be able to go in and kind of see the same way that you might use a brain scan to sort of figure out where a human's brain is lighting up when they feel certain emotions. You could actually go in and kind of see the deception inside the model, which would be really important if these systems got super powerful and did start to behave in deceptive ways. I agree with that. And can I actually just say Kevin, do you know one thing I did this week during the sort of dedicated time I'm now trying to set aside to interact with these large language models? What did you do? I tried to get it to tutor me about interpretability. I just, I asked it about, what are the general problems in interpretability? And it would give me like 14 of them. And I would say, well, let's like drill down on these three, like tell me the history of this problem. And, you know, it is a large language model. So there's always the risk that it is hallucinating. But I do think that if what you just want is kind of the, if you want to get a flavor of something that you don't have to stake your life on, getting pretty good. And I will confess that one of the ways that I prepared for this segment was by uploading one of these PDF research papers into an AI chat button, asking it to summarize it for me at an eighth grade level. And it would be so funny if the AI actually already was deceptive and was just like, Oh yeah, Kevin, you've already figured us out. We would never trick you. Yeah, you got us. Exactly.

When we come back, a strong new challenge to the communist manifesto.

Well, Kevin, there's a hot new manifesto sweeping Silicon Valley. We love a manifesto. Now, have you read the techno-optimist manifesto by area businessman Mark Andreessen? I did. It was very long. It was like 5,000 words. Mark Andreessen is of course the venture capitalist who's the co-founder
of Andreessen Horowitz and was the co-founder of Netscape. So a person who has drawn a lot of attention in the tech world and who also likes to prod the tech world to adopt his ideas through these lengthy manifestos that he writes. Yeah, and he has said a bunch of things in the past that turned out to be true. One of his more famous pieces was called Software is Eating the World, which was this basically true observation that every industry was being transformed by technology. In 2020, he wrote another very widely read piece called It's Time to Build, which I don't really know how profound it was, but people did spend the next few weeks talking about what they were going to be building to impress Uncle Mark. Totally. So this manifesto came out this week
and it begins with a section called Lies. He writes, quote, we are being lied to. We are told that technology takes our jobs, reduces our wages, increases inequality, threatens our health, ruins the environment, degrades our society, corrupts our children, impairs our humanity, threatens our future, and is ever on the verge of ruining everything. Yeah. And I would just note that
this rhetorical structure of beginning with your being lied to is borrowed from the Tucker Carlson program on Fox News. Yeah. And in fact, I would describe this text overall as something between like a Tucker Carlson monologue and almost like a religious text. Like he actually does say further down in it that he is here to bring the good news. He says, quote, we can advance to a far superior way of living and of being. We have the tools, the systems, the ideas, we have the will. It is time once again to raise the technology flag. It is time to be techno optimists. And am I right that bringing the good news was previously associated with the apostles of Jesus Christ?
That's true. That's true. Okay. So I would say the overall thrust of this manifesto is that people are being mean to technologists and they're being critical of new technology and that actually what we need to do as a society is recognize the technology is the only sort of force for progress and innovation and improvement in society and that the people who build technology should be celebrated as heroes and that we should essentially accelerate. We should go faster. We should remove all of the barriers to technological adoption and essentially usher ourselves into the glorious techno future. So what did you make of this manifesto, Casey? Well, I had a very emotional reaction. Then I took a step back and then I tried to have a more
rational reaction, which is I think maybe where I would like to start. No, no, no. I want to hear about your emotional reaction. Were you slamming things against your keyboard? Did you have to like shut your laptop in a fit of rage? I read it and I thought like he's lost it
because it is so messianic and it is fervor for technology absent any concern for potential misuse that I think if founders took this as gospel and it is very much intended as a gospel, then I think we could be in trouble. So yeah, my first emotional reaction was this is bad and irresponsible. Yeah, and I think I know exactly which part of this manifesto provoked that reaction to you. I believe it was probably the part where he spelled out who the enemies of technology are and it was basically everyone in the field of trust and safety, anyone concerned with tech ethics or risk management, people involved in ESG or social responsibility. Basically,
if you are a person in the technology industry whose job it is to make technology safer or more socially responsible, you are among the enemies that Marc Andreessen lists off in this manifesto.

Am I correct? Yes, those are absolutely the enemies that he lists. I think you missed the communists. That was the other big one. He's very concerned about the communists. He did say that everyone who objects to technological progress is either a Luddite or a communist and to that, I say why not both? Communism comes up so much in this manifesto that I had to check and make sure it wasn't written in like 1954, like before the McCarthy hearings. I was like, where is this vibrant communist party in America that Marc Andreessen is so concerned about? We're laughing, but I think this is a kind of document that captures a certain mood among a set of Silicon Valley investors and entrepreneurs who believe that they are basically being unfairly scapegoated for the problems of society, that they are building tools that move us forward as a civilization and that their reward for doing that has been that journalists like write mean stories about them and people are mean to them on social media and that we should sort of embrace the natural course of progress and get out of the way and that we would all be better for that. But I want to just sort of set the scene here by just saying that these ideas in and of themselves are not new, right? There has been this call for many years for what used to be called accelerationism, which is this idea that was kind of popularized by some philosophers that essentially technology and capitalism, they were sort of pushing us forward in this inexorable path of progress. And that instead of trying to regulate technology or place barriers or safeguards around it, that we should just get out of the way and let history run its course. And this was popularized by a guy named Nick Land who wrote a bunch of controversial essays and papers several decades ago. It was sort of adopted as a rallying cry by some people in the tech industry. It sort of says we are headed into this glorious techno capital future. Nick Land calls it the techno capital singularity, the sort of point at which we will no longer need things like democracy or human rights because the market will just take care of all of our needs. Well, I mean, how do we want to take this conversation, Kevin? Because I feel like I could just sort of jump in and start dunking. There are things in this that I agree. Is it worth just taking a moment to say that like, yes, there are things about this that we do agree with? Yeah, totally. I mean, I think you and I are both relatively optimistic about technology. I don't think we would be doing this show if we just thought all technology was bad. I'm not a Luddite. You're not a Luddite. I firmly believe that technology has solved many large problems for humanity over the centuries and will continue to solve large problems into the future. I'm not a person who's reflexively opposed to optimism around technology. But I do think that this manifesto goes a lot farther than I would ever feel comfortable going in sort of asserting that technology is automatically going to improve lives. And this is a thing you hear a lot from people who are optimistic about AI. They say, well, yeah, it'll destroy some jobs and make some people's lives harder and cause some problems. But ultimately, we will all be better off. And they justify that by saying, well, that's always how technology has worked, right? None of us would switch places with our great, great grandparents. We don't want to be subsistence farmers again. We don't want to rewind the clock on technology. So technology just always improves our lives. And what I would say to that is, well, I think that's true in the long run, but in the short run or in individual people's lives, it is not always true. I've been reading this book by the economists, Daron Osamoglu and Simon Johnson called Power and Progress. And you could almost
read that book as sort of a rebuttal to this Mark Andreessen essay, because it's all about how

the gains of improved technology are not automatically shared with people. Progress doesn't automatically create better living standards. Technology is kind of this wrestling match between machines and humans and politicians and policymakers and workers. And we just have to focus our energy there and making these technological gains as widespread as possible, rather than just sort of assuming that all we have to do is invent new technology and then everyone's going to be automatically better off. Yeah. And I think one of the main challenges to Andreessen's view of the world is just the rampant inequality in the United States in particular, right? If all it took was capitalism and technology to lift every single person out of poverty, it seemed like we'd be further along than we are right now. And instead what we've seen over the past half a century or so is that the rich get richer and the middle class is shrinking. So I just think empirically that this argument has a lot of challenges standing up. And it's maybe worth asking like, well, why does Mark Andreessen feel this way? Well, think about how his life works. Like he raised a bunch of money from limited partners and he gives it away to technologists, and the technologists make successful companies. And now Mark Andreessen is a billionaire. So like this approach to life is working great for Mark Andreessen. I think the issue is that not everyone is a venture capitalist and not everyone has such an uncomplicated relationship with the technology in their life. And so on top of everything else in this piece, I just feel like there is a solipsism to this essay that I found very concerning. Yeah. I think it's obviously a product of someone whose optimism has paid off, right? Venture capitalists are not paid to bet against things. They are paid to bet on things. And when those things work out, they make a lot of money. So he is incentivized to be optimistic. But I also think it's part of this sort of undercurrent of the conversation, especially around AI right now. One movement, I don't know if you'd call it a movement. It's basically, you know, people on the internet, but you might call it like a collective of people who believe in what they call effective accelerationism. You might have seen people on Twitter putting like E slash ACC or EAC on their X handles. And Mark Andreessen is one of those people.

And this is a movement. We can talk more about it on some other episodes sometime. But basically, it's sort of a response to effective altruism, which among other things advocates for sort of slowing down AI for being very careful about how AI is developed for putting very tight guardrails around AI systems. And effective accelerationists basically say all of that is just hindering us from realizing our full potential. And we should get all of these sort of hall monitors out of the way and just march forward consequences be damned. But Casey, can I ask you a question about the kind of motivation for someone to write a document like this? Because what I got from this document, the sort of tone of it is written from the point of view of someone who has a very large chip on their shoulder, right? Mark Andreessen, he is clearly so angry at all of the people who criticize technology, technology companies, tech investors. And from where I sit, looking at the world as it exists today, like he has won, you know, yes, venture capitalists, technologists, they are the richest and most powerful people in the world. They control vast swaths of the global economy. Why do you think there is still such a desire among this group of people to not only sort of win the spoils of technological progress, but also to sort of be respected and hailed as heroes? Why are they so thirsty for society's approval? I mean, partly that's just a basic human thing, right? You want people to think you're right. Also, these are the most competitive people in the entire world. And even after they have won, they're not going to be gracious winners.

They want to just kind of keep beating up on the losers. You know, we should also say there's just a strategic business purpose to Andreessen writing like this. Andreessen's job is to give away other people's money to try to get a return on it. It turns out a lot of people can do that job. So how do you get the most promising founders in the world to take your money? Well, in part, you do it by sucking up to them, flattering them, telling them that they're a race of Nietzschean supermen and that the future of civilization depends on their B2B SaaS company,
right? And he is just, you know, really, really going after that crowd with this piece. And I think a secondary thing is it's hugely beneficial to him to annoy people like us, right? When the hall monitors and the scolds of the world hop on their podcasts and say, shame on this man for his blinkered view of society, he gets to take all those clips and send them to the founders and say, look, look at all the people that are lining up against you, the race of Nietzschean supermen, as you try to build your SaaS companies. So I am aware that even in talking about this, we are just kind of pawns in a game. And I still think it is worth doing it because one of the first things that we ever said on this podcast was technology is not neutral. And as I read through, I'm sorry, as I slogged through all 5,000 words of this thing, it kept occurring to me how easy it is to sit back and say technology will save us. It is much harder to talk about like, oh, I don't know, here's Clearview AI. It scraped billions of faces without anyone's permission and is now being used to create a global penopticon that's throwing innocent people in prison. Is that a good technology mark? Should we accelerate that? He doesn't want to engage with that, right? He wants to do content marketing for his venture firm. Yeah, I agree with that. But at the same time, I also want to take this idea of techno optimism seriously, because I think this is a flag that a growing number of people in tech are waving, that we sort of need a societal reset when it comes to
how we think about technology where, and this is a part that I would actually agree with. I do think we should celebrate breakthroughs in scientific progress. I thought we should have had a ticker
tape parade in every city in the world for the inventors of the COVID vaccines. I think that we should lift up on our shoulders people who do make material advances in science and technology that
improve the lives of a lot of people. I am on board with that. Where I fall off is this idea that anyone who opposes anything having to do with technology wants it to be sensibly regulated, wants to have a trust and safety team, wants to talk about ethics, is somehow doing it for cynical or knee-jerk reactions, that they are communists, that they are Luddites, that they hate technology. In fact, some of the people who work in tech ethics and trust and safety love technology more than any other people I know. And that is just a thing that I don't see reflected in this manifesto. That's right. They have arguably made greater sacrifices because they have chosen the part of this job that is not fun at all, but they're doing it because they think that we can live on a better and safer internet. And that's why I think that them and we are the real techno optimists. Because we believe that technology can make us happier, more productive, can help society
grow. We just also believe that doing that in the right way just requires all of society to be given a seat at the table. It cannot all be left up to people with a profit motive to design a perfect

future society. You have to take other perspectives into account. And what makes me an optimist is I think we can do that. And in fact, I think in general, the tech industry has gotten much better at that, right? Like there is this reactionary quality to this essay that you can sort of hear him saying like, you know, I didn't used to have to listen to all these jackals when I would build my technology. I didn't used to have to take into account all of these other voices. Think about how the work that I was doing affected people that are unlike me. But now I do and it's infuriating me. But I think that is a good thing because I think an optimistic view of the world is one that takes into account other views and perspectives. So I think that regardless of what Andreessen asked
to say, like, I do think by hook and by crook, we're going to build a better society a little by little, but it is not going to be through venture capital alone. Right. Right. I think it's a great point. And I also think that the sort of thing that I'm coming away from this thinking is like technology is not just about tools. It is also a collaboration with people and communities and regulators. There is a social acceptance that is required for any technology to work in society. For example, one, one example that Andreessen brings up in his essay is nuclear power, right, which was invented in the 20th century. But then there was this movement that sort of had concerns about the use of nuclear power. And so they were able to sort of shut down nuclear power plants to sort of stop the proliferation of nuclear power. He thinks that's a very bad thing. I would actually, you know, agree that we should have more nuclear power. But that was an argument that that sort of
the pro nuclear power folks lost in the marketplace of ideas. And so my response to that is not that no one should be able to argue against nuclear power. It's that the pro nuclear power people have to do a better job of selling it to the American public. You actually do have to fight for your ideas and convince people that they are good, because not everyone is going to be sort of reflexively and automatically supportive of every new thing that comes out of Silicon Valley.
That's right. I mean, democracy is exhausting. I think that's why it's so unpopular. You know, it's like you have to fight really hard to advance your ideas. But at the end of the day, you know, it's the only system of government I want to live under.
Right. I also think one thing that was just blatantly missing from this is just the sort of market demand for some of these things that Mark Andreessen decries as sort of nanny state paternalism. This is my favorite, the capitalist critique of Mark Andreessen's manifesto, because it's shockingly easy to make. Totally. I mean, he thinks that trust and safety is this kind of piece of sort of activism within the tech industry. It actually emerged as a demand from advertisers who wanted platforms that they could safely throw ads onto without being next to all kinds of toxic garbage. So trust and safety emerge not out of some desire to like control speech and impose values, but out of actual demand from the customers of these platforms, the advertisers who wanted to be on them. You could say the same about AI ethics or AI safety. This is not something that is emerging kind of out of ideology alone. Companies want these chatbots to be safe so that they can use them without them spewing out all kinds of nonsensical or toxic garbage. They do not want badly behaved chatbots. And so I think it's just missing this whole element of like these things that he hates, that he thinks are anti-progress, are actually preconditions for the success of technology. You cannot have a successful chatbot that succeeds in the marketplace and makes money for its investors if it's not well behaved, if it's,

if it doesn't have a trust and safety or an ethics process behind it.
Just to underscore this point, Mark Andreessen sits on the board of Metta and has for a really long time. Metta has a really big trust and safety team. The existence of that trust and safety team is a big reason why Metta makes a lot of money every quarter. And it takes that money and it invests it into AI and building all these future technologies, right? So if you want to be an EAC, okay, then you should actually endorse trust and safety because it is fueling the flywheel that is generating the profits that you need to invent the future.
100%. All right, let's leave it there with Mark Andreessen and his manifesto. Casey, are you going to write on it? Wait, don't we want to challenge him to a debate?
I feel like in these things, it always ends with we challenge him to a debate.
Well, I feel like maybe a cage match would be more appropriate. I don't know, we could take him. Maybe if we fight him both at once, he's a tall man.
All right, Casey, when we come back, let's talk about a project that actually makes me more optimistic about where technology is headed. It involves ancient scrolls and AI.
All right, how should we get into it? Well, I think this is sort of the natural continuation of our ongoing discussion about ancient scrolls. Here, how about this?
You heard of rock and roll? Well, today we're going to rock and scroll. Yeah, that's right, because a contest to decode an ancient scroll is sweeping the nation. And it turns out, Kevin, that someone was able to solve it using the power of artificial intelligence.
That's right. So this is something called the Vesuvius challenge. And it's called that because it has to do with a series of rare scrolls that were buried and preserved in a volcanic eruption of Mount Vesuvius in the year 79 AD. Now, Casey, you're a little bit older than me, but you weren't actually around in 79 AD, right? No, I wasn't. And by the way, this is not the same Vesuvius challenge as the one that went viral on TikTok where people tried to throw their friends into rivers of liquid hot magma that they wound up having to take off the app. This is a different Vesuvius challenge. No, so the Vesuvius challenge is this kind of crowdsourced scientific research contest that has sort of been taking Silicon Valley or at least a chunk of Silicon Valley by storm. It was started by this guy, Nat Friedman, the former CEO of GitHub, who's now a big AI investor,
and Daniel Gross, his investing partner, a founder of Q, John and Patrick Collison, the co-founders of Stripe, kicked in some money as well as people like Toby Lutke, the founder of Shopify, Aaron Levy, one of the co-founders of Box. A bunch of people from the tech industry have contributed money toward this challenge. And basically, the story is that these scrolls that were preserved in the eruption of Mount Vesuvius have long been a source of fascination for historians because there's some of the few examples of written material from ancient Rome that we have, but they're so charred and fragile that a lot of them have never been opened or read. They look like little burnt coal burrito things. And for many years, historians have been trying to open these scrolls to read what was inside them, but because they had been so carbonized and turned into essentially ash, when they would try to open one, it would just crumble and fall apart. So for a long time, there was this question of, will these scrolls that are called the Herculaneum
Scrolls, will we ever be able to actually open them and read what's inside?
Once a week, I feel like I'm pulling you aside saying, Kevin, where are we on the Herculaneum

Scrolls? Exactly. So Brent Seals is a computer scientist at the University of Kentucky, and he has spent the last 20 years working on the technology that is now being used in this challenge, this project to open and read what is inside this trove of literature from ancient Rome. And he developed an approach that uses some imaging technology and some AI to uncover things in these scans without actually physically opening the scrolls. You can just kind of scan the insides of them and then use AI to sort of try to decipher it. And he is sort of the founding father of this Vesuvius challenge, which has now been thrown open. People have been invited to compete to try to decipher these scrolls. And you know, it's like we're using the technology of the future, Kevin, to talk to the past. Exactly. So last week, the organizers of the Vesuvius challenge announced that one of the sort of intermediate prizes had been won. There's a big prize at the end, $700,000. This was a $40,000 prize, and it was won by a 21-year-old student, Luke Farator, who used an AI program that he had developed to read the first word from these scrolls. And that word, do you know what that was? The word, I believe, was purple. It was purple. So to talk about how this challenge works, what this AI technology does, and what this might mean for history, we've invited Brent Seals onto the show. Brent Seals, welcome to Hard Fork. Well, thank you so much. First of all, can I just ask you, you are a computer scientist, not a historian or a classicist or a scholar of ancient Rome or ancient Greece. How did you get interested in these ancient scrolls? I am an imaging specialist, and I came through my graduate work as a computer vision specialist. The internet occurred, though, and pretty soon images really became about the material that we wanted to digitize and make available in libraries and museums. And that was what pulled me in to the world of antiquities. We did our first imaging at the British Library with the Beowulf manuscript, and we took images of it and created a digital edition. And during that time, the conservators said, it's fine to take photos of this beautiful manuscript, but how about this one over here that's so damaged you can't even make a digital copy of it? What are you going to do about that? And it occurred to me, wow, museums and libraries are packed full of stuff that's never going to make it on the internet, because we don't even know how to make a digital copy of it. What does that mean? And in terms of the significance of these particular scrolls, the Herculaneum scrolls, why are they so important for historians? What might they contain or reveal? Well, there are a few reasons why Herculaneum is so enigmatic. It's really the only library from antiquity that survived, the only one. But another thing about it is that there are pieces of it that are still completely unopened. How many of these unopened scrolls are there? The unopened ones number in the three, four, 500 range. And counting is hard because they're fragmentary. Even the unopened ones might be two scrolls that are in three pieces or one scroll in two pieces. So how do you really count? And it's the case that the lettering that is used, the words that are used, we have an ability to read it as long as we can make out the characters. So the game here is just getting the scrolls in a position where you can actually see the characters that are written down. Is that right? Yeah. But I mean, we're using diminished imaging capabilities because the scrolls are completely wrapped up and we can't physically open them. If we could photograph them every surface, we would be able to read everything that's there, but you can't open them for photography. So what we've been able to figure out is how software and AI and machine learning can overcome the

fact that the diminished modality of X-ray creates a real challenge. Yeah. Yeah. So you've been at this project of trying to read these scrolls using technology without opening them for many years. This is not a recent project that you started. So what are the big sort of milestones or moments in your quest that stand out to you? Probably the most significant early moment was simply having the chance to take a crack at it in 2009. We looked inside a scroll for the first time ever. And so that was an incredible moment, but unfortunately, it wasn't a moment where we could read anything.
We saw the internal structure and we knew that the imaging would work at least to show us the structure, but it became a real challenge to figure out how to move on from that.
And can you explain just in basic terms how you were able to read the structure without opening the scrolls? The tomography that we use is exactly what you would get if you went to the doctor's office for an analysis of, say, a broken bone or a fracture. It's X-ray in 360 full round detail, but at a scale that is 10 times or even 100 times more precise than what you get at the doctor's office. So where does the AI part come into this process? What has AI been useful for in terms of deciphering what these scans of these scrolls are able to detect? The ink that was used in Herculaneum, when you look at that ink with X-ray, looks exactly like the papyrus that it's written on.
It'd be like you wrote a memo to me in a white pen on a white piece of paper, and then I'm supposed to read that because the visible light would not show me the difference between white ink and white paper. X-rays don't show us clearly the difference. And for a long time, people basically said, seals is trying to do something that actually isn't physically possible because X-rays just won't show the difference between the density of the ink and the density of the papyrus.
Wait, so you had like haters who were saying he's never going to figure this out? Were they like posting on Reddit or where was this happening? I wouldn't call them haters, but they sound like haters. Okay, so let's call them skeptics or let's call them even, you know, just nervous Nellies who are pretty sure we're not going to achieve what we're saying.
So how did the AI help solve that problem of there not being a difference between the ink that was used on the scrolls and the papyrus itself? The thinking was to find a way for AI to bring its power to bear on this problem. Maybe the machine learning framework can be trained to detect that we're just not seeing with the naked eye. So we started to say, what if we showed the computer machine learning framework, examples of ink and examples of not ink, whether
or not we can see any difference, right? And then see if we can learn it. And it can. It does. We've done those experiments. That was really the key that unlocked the whole thing. Yeah, like if you've used Google Photos and you type dog into the search bar, Google's machine learning model has learned
what is a dog and what is not a dog. And that's just the easy way of finding all the dogs in your photo. Of course, most of us can spot a dog with a naked eye. What you're saying is this technology has now advanced to the point where something that is much harder to discern to the naked eye now can be discerned. That's exactly right. And it's x-ray. So we are already a little bit at a disadvantage in seeing what that looks like because it looks different in x-ray. Everything does. Some of the ink actually is visible to the naked eye. And so the haters, if you will, who said you'll never see carbon ink, were wrong on two fronts. I mean, first of all, you can straight

up see some of it with the naked eye. So they were completely wrong on that. And then the second thing is that the subtlety of it is actually teasable with the machine learning. And that's the part of it that you can and all see with the naked eye. And we can still see that too.
All right. Well, so let's talk about the Vesuvius challenge. What was the actual Vesuvius challenge? We'd done all this work to confirm that the machine learning worked. We had the data in the can. But what we wanted to do is accelerate the work rather than the three or four or five people I have at my research team sort of hammering away. So Nat Friedman, who is former CEO of GitHub, pitched me the idea of a contest. And it turned out to be a brilliant idea, really. And I'm so glad we teamed up with him because it's hard work. And the composite of all those people is a magnificent accelerated step forward. And so what are all these people doing? Like, are they all running their own machine learning models? What are they doing?
They're doing exactly that. I mean, we made a platform for them of a tutorial. And we released all of our experimental code so that they could sort of run examples. And then we made all the data available so that really all they needed to do is exactly what you're saying, hey, I have a machine learning model. I'm going to try it. Hey, I know where some ink is. I'm going to make my own
labels and I'm going to run that. Right. So now you have this group of sort of technologists in Silicon Valley who are funding this prize to sort of incentivize participation in this challenge. You also have people, college students and other people with sort of technical expertise, actually writing AI programs to help decipher these scrolls. What do you think appeals to these tech people about this project? Because from where I sit, you know, we talked to a lot of tech people on this show. They're very interested in the future. They're not always so interested in the past and especially the ancient past. So what about this project to recover these manuscripts from antiquity? Do you think has captured the attention of so many people in tech?
You know, I don't really know. Kevin, I was hoping you would tell me because I've been so embroiled in
it that I have my own passion and it's hard to think outside of it. I have a theory. I know that the narratives that we construct and that people play for entertainment in those video game circles are very strong and a lot of times they have components that go back. Somebody's running through, you know, medieval Venice, for example. And I think there may be a strand that's intriguing that is about the mystery of the Roman Empire. That may be a thing.
Well, and we know that men are always thinking about the Roman Empire.
Is that true? That's a meme, at least, that's been going around where you have to ask men in your life how often they think about the Roman Empire. Has no one asked you this yet?
No, I didn't know that. Oh, they will soon. They will. Some Gen Z student is going to ask you how often you think about the Roman Empire and your answer will probably be every day.
I think about it all the time, yeah.
I mean, my theory, and this could be totally wrong, is that it is just,
it is a puzzle and people love puzzles in tech, but it is also an area where something that used to be impossible or was thought to be impossible suddenly became possible because of AI. And that is just kind of one of these technically sweet problems that these folks love to work on. Also, a lot of them are big history nerds. So, I think it's probably a combination of those things.
Well, there is a deeply resonant, you know, thread that's going through the community,

and it's incredibly varied. I mean, we're seeing in the Discord channel,
paperologists talking with 21-year-old computer scientists and kicking ideas back and forth.
I mean, I love that so much because I think that our future world is going to require that we break down all these barriers of disciplines and also of interests so that we get better communication.
Tell me about the role that Discord has played in your project. I haven't heard about a lot of science that has a Discord component to this point.
I think it's been hugely influential. I mean, the competitors,
unlike what I thought would happen, I figured everyone would hold things pretty close to the chest, but at least in the early part of this competition, they readily shared
some small victories and defeats and dead ends and so forth. It's been really collaborative,
and I love seeing that. Has there been a culture clash between some of the sort of academics who have been working on this problem and now these like college kids who are interested in AI and on Discord sharing screenshots of their latest, you know, papyrus fragment? Has there been
any tension between those groups? Oh, yeah. Oh, yeah. So, a paperologist will go into the Discord and
say, I'm not even going to do a reading because these images are changing so fast and you guys don't understand papyrology. You have to be really careful. And then you got on the other side, you have the 21-year-old kid who's like, I think that's a pie and I think that is too. Hey, look at it.
This is important information for our listeners. If you're talking to a papyrologist, just don't tell them that you're looking at pie, okay? You want to say this might be pie. It might be worth considering that this is pie, but if you're too confident, you are going to trigger them and you'll find yourself in a flame war. That's right. It is kind of triggering to have people who don't know anything about the language and I would be one of them to say to them, hey, how about this reading?
You know, right? Yeah. Yeah. Let's talk about this prize winner, Luke Farator. He recently won $40,000 for deciphering a fragment of one of these scrolls. So, talk to us about his contribution. What was your reaction when he told you that he had made this discovery and what did he actually find? I almost fell off my chair when I saw on my cell phone this word popping out starting with pie. And if you've seen the image, the pie is so clear that I could have written it with crayon today. I mean, it's just right there. So, Luke discovers the word purple and I have to ask, Brent, what do you think was purple back in ancient times that they're talking about? Well, I have a pretty good guess because I'm as much of an armchair paparologist as anyone else and so, you know, I have Google. Yeah. Pliny the Elder wrote about purple coming from the mollusks, you know, as the Tyrian purple, the dye that the Romans used to make incredibly expensive clothing, robes that the wealthy wore. It could be referring to that. We also have these passages in the early Gospels where Jesus himself was on his way to being crucified and he was mocked
by being clothed in purple, which was a sign of somebody being powerful and wealthy and a king and it was mockery, right? Because he was actually going to be executed. So, those are the contexts we know, but what's really intriguing is that we don't know what this context is until we read a little bit more of the passages. Wow. Give us a sense of the progress that has been made so far.
So, this prize that Luke won, this $40,000 prize for deciphering this word, that was sort of an incremental prize on the way to the grand prize, which is $700,000, which is supposed to go to

the first team that reads four passages of text from the inside of two intact scrolls. This has to be done by December 31st. How close or far do you think we are from that? Do you think this prize will be won in the next few months before the end of the year? What's been revealed now as part of the first letters prize for just $40,000 is about 25% of what we need to win the grand prize. So, that's a substantial step. So, what I think is going to happen is that things are going to accelerate and it's an existence proof that the prize is winnable. So, the odds just went way up that even by December 31st, which I think is still a really aggressive deadline, actually, but even by then we may have people submitting and we have to evaluate that.
Why is it hard for these techniques to generalize? Now that you have an AI system that can read purple on one line, why can't you just feed all the hundreds of other scrolls into that same model and get back all of the other texts? What makes it hard to make these gains now that you have the basic technology? The way machine learning works is that you're sampling from a really big probability distribution of the world and you just don't have those samples when you begin the process. It's almost always incremental. So, if you're going to do the dog recognition, like Casey mentioned, and you have pictures of dogs from only one kind of camera, you probably be able to learn how to recognize a dog in that one camera, but there are differences between cameras, right? And you have to learn those differences. So, the next guy takes a picture with a different camera, can't find the dog. And it's exactly the same with this. The papyrus varies a lot in the way it was manufactured. It varies a lot in the way that you capture the x-ray images, the way that the papyrus twists and turns. There are a lot of variations. And until we can sample that distribution of all of those variations very well, we're just going to be ramping up until that point. It would be really funny if you finally decode this thing. And the first words are, this is very private. I hope no one ever finds it. I've heard many variations on that.
My worry is, what if it's just someone's grocery list? What if this is not a document of any significance? Right. Pick me up 14 purple grapes down at the bazaar on your way home, honey.
Now, what are Casey and my chances of winning the $700,000 grand prize by December 31st if we start today? Yeah. How dumb can you be and still contribute to this project? Well, I'm part of it.
So, I mean, there you go. I think you have as good a chance as anyone else relative to the amount of work that you're willing to put in. Brett, we're joking around, but I really want you to sort of close this out by just giving us a sense of the stakes here. Why is this a project that is worth hundreds of thousands of dollars of investors' money, that is worth the effort and labor of all of these researchers that are working on this crowdsourced science project? What is the pot of gold at the end of the rainbow? What do you hope humanity will gain from knowing what's in these scrolls? Well, we don't know what's there, and so revealing it is important. It's technical virtuosity in a way that we can feel good about instead of feeling frightened about. This is a good application of AI revealing the past, restoring the past.
So, there's kind of a redemptive piece here that says, I have something damaged and I'm redeeming it and we're going to be able to read it, pulling it back from oblivion. I think there's a broader thing here. When we go back far enough in history, we lose all the political boundaries and we lose all of the separators between us, and we find common humanity. And to be able to go back and think and talk and dialogue on those terms is tremendously important. And maybe this is just a proxy for us being able to do that. And it's really not about the material itself so much as our common humanity and being led to the important thing about our common humanity.

I love that. Well, I was going to ask a joke question, but then what Brent said was so lovely that now I don't want to. I think that's such a lovely note to end on.
Well, Brent, thank you so much for coming on and great to talk to you.
Thank you, Brent. This is great. Good to talk to you, Kevin. Casey, thank you.
Hard Fork is produced by Rachel Cohn and Davis Land. We're edited by Jen Poyant.
This episode was fact checked by Will Peischel. Today's show was engineered by Alyssa Moxley.
Original music by Rowan Nemistow and Dan Powell. Special thanks to Paula Schumann, Puiwing Tam, Nelga Logli, Kate LaPresti, Ryan Manning, Dylan Bergeson, and Jeffrey Miranda.
As always, you can email us at heartfork at nytimes.com.
You