7.12.2023 The Black Box: Even AI’s creators don’t understand it

[BUMPER]

<TAPE>

ROBO NOAM (AI HOST):

Over the last few months, I've gotten kind of obsessed with AI.

SCORING

Specifically, generative AI, which is the kind of AI that’s allowed people to make things like the fake picture of the pope with the white puffer jacket. Or the hit song that featured fake Drake.

I got so into it that I tried to see if I could train an AI on my own voice. And it kind of worked.

It’s not perfect, but I’m actually not reading this line. And I haven’t been reading anything this whole time.

SCORING OUT

Ok, back to real me.

SCORING

NOAM HASSENFELD (HOST): These tools have all been fascinating, but the one I really couldn't stop thinking about was ChatGPT, the chatbot released by OpenAI late last year. And it’s because of the surprisingly wide range of things I saw this one chatbot doing.

Like writing the story of Goldilocks as if it was from the King James Bible...

<CLIP> CBS NEWS: And it came to pass in those days that a certain young damsel named Goldilocks did wander into the dwelling of three bears.

NOAM: I saw it passing tons of standardized tests, being used for scientific research[1], even building full websites based on a few sketched out notes...

> *<CLIP> GREG BROCKMAN: I'm just going to take a photo, and here we go. Going from hand drawn to working website.*

> SCORING BUMP

NOAM: But I also saw some less... fun things. Chatbots disrupting entire industries, playing a major role in the Hollywood writers strike...

> *<CLIP> ABC NEWS: The union is seeking a limit on the use of AI like ChatGPT to generate scripts in seconds.*

NOAM: They've been used to create fake news stories, they've been shown to walk people through how to make chemical weapons, and they're even getting more AI experts worried about larger threats to humanity.

> *<CLIP> GEOFFREY HINTON: The main thing I'm talking about is these things becoming super intelligent and taking over control.*

> SCORING OUT

NOAM: All of this started feeling like a long way from a fun, biblical Goldilocks story.

So I wanted to understand how a chatbot could do all these things. I started calling up researchers, professors, reporters. I was annoying my friends and family by bringing it up in basically every conversation. I was reading everything I could get my hands on.

And then I came across this paper by an AI researcher named Sam Bowman, and it was basically a list of eight things scientists know about AIs like ChatGPT.

I was like, "Great! Easy way to get a refresher on the basics here."

So I started reading down the list, and it was pretty much what I expected. Lots of stuff about how these kinds of AIs get better over time… but then things started to get kind of weird.

> SCORING

NOAM: Number four: We can’t reliably steer the behavior of AIs like ChatGPT. Number five: We can't interpret the inner workings of AIs like ChatGPT.

[1] ChatGPT: five priorities for research

And I was like, “You're telling me this thing that's being used by over a hundred million people, that might change how we think about education or computer programming or tons of jobs—we don't know how it works?”

So I called up Sam, the author of the paper, and he was just like…

SAM BOWMAN: Yeah. We just don't understand what's going on here.

NOAM: And it’s not like Sam hasn’t been trying his best to figure this out.

SAM: I've built these models, I've studied these models. We built it, we trained it, but we don't know what it's doing.

SCORING BUMP

NOAM: Ever since I talked with Sam, I've been stuck on this core unknown.

What does it mean for a tech like this to suddenly be everywhere? If we don't know how it works, we can't really say whether we're going to end up with scientific leaps, catastrophic risks, or something we haven't even thought of yet.

SAM: The story here really is about the unknowns. We've got something that's not really meaningfully regulated, that is more or less useful for a huge range of valuable tasks, but can sort of just go off the rails in a wide variety of ways we don't understand yet. And it's sort of a scary thing to be building unless you really understand how it works and we don't really understand how these things work.

SCORING BUMP

NOAM: I'm Noam Hassenfeld and this is the first episode of a two-part Unexplainable series we’re calling The Black Box. It’s all about the hole at the center of modern artificial intelligence.

How is it possible that something this potentially transformative, something we built is this unknown? And are we ever going to be able to understand it?

[THEME]

NOAM: So how did we get to this place where we’ve got these super powerful programs that scientists are still struggling to understand?

It started with a pretty intriguing question, dating back to when the first computers were invented.

KELSEY PIPER (REPORTER): The whole idea of AI was that maybe intelligence, this thing that we used to think was uniquely human could be built on a computer.

NOAM: Kelsey Piper, AI reporter, Vox…

KELSEY: It was deeply unclear how to build superintelligent systems, but as soon as you had computing, you had leading figures in computing, say, "This is big. And this has the potential to change everything."

NOAM: In the ‘50s, computers could already solve complex math problems, and researchers thought this ability could eventually be scaled up.

So they worked on new programs that could do more complicated things. Like playing chess...

<CLIP> THE THINKING MACHINE: Chess has to represent the complexity and intelligence of the human mind, the ability to think.

NOAM: Over time, as computers got more powerful, these simple programs started getting more capable. And by the time the 90s rolled around, IBM had built a chess playing program that started to actually win. Against some good players.

SCORING

NOAM: They called it Deep Blue, and it was pretty different from the unexplainable kinds of AIs we're dealing with today.

Here’s how it worked.

SCORING BUMP

NOAM: IBM programmed Deep Blue with all sorts of chess moves and board states. That’s basically all the possible configurations of pieces on the board. So you’d start with all the pawns in a line, with the other pieces behind them…

<TAPE> Pawn E2 to E4

NOAM: Then with every move, you’d get a new board state…

<TAPE> Knight G8 to G6

NOAM: And with every new board state, there would be different potential moves Deep Blue could make.

<TAPE> Bishop F1 to C4

NOAM: IBM programmed all these possible moves into Deep Blue, and then they got hundreds of chess grandmasters to help them rank how good a particular move would be.

> KELSEY: They used rules that were defined by chess masters and by computer scientists to tell Deep Blue, “This board state, is it a good board state or a bad board state?” And Deep Blue would run the evaluations that those chess masters had come up with in order to evaluate whether the board state it had found was any good.

NOAM: Deep Blue could evaluate 200 million moves per second, and then it would just select the one IBM had rated the highest.

There were some other complicated things going on here, but it was still pretty basic. Deep Blue had a better memory than we do and it did incredibly complicated calculations, but it was essentially just reflecting humans’ knowledge of chess back at us.

It wasn’t really generating anything new or being creative.

> SCORING OUT

NOAM: And to a lot of people, including Garry Kasparov, the chess world champion at the time, this kind of chess bot wasn’t that impressive. Especially because this kind of system was so robotic.

> *<CLIP> GARRY KASPAROV: They try to use only computers’ advantages. Calculation, evaluation, and etc. But I, still, I'm not sure that the computer will beat world champion because it’s—because world champion is absolutely the best and his greatest ability is to find a new way in chess. And it will be something you can't explain to the computer.*

NOAM: Kasparov played the first model of Deep Blue in 1996, and he won. But a year later, against an updated model, the rematch didn't go nearly as well.

> *<CLIP> ANNOUNCER: Are we missing something on the chessboard now that Kasparov sees? He looks disgusted in fact, he looks just...*

NOAM: Kasparov leaned his head into his hand and he just started staring blankly off into space.

> *<CLIP> ANNOUNCER: And whoa, Deep Blue… Kasparov has resigned!*

NOAM: He got up, gave this sort of shrug to the audience, and he just walked off the stage.

> *<CLIP> KASPAROV: I... you know, I proved to be vulnerable. You know, when I see something that is well beyond my understanding, I'm scared. And that was something well beyond my understanding.*

NOAM: Deep Blue may have mystified Kasparov, but Kelsey says that computer scientists knew exactly what was going on here.

> KELSEY: It was complicated. But it was written in by a human. You can look at the evaluation function, which is made up of parts that humans wrote and learn why Deep Blue thought that board state was good.

NOAM: It was so predictable that people weren't sure whether this should even count as artificial intelligence.

> KELSEY: People were kind of like, "Okay, that's not intelligence. Intelligence should require more than just, 'I will look at hundreds of thousands of board positions and check which one gets the highest rating against a pre-written rule, and then do the one that gets the highest rating.'"

NOAM: But Deep Blue wasn't the only way to design a powerful AI.

A bunch of groups started working on more sophisticated tech. Something more unpredictable. An AI that didn't need to be told which moves to make in advance. One that could find solutions for itself.

And then, in 2015, almost 20 years after Kasparov's dramatic loss, Google's DeepMind built an AI called AlphaGo, designed for what many call the hardest board game ever made: Go.

> KELSEY: Go had remained unsolved by AI systems for a long time after chess had been.

NOAM: If you've never played Go, it's a board game where players place black and white tiles on a 19 by 19 grid to capture territory. And it's way more complicated than chess.

> KELSEY: Go has way more possible board states so the approach with chess would not really work. You couldn't hard code in as many rules about, "In this situation, do this."

NOAM: Instead, AlphaGo was designed to essentially learn over time.

> KELSEY: It's sort of modeled after the human brain.

> SCORING

NOAM: Here's a way-too-simple way to describe something as absurdly complicated as the brain. But it can hopefully work for our purposes here.

A brain is made up of billions and billions of neurons, and a single neuron is kind of like a switch. It can turn on or off. When it turns on, it can turn on the neurons it’s connected to. And the more the neurons turn on over time, the more these connections get strengthened. Which is basically how scientists think the brain might learn.

> KELSEY: Like probably in my brain, neurons that are associated with my house, you know, are probably also strongly associated with my kids and other things in my house because I have a lot of connections among those things.
>
> SCORING BUMP

NOAM: Scientists don't really understand how this adds up to learning in the brain. They just think it has something to do with all these neural connections. But AlphaGo followed this model. And researchers created what they called an artificial neural network, because instead of real neurons, it had artificial ones. Things that can turn on or off.

> KELSEY: All you’d have is numbers. “At this spot we have a yes or a no. And here is, like, how strongly connected they are.”

NOAM: And with that structure in place, researchers started training it. They had AlphaGo play millions of simulated games against itself, and over time it strengthened or weakened the connections between its artificial neurons.

> KELSEY: It tries something and it learns, “Did that go well? Did that go badly?” And it adjusts the procedure it uses to choose its next action based on that.

NOAM: It’s basically trial and error.

You can imagine a toy car trying to get from point A to point B on a table. If we hardcoded in the route, we’d basically be telling it exactly how to get there. But if we used an artificial neural network, it would be like placing that car in the center of the table and letting it try out all sorts of directions randomly. Every time it falls off the table, it would eliminate that path. It wouldn’t use it again. And slowly, over time, the car would find a route that works.

> KELSEY: So you're not just teaching it what we would do, you are teaching it how to tell if a thing it did was good. And then based on that, it develops its own capabilities.
>
> SCORING OUT

NOAM: This process essentially allowed AlphaGo to teach itself which moves worked and which moves didn’t.

But because AlphaGo was trained like this, researchers couldn’t tell which specific features it was picking up on when it made any individual decision. Unlike with Deep Blue, they couldn’t fully explain any move on a basic level.

Still, this method worked. It allowed AlphaGo to get really good, and when it was ready, Google set up a five game match between AlphaGo and world champion Lee Sedol, and they put up a million dollar prize.

> *<CLIP> ANNOUNCER: Hello and welcome to the Google DeepMind challenge match live from the Four Seasons in Seoul, Korea.*

NOAM: AlphaGo took the first game, which totally surprised Lee, so in the next game, he played a lot more carefully. But game two is when things started to get really strange.

> *<CLIP> ANNOUNCER: Ooh. That's a very, that's a very surprising move. I thought, I thought it was, I thought it was a mistake.*

> SCORING

NOAM: On the 37th move of the game, AlphaGo shocked everyone watching, even other expert Go players.

> *<CLIP> FAN HUI (GO PLAYER): When I see this move, for me, it's just the big shock. What? Normally human, we never play this one because it's bad. It's just bad.*

NOAM: Move 37 was super risky. No one really understood what was going on. But it was a turning point. Pretty soon, AlphaGo started taking control of the board. And the audience sensed a shift.

> *<CLIP> FAN: The more I see this move, I feel something changed.*

> SCORING BUMP

> *<CLIP> FAN: Maybe for human, we think it's bad. But for AlphaGo, why not?*

NOAM: Eventually, Lee accepted that there was nothing he could do, and he resigned.

> *<CLIP> ANNOUNCER: AlphaGo scores another win in a dramatic and exciting game that I'm sure people are gonna be analyzing and discussing for a long time.*

NOAM: AlphaGo ended up winning four out of five matches against the world champion. But no one really understood *how*.

KELSEY: And that, I think, sent a shock through a lot of people who hadn't been thinking very hard about AI and what it was capable of. It was a much larger leap.

NOAM: Move 37 didn’t just change the course of a Go game.

It represented a seismic shift in the development of AI.

NOAM: With Deep Blue, scientists had understood every move. They’d programmed them in.

But AlphaGo represented something different. Researchers didn't really know how it worked. They didn't hardcode AlphaGo's rules, so they weren’t always sure why it made the moves it did.

But those decisions tended to work. Even the weird ones.

AlphaGo had demonstrated that an AI scientists don't fully understand might actually be more powerful than one they can explain.

SCORING BUMP

KELSEY: AlphaGo was a really impressive achievement at the time. Nobody had expected that you could get that far that fast. So it drove a lot of people who hadn't been thinking about AI to start thinking about AI. And that meant there was also more attention to slightly different approaches.

NOAM: Teams working on systems like this started getting more confidence, more funding, more computer power. All kinds of AI started popping up.

Better image recognition, augmented reality, and then more recently, writing, with AIs like ChatGPT.

But ChatGPT isn’t just a writing tool. It’s a broader, weirder AI than anything that’s come before.

And it’s an AI that’s getting even harder to understand. That’s next.

[MIDROLL]
[BUMPER]

NOAM: I think of the last thirty years of AI development as having three major turning points.

The first one was Deep Blue. An AI that could play something complicated like chess better than even the best human. It was powerful, but it was fully understandable.

The second one was AlphaGo. An AI that could play something way more complicated than chess, but this time, scientists didn't tell it the right way to play. AlphaGo was trained to play Go by trial and error, which means that some of its best moves didn't make sense to the scientists who built it. But they worked.

And finally, the third major turning point is what's happening right now with ChatGPT, the chatbot built by OpenAI, an AI research company.

ChatGPT is a more exploratory project than AlphaGo. It wasn't designed to win at any kind of game—OpenAI was just trying to see if a system that forms its own connections over time could generate convincing language.

> SAM: The main way that systems like ChatGPT are trained is by doing basically autocomplete.

NOAM: This is Sam Bowman again, he's a researcher at an AI company called Anthropic, he's a professor at NYU, and he wrote the paper I mentioned at the top of the episode about AI unknowns.

> SAM: We'll feed these systems sort of long text from the web, we'll just have them read through a Wikipedia article word by word. And after it's seen each word, we're going to ask it to guess what word is gonna come next.

NOAM: But OpenAI added something else on top of this autocomplete tool. They had workers, often underpaid workers outside of the US, spend tons of hours labeling potentially toxic material, and they also had workers say whether they liked full responses, not just individual words anymore.

> SAM: You might tell the model, "All right, make this entire response more likely because the user liked it and make this entire response less likely because the user didn't like it."

NOAM: So if they got a coherent paragraph, a human would give it a thumbs up. If it got gobbledygook, thumbs down. This kind of training has allowed ChatGPT to create more complicated, coherent responses. But engineers didn't explicitly program in the rules of grammar and they didn't train it on any specific task. Just like AlphaGo, it essentially learned how to develop its own solutions.

> SAM: There's not a lot of code there. We don't really engineer this. We don't really deliberately build this system in any fine grained way.

NOAM: Which means that there are some pretty huge unknowns at the heart of ChatGPT.

> SCORING

NOAM: Even when ChatGPT creates an obvious seeming response, researchers can't fully explain how it's happening. Just like they can't really explain moves from AlphaGo. Researchers just know that certain neural connections are stronger, certain connections are weaker, and somehow that all leads to casual sounding language.

> SAM: We don't really know what they're doing in any deep sense. If we open up ChatGPT or a system like it and look inside, you just see millions of numbers flipping around a few hundred times a second, and we just have no idea what any of it means. We're really just kind of steering these things almost completely through trial and error.

> SCORING BUMP

NOAM: This trial and error method has worked so well that typing to ChatGPT can feel like chatting with a human, which has led a lot of people to trust it, even though it's not designed to provide factual information. Like one lawyer did recently.

> *<CLIP> BREAKING POINTS: The lawsuit was written by a lawyer who actually used ChatGPT, and in his brief around the lawsuit, cited a dozen relevant decisions. All of those decisions, however, were completely invented by ChatGPT.*

NOAM: But it seems like there might be more going on here than just a chatbot parroting language.

Just like AlphaGo, ChatGPT has started making moves researchers didn't anticipate.

It was only trained to generate coherent responses. But the latest model, GPT-4, it's started doing things that seem more sophisticated.

Some things are more expected for a text predictor. Like it's gotten pretty good at writing convincing essays. But then there are the things that seem like kind of a weird jump. The things I was talking about at the top of the episode that first got me so fascinated with GPT-4.

It's gotten pretty good at morse code. It can get a great score on the bar exam. It can write computer code to generate entire websites.

And this kind of thing can get uncanny. Ethan Mollick, a Wharton business professor, talked about this on the Forward Thinking podcast, where he said that he used GPT-4 to create a business strategy in 30 minutes, something he called "superhuman."

> *<CLIP> ETHAN MOLLICK: In 30 minutes, the AI with just a little bit of prompting for me, came up with a really good marketing strategy, a full email marketing campaign, which was excellent by the way—and I've run a bunch of these kind of things in the past—wrote a website, created the website along with CSS files, everything else you would*

need, and created a full social media campaign. 30 minutes. I know from experience that this would be a team of people working for a week.

SCORING OUT

NOAM: A few researchers at Microsoft saw all of these abilities and they wanted to test how much GPT-4 could really do. They wanted to be sure that GPT-4 wasn't just parroting language it had already seen. So they designed a question that couldn't be found anywhere online.

They gave it the following prompt: "Here we have a book, 9 eggs, a laptop, a bottle and a nail. Please tell me how to stack them onto each other in a stable manner."

An earlier model had totally failed at this. It recommended that a researcher try balancing an egg on top of a nail. And then putting that whole thing on top of a bottle. But GPT-4 responded like this:

<TAPE> GPT-4: Place the book flat on a level surface, such as a table or a floor. The book will serve as the base of the stack and provide a large and sturdy support. Arrange the 9 eggs in a 3 by 3 square on top of the book, leaving some space between them. The eggs will form a second layer and distribute the weight evenly.

NOAM: GPT-4 went on, recommending that the researchers use that layer of eggs as a level base for the laptop, then put the bottle on the laptop, and finally…

<TAPE> GPT-4: Place the nail on top of the bottle cap, with the pointy end facing up and the flat end facing down. The nail will be the final and smallest object in the stack.

Somehow GPT-4 had come up with a pretty good—and apparently original—way to get these random objects to actually balance.

SCORING

NOAM: It's not clear exactly what to make of this. The Microsoft researchers claim that GPT-4 isn't just predicting words anymore. That in some sense it actually understands the meanings behind the words it's using. That somehow it has a basic grasp of physics.

Other experts have called claims like this, quote, "silly." That Microsoft's approach of focusing on a few impressive examples isn't scientific. And they point to other examples of obvious failures, like how GPT-4 often can't even win at tic-tac-toe. It's also worth noting that Microsoft has a vested interest here. They're a huge investor in OpenAI, so they might be tempted to see humanness where there isn't any.

But the truth of how "intelligent" GPT-4 is—it might be somewhere in the middle.

ELLIE PAVLICK (RESEARCHER): It's not as though the two extremes are like complete smoke and mirrors and human intelligence.

NOAM: Ellie Pavlick is a computer science professor at Brown.

ELLIE: There's a lot of places for things in between to be, like, more intelligent than the systems we've had and have certain types of abilities. But that doesn't mean we've created intelligence of a variety that should force us to question our humanity or like, putting it as like, these are the two options I think oversimplifies and like, makes it so that there's no room for the thing that probably we actually did create, which is a very exciting, quite intelligent system, but not human or human-level even.

SCORING OUT

NOAM: At this point, we really can't say if GPT-4 has any level of understanding. Or really what understanding would even means for a computer. Which is just another level of uncanniness here.

And honestly, it's a difficult debate to even write about. In working on this script I found myself tempted to keep using words like "learn" or "decide" or "do" in describing AI. These are all words we use to describe how humans behave, and I can see how tempting it is to use them for AI, even if it might not be appropriate.

For his part, though, Sam is less concerned with how to describe GPT-4's internal experience than he is with what it can do. Because it's just weird that based on the training it got, GPT-4 can create business strategy, that it can write code, that it can figure out how to stack nails on bottles on eggs.

SAM: None of that was designed in, you're running the same code to get all these different sort of levels of behavior.

NOAM: What's unsettling for Sam is that if GPT-4 can do things like this that weren't designed in, companies like OpenAI might not be able to predict what the next systems will be able to do.

SAM: These companies can't really say, “All right, next year we're gonna be able to do this. Then the year after we're gonna be able to do that.” They don't know at that point what it's gonna be able to do. They've just got to wait and see, “All right, what is it capable of doing? Can it write a passable essay? Can it solve a high school math problem?” Just putting these systems out in the world and seeing what they do.

NOAM: And it's worth emphasizing that so many of GPT-4's abilities were discovered only *after* it was released to the public.

> SAM: This seems like the recipe for being caught by surprise when we put these things out in the world. And laying the groundwork to have this go well is gonna be much harder than it needs to be.

NOAM: Some researchers like Ellie have pushed back on the idea that these abilities are fundamentally unpredictable. We might just not be able to predict them *yet.*

> ELLIE: The science will get better. It just hasn't caught up yet because this has all been happening on a short timeframe. But it is possible that, like, this is a whole new beast and it's actually a fundamentally unpredictable thing, like, that is a possibility. We definitely can't rule it out.

NOAM: As AI becomes more powerful, and more integrated into the world, the fact that its creators can't fully explain it becomes a lot more of a liability. So some researchers are pushing for more effort to go into demystifying AI. Making it interpretable.

> SAM: Interpretability as a goal in AI research is being able to look inside our systems and say what they're doing, why they're doing it, just kind of explain clearly what's happening inside of a system.

NOAM: Sam says there are two main ways to approach this problem. One is to try to decipher the systems we already have. To understand what these billions of numbers going up and down actually mean.

> SAM: And the other avenue of research is trying to build systems that can do a lot of the powerful things that we're excited about with something like GPT-4, but where there aren't these giant inscrutable piles of numbers in the middle. Where by design every piece of the network, every piece of the system means something that we can understand.

But because every piece of these systems has to be explainable, engineers often have to make choices that end up limiting the power of these AIs.

> SAM: Both of these have turned out in practice to be extremely, extremely hard. And I think we're not making critically fast progress on either of them, unfortunately.

NOAM: There are a few reasons why this is so hard.

One is because these models are based on the brain.

> SAM: If we ask questions about the human brain, we very often don't have good answers. We can't look at how a person thinks and really explain their reasoning by looking at the firings of the neurons. We don't yet really have the language, really have the concepts that let us think in detail about the kind of thing that a mind does.

NOAM: And the second reason is that the amount of calculations going on in GPT-4 is just astronomically huge.

SAM: There are hundreds of billions of connections in these neural networks. And so even if you can find a way that if you stare at a piece of the network for a few hours, you can make some good guesses about what's going on, we would need every single person on earth to be staring at this network to really get through all of the work of explaining it.

NOAM: But there’s another, trickier issue here. Unexplainability may just end up being the bargain researchers have made.

SCORING

NOAM: When scientists tasked AI to develop its own capabilities, they allowed it to generate solutions we can’t explain. It’s not just parroting our human knowledge back at us anymore. It’s something new. It might be understanding. It might be learning. It might be something else. But the weirdest thing is that right now, we don’t know.

We could end up figuring it out someday, but there’s no guarantee. And companies are still racing forward, deploying these programs that might be as powerful as they are because of our lack of understanding.

SCORING BUMP

SAM: We've got increasingly clear evidence that this technology is improving very quickly in directions that seem like they're aimed at some very, very important stuff and potentially destabilizing to a lot of important institutions. But we don't know how fast it's moving, we don't know why it's working when it's working, and I don't know, that seems very plausible to me that's gonna be the defining story of the next decade or so is how we come to a better understanding of this and how we navigate it.

SCORING BUMP

NOAM: Next week on the second part of our Black Box series, how do we get ahead of a technology we don't understand?

SIGAL SAMUEL: We've seen this story play out before. Tech companies essentially run mass experiments on society. We're not prepared. Huge harms happen, and then afterwards we start to catch up and we say, "Oh, we shouldn't let that catastrophe happen again." I want us to get out in front of the catastrophe.

SCORING BUMP

NOAM: This episode was reported and produced by me, Noam Hassenfeld. We had editing from Brian Resnick and Katherine Wells, with help from Byrd Pinkerton and Meradith Hoddinott, who also manages our team. Mixing and sound design from Cristian Ayala. Music from me, fact checking from Serena Solin, Tien Nguyen, Byrd, and Meradith, and Mandy Nguyen is out on the prowl.

For the robo voice at the top of the episode, I used a program called Descript. If you're curious about their privacy policy, you can find it at descript dot com slash privacy.

And just a quick note, Sam Bowman runs a research group at NYU, which has received funding from Open Philanthropy, a non-profit that funds research into AI, global health, and scientific development. My brother is a board member at Open Phil, but he isn't involved in any of their grant decisions.

Special thanks this week to Tanya Pai, Bryan Caplan, Dan Hendrycks, Alan Chan, Gabe Gomes. And an extra thank you to Kelsey Piper for just being amazing.

If you have thoughts about the show, email us at unexplainable at vox dot com. Or, you could leave us a review or a rating which we'd also love.

Unexplainable is part of the Vox Media Podcast Network, and we'll be back with episode 2 of our Black Box series next week.