7.19.2023 The Black Box: In AI we trust?

[BUMPER]

SCORING - CYBORG

NOAM HASSENFELD (HOST): **I went to see the latest Mission Impossible movie this weekend, and it had a bad guy that felt very… 2023.**

<CLIP> MISSION IMPOSSIBLE 7: The entity has since become sentient.

NOAM: **An AI becoming superintelligent and turning on us.**

<CLIP> You're telling me this thing has a mind of its own?

NOAM: **And it's just the latest entry in a long line of supersmart AI villains.**

<CLIP> 2001: A SPACE ODYSSEY: Open the pod bay doors, Hal.

NOAM: Like in 2001: A Space Odyssey…

<CLIP> 2001: A SPACE ODYSSEY: I'm sorry, Dave. I'm afraid I can't do that.

NOAM: Or Ex Machina…

<CLIP> EX MACHINA: Ava, go back to your room!

NOAM: Or maybe the most famous example, Terminator.

<CLIP> TERMINATOR: They say it got smart. A new order of intelligence. It decided our fate in a microsecond.

SCORING OUT

NOAM: But AI doesn't need to be superintelligent in order to pose some pretty major risks.

Last week on the first episode of our Black Box series, we talked about the unknowns at the center of modern AI. How even the experts often don't understand how these systems work or what they might be able to do.

And it's true that understanding isn't necessary for technology. Engineers don't always don't understand exactly how their inventions work when they first design them.

But the difference here is that researchers using AI often can't predict what outcome they're going to get. They can't really steer these systems all that well.

And that's what keeps a lot of researchers up at night. It's not Terminator. It's a much likelier and maybe even stranger scenario.

It's the story of a little boat.

SCORING - A VIVID DREAM

NOAM: Specifically, a boat in this retro looking online video game.

It's called Coast Runners and it's a pretty straightforward racing game—there are power ups that give you points if your boat hits them, there are obstacles to dodge, there are these kind of lagoons where your boat can get all turned around. And a couple years ago the research company OpenAI wanted to see if they could get AI to teach itself how to get a high score on the game. Without being explicitly told how.

<CLIP> DARIO AMODEI (RESEARCHER): We were supposed to train a boat to complete a course from start to finish.

NOAM: This is Dario Amodei, he used to be a researcher at OpenAI, now he's the CEO of another AI company called Anthropic. And he gave a talk about this boat at a think tank called the Center for a New American Security.

<CLIP> AMODEI: I remember setting it running one day just telling it to teach itself. And I figured that it would learn to complete the course.

SCORING BUMP

NOAM: Dario had the AI run tons of simulated races over and over, but when he came back to check on it, the boat hadn't even come close to the end of the track.

<CLIP> AMODEI: What it does instead, this thing that's been looping, is it finds this isolated lagoon and it goes backwards in the course.

SCORING BUMP

NOAM: The boat wasn't just going backwards in this lagoon. It was on fire. Covered in pixelated flames. Crashing into docks and other boats. And just spinning around in circles.

SCORING BUMP

NOAM: But somehow the AI's score was going... up.

> *<CLIP> AMODEI: It turns out that by spinning around in this isolated lagoon in exactly the right way, it can get more points than it could possibly ever have gotten by completing the race in the most straightforward way.*

NOAM: When he looked into it, Dario realized that the game didn't award points for finishing first. For some reason, it gave them out for picking up power ups.

> *<CLIP> AMODEI: Every time you get one you increase your score, and they're kind of laid out mostly linearly along the course.*

NOAM: But this one lagoon was just full of these powerups, and the powerups would regenerated after a couple seconds, so the learned to AI time its movement to get these powerups over and over by exploiting this weird game design.

> SD BUMP

> *<CLIP> AMODEI: There's nothing wrong with this in the sense that we asked it to find a solution to a mathematical problem, how do you get the most points, and this is how it did it. But, you know, if this was a passenger ferry or something, you wouldn't want it spinning around, setting itself on fire, crashing into everything.*

> SCORING OUT

NOAM: This boat game might seem like a small, glitchy example, but it illustrates one of the most concerning aspects of AI.

It's called the alignment problem. Essentially, an AI's solution to a problem isn't always aligned with its designers' values. How they might want it to solve the problem.

And like this game, our world isn't perfectly designed. So if scientists don't account for every small detail in our society when they train an AI, it can solve problems in unexpected ways. Sometimes even harmful ways.

> *<CLIP> AMODEI: Something like this can happen without us even knowing that it's happening. Where our system has found a way to do the thing we think we want in a way that we really don't want.*

NOAM: The problem here isn't with the AI itself—it's with our expectations of it. Given what AIs can do, it's tempting to give them a task, and assume it won't end up in flames.

But despite this risk, more and more institutions, companies, and even militaries are considering how AI might be useful to make important real world decisions.

Hiring. Self-driving cars. **Even battlefield judgment calls.**

SCORING

NOAM: Using AIs like this can almost feel like making a wish with a super annoying, super literal genie. You've got the potential for a wish, but you need to be extremely careful.

<CLIP> WHAT WE DO IN THE SHADOWS: This reminds me of the tale of the man who wished to be the richest man in the world, but was then crushed under a mountain of gold coins.

SCORING BUMP

NOAM: I'm Noam Hassenfeld and this is the second episode of The Black Box, Unexplainable's series on the unknowns at the heart of AI.

If there's so much we still don't understand about AI, how can we make sure it does what we want, the *way* we want? And what happens if we can't?

[THEME]

NOAM: So given the risks here, that AI can solve problems in ways its designers don't intend, it's easy to wonder why anyone would want to use AI to make decisions in the first place.

It's because of all this promise. The positive side of this potential genie.

Here's just a couple examples. Last year, an AI built by Google predicted almost all known protein structures. It was a problem that had frustrated scientists for decades, and this development has already started accelerating drug discovery.

AI has helped astronomers detect undiscovered stars, it's allowed scientists to make progress on decoding animal communication, and like we talked about last week, it was able to beat humans at Go, arguably the most complicated game ever made.

In all of these situations, AI has given humans access to knowledge we just didn't have before.

KELSEY PIPER (REPORTER): So the powerful and compelling thing about AI, you know when it's playing Go, is sometimes it will tell you a brilliant Go move that you would never have thought of. That no Go master would ever have thought of. And that does advance your goal of winning the game.

NOAM: This is Kelsey Piper, she's a reporter for Vox who we heard from last episode. And she says that this kind of innovation is really useful, at least in the context of a game.

KELSEY (REPORTER): But when you're operating in a very complicated context like the world, then those brilliant moves that advance your goals might do it by having a bunch of side effects or inviting a bunch of risks that you don't know, don't understand, and aren't evaluating.

NOAM: Essentially, there's always that risk of the boat on fire.

SCORING - RIGHT ON TIME

NOAM: And we've already seen this kind of thing happen outside of a video game.

NOAM: Just take the example of Amazon, back in 2014.

KELSEY: So Amazon tried to use an AI hiring algorithm, looked at candidates and then recommended which ones would proceed in the interview process.

SCORING BUMP

NOAM: Amazon was developing an AI to help them with hiring. They fed it ten years worth of submitted resumes and they instructed it to find patterns that were associated with stronger candidates.

KELSEY: And then an analysis came out finding that the AI was biased. It had learned, you know, that Amazon generally preferred to hire men, so it was happily more likely to recommend Amazon men.

NOAM: Amazon never actually used this AI in the real world. They only tested it. But a report by Reuters found exactly which patterns the AI might have internalized.

<CLIP> REUTERS: The technology thought. "Oh Amazon doesn't like any resume that has the word 'women's' in it." An all women's university, captain of a women's chess club, captain of a women's soccer team.

NOAM: Essentially, when they were training their AI, Amazon hadn't accounted for their own flaws in how *they'd* been measuring success internally. Kind of like how OpenAI hadn't accounted for the way the boat game gave out points based on power ups, not based on who finished first.

KELSEY: And of course when Amazon realized that, they took the AI out of their process.

NOAM: But it seems like they might be getting back in the AI hiring game. According to an internal document obtained by former Vox reporter Jason Del Rey, Amazon's been working on a

new AI system for recruitment. At the same time, they've been extending buyout offers to hundreds of human recruiters.

And these flaws aren't unique to hiring AIs. The way AIs are trained has led to all kinds of problems. Take what happened with Uber in 2018, when they didn't include jaywalkers in the training data for their self-driving cars, and then a car killed a pedestrian.

> *<CLIP> CBS NEWS: Tempe, Arizona Police say 49-year-old Elaine Herzberg was walking a bicycle across a busy thoroughfare, frequented by pedestrians Sunday night. She was not in a crosswalk.*

NOAM: And a similar thing happened a few years ago with a self-training AI Google used in its photos app.

> *<CLIP> INDIA TODAY: The company's automatic image recognition feature in its photo application identified two black persons as gorillas and in fact, even tagged them as so.*

NOAM: According to former Google employees, this may have happened because Google had a biased data set. They may just not have included enough Black people.

> KELSEY: The worrying thing is if you're using AIs to make decisions and the data they have reflects our own biased processes, like a biased justice system that sends some people to prison for crimes where it lets other people off with the slap on the wrist or a biased hiring process, then the AI is gonna learn the same thing.
>
> SCORING OUT

NOAM: But despite these risks, more companies are using AI to guide them in making important decisions.

> KELSEY PIPER: This is changing very fast. Like there are a lot more companies doing this now than there were even a year ago. And there will be a lot more in a couple more years.

NOAM: Companies see a lot of benefits here. First, AI is cheap. Systems like ChatGPT are currently being heavily subsidized by investors, but at least for now, AI is way cheaper than hiring real people.

> KELSEY: If you want to look over thousands of job applicants, AI is cheaper than having humans screen those thousands of job applicants. If you wanna make salary decisions, AI is cheaper than having a human whose job is to think about and make those salary decisions. If you wanna make firing decisions, those get done by algorithm because it's easier to fire who the algorithm spits out than to have human judgment and human analysis in the picture.

NOAM: **Kelsey's seen this kind of thing firsthand.**

> KELSEY: I was helping a friend apply for a bunch of minimum wage jobs a while ago, and a bunch of them wanted these videos where you talk about how cheerful you are about the job and then an AI would look at them to, like, analyze your face and decide to hire you. And that's just infuriating. Like the idea that you have to make this video and then the computer is deciding if you're a cheerful enough person. It's obnoxious.

NOAM: **But whether it's obnoxious or not, Kelsey says a lot of companies might be ok with this kind of thing or with even bigger risks like potential boats on fire.**

> KELSEY: It's so much cheaper that that's like a good business trade-off. And so we hand off more and more decision making to AI systems for financial reasons.

NOAM: The second reason behind this push to use AI to make decisions is because it could offer a competitive advantage.

> KELSEY: Companies that are employing AI—in a very winner-take-all capitalist market—they might outperform the companies that are still relying on expensive human labor. And the companies that aren't are much more expensive, so fewer people want to work with them, and they're a smaller share of the economy. And you might have huge, like, economic behemoths that are making decisions almost entirely with AI systems.

NOAM: But it's not just companies. Kelsey says competitive pressure is even leading the military to look into using AI to make decisions.

> KELSEY: I think there is a lot of fear that the first country to successfully integrate AI into its decision making will have a major battlefield advantage over anyone still relying on slow humans. And that's the driver of a lot in the military, right? If we don't do it, somebody else will, and maybe it will be a huge advantage.

NOAM: This kind of thing may already have happened in actual battlefields. In 2021, a UN panel determined that an autonomous Turkish drone may have killed Libyan soldiers without a human controlling it or even ordering it to fire. And lots of other countries including the US are actively researching AI-controlled weapons.

> KELSEY: You don't wanna be the people, you know, still fighting on horses when someone else has invented fighting with guns. And you don't wanna be the people who don't have AI when the other side has AI. So I think there's this very powerful pressure not just to figure this out, but to have it ready to go.

NOAM: And finally, the third reason behind the push toward AI decision making is because of the promise we talked about at the top. AI can provide novel solutions for problems humans might not be able to solve on their own.

Just look at the Department of Defense. They're hoping to build AI systems that, quote, "function more as colleagues than as tools." And they're studying how to use AI to help soldiers make extremely difficult battlefield decisions. Specifically when it comes to medical triage.

> *<CLIP> KATHLEEN FISHER (DARPA DIRECTOR): I'm going to talk about how we can build AI-based systems that we would be willing to bet our lives with and not be foolish to do so.*

NOAM: AI has already shown an ability to beat humans in war game scenarios, like with the board game Diplomacy. And researchers think this ability could be used to advise militaries on big decisions like strategic planning. Cybersecurity expert Matt Devost talked about this on a recent episode of On The Media:

> *<CLIP> MATT DEVOST: I think it'll probably get really good at threat assessment. I think analysts might also use it to help them through their thinking, right? They might come up with an assessment and say, tell me how I'm wrong. So I think there'll be a lot of unique ways in which the technology is used in the intelligence community.*

NOAM: But this whole time, that boat on fire possibility is just lurking.

> SCORING

NOAM: One of the things that makes AI so promising—the novelty of its solutions—is also the thing that makes it so hard to predict.

Kelsey imagines a situation where AI recommendations are initially successful, which leads humans to start relying on them uncritically, even when the recommendations seem counterintuitive. Humans could assume the AI sees something they don't, so they follow the recommendation anyway.

We've already seen something like this in a game context with AlphaGo like we talked about last week, so the next step is imagining it happening in the world.

> SCORING BUMP

NOAM: And we know that AI can have fundamental flaws. Things like biased training data or strange loopholes engineers haven't noticed. But powerful actors relying on AI for decision making might not notice these faults until it's too late.

> KELSEY: And this is before we get into the AI, like, being deliberately adversarial.

NOAM: This isn’t the Terminator scenario with AI becoming superintelligent and wanting to kill us. The problem is more about humans. And our temptation to rely on AI uncritically.

KELSEY: This isn't the AI trying to trick you. It's just the AI exploring options that no one would've thought of that get us into weird territory that no one has been in before. And since they're so untransparent, we can't even ask the AI, "Hey, what are the risks of doing this?"

SCORING BUMP

NOAM: So if it’s hard to make sure that AI operates in the way its users intend, and more institutions feel like the benefits of using AI to make decisions might outweigh the risks, what do we do?

What can we do?

KELSEY: There's a lot that we don't know, but I think we should be changing the policy and regulatory incentives so that we don't have to learn from a horrible disaster, and so that we, like, understand the problem better and can start making progress on solving it.

NOAM: How to start solving a problem that you don't understand. After the break.

[MIDROLL]
[BUMPER]

NOAM: So here's what we know.

Number one: Engineers often struggle to account for all the details in the world when they program an AI. They might want it to complete a boat race and end up with a boat on fire. A company might want to use it to recommend a set of layoffs only to realize that the AI has built-in biases.

Number two: Like we talked about in the first episode of this series, it isn’t always possible to explain why modern AI makes the decisions it does, which makes it difficult to predict what it’ll do.

And finally, number three: We've got more and more companies, financial institutions, even the military considering how to integrate these AIs into their decision making. There’s essentially a race to deploy this tech into important situations, which only makes the potential risks more unpredictable.

SCORING - THE BUTTERFLY DRANK TOO MUCH

NOAM: Unknowns on unknowns on unknowns.

So what do we do?

> SIGAL SAMUEL: I would say at this point it's sort of unclear.

NOAM: Sigal Samuel writes about AI and ethics for Vox, and she’s about as confused as the rest of us here. But she says there's a few different things we can work on.

The first one is interpretability. Just trying to understand how these AIs work, but like we talked about last week, interpreting modern AI systems is a huge challenge..

> SIGAL: Part of how they're so powerful and they're able to give us info that we can't just drum up easily ourselves is that they're so complex. So there might be something almost inherent about lack of interpretability being an important feature of AI systems that are gonna be much more powerful than my human brain.

NOAM: So interpretability may not be an easy way forward.

But some researchers have put forward another idea. Monitoring AIs by using *more* AIs. At the very least just to alert users if AIs seem to be behaving kind of erratically.

> SIGAL: But it's a little bit circular because then you have to ask, "Well, how would we be sure that our helper AI is not tricking us in the same way that we're worried our original AI is doing?"

NOAM: So if these kinds of tech solutions aren’t the way forward, the best path could be political, just trying to reduce the power and ubiquity of certain kinds of AI.

A great model for this is the EU, which recently put forward some promising AI regulation.

> SIGAL: The European Union is now trying to put forward these regulations that would basically require companies that are offering AI products in, like, especially high-risk areas to prove that these products are safe.

NOAM: This could mean doing assessments for bias, requiring humans to be involved in the process of creating and monitoring these systems, or even just trying to reasonably demonstrate that the AI won't cause harm.

> SIGAL: We've unwittingly bought this premise that they can just bring anything to market when we would never do that for other similarly impactful technologies, like, think about medication, you’ve gotta get your FDA approval, you've gotta jump through these hoops. Why not with AI?

SCORING OUT

NOAM: Why not with AI?

Well, there's a couple reasons regulation might be pretty hard.

First, AI is totally different from something a medication that the FDA would approve. The FDA has clear, agreed upon hoops to jump through: clinical testing. That's how they assess the dangers of a medicine before it goes out into the world.

But with AI, researchers often don't know what it can do until it's been made public. And if even the experts are often in the dark, it may not be possible to prove to regulators that AI is safe.

The second problem here is that even aside from AI, big tech regulation doesn't exactly have the greatest track record of really holding companies accountable. Which might explain why some of the biggest AI companies like OpenAI have actually been publicly calling for more regulation.

> SIGAL: The cynical read is that this is very much a repeat of what we saw with a company like Facebook—now Meta—where people like Mark Zuckerberg were going to Washington DC and saying, "Oh, yes, we're all in favor of regulation. We'll help you, we wanna regulate too."

NOAM: When they heard this, a lot of politicians said they thought Zuckerberg's proposed changes were vague and essentially self-serving. That he just wanted to be seen supporting the rules. Rules which he never really thought would hold them accountable…

> SIGAL: Allowing them to regulate in certain ways, but where really they maintain control of their data sets. They're not being super transparent and having external auditors. So really they're getting to continue to drive the ship and make profits, while creating the semblance that society or politicians are really driving the ship.

NOAM: Regulation with real teeth seems like such a huge challenge that one major AI researcher even wrote an op-ed in Time Magazine calling for an indefinite ban on AI research. Just shutting it all down.

But Sigal isn't sure that's such a good idea.

> SIGAL: I mean, I think we would lose all the potential benefits it stands to bring. So, drug discovery, you know, cures for certain diseases, potentially huge, economic growth that if it's managed wisely—big if—could help alleviate some kinds of poverty. I mean, at least potentially it could do a lot of good. And so you don't necessarily wanna throw that baby out with the bath water.

NOAM: At the very least, Sigal does want to turn down the faucet.

> SIGAL: I think the problem is we are rushing at breakneck speed towards more and more advanced forms of AI when the AIs that we already currently have, we don't even know how they're working.

NOAM: When ChatGPT launched, it was publicly deployed faster than any other technology in history. Twitter took 2 years to reach a million users. Instagram took 2 and a half months. ChatGPT took 5 days.

And there are so many things researchers learned ChatGPT could do only after it was released to the public.

> SIGAL: There's so much we still don't understand about them so what I would argue for is just slowing down.

NOAM: Slowing down AI could happen a whole bunch of different ways.

> SIGAL: So you could say, you know, we're gonna stop working on making AI more powerful for the next few years, right? We're just not gonna try to develop AI that's got even more capabilities than it already has.

NOAM: AI isn't just software. It runs on huge, powerful computers. It requires lots of human labor. It costs tons of money to make and operate, even if those costs are currently being subsidized by investors.

So the government could make it harder to get the types of computer chips necessary for huge processing power. Or it could give *more* resources to researchers in academia, who don't have the same profit incentive as researchers in industry.

> SIGAL: You could also say, "All right, we understand researchers are gonna keep doing the development and trying to make these systems more powerful, but we're gonna really halt or slow down deployment and, like, release to commercial actors or whoever."

NOAM: Slowing down the development of a transformative technology like this, it's a pretty big ask, especially when there's so much money to be made. It would mean major cooperation, major regulation, major complicated discussions with stakeholders that definitely don't all agree. But Sigal isn't hopeless.

> SIGAL: I'm actually reasonably optimistic.
>
> SCORING - FIVE YEARS IN A MINUTE

SIGAL: I'm very worried about the direction AI is going in. I think it's going way too fast. But I also try to look at things with a bit of a historical perspective.

NOAM: Sigal says that even though tech progress can seem inevitable, there is precedent for real global cooperation.

SIGAL: We know historically there are a lot of technological innovations that we could be doing that we're not because societally it just seems like a bad idea, human cloning or like certain kinds of genetic experiments, like humanity has shown that we are capable of putting a stop or at least a slowdown on things that we think are dangerous.

NOAM: **But even if guardrails are possible, our society hasn't always been good about building them when we should.**

SIGAL: The fear is that sometimes society is not prepared to design those guardrails until there's been some huge catastrophe like Hiroshima, Nagasaki, just horrific things that happen and then we pause and we say, "Okay, maybe we need to go to the drawing board." Right? That's what I don't want to have happen with AI. We've seen this story play out before. Tech companies or technologists 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. Hopefully that will be by slowing down the whole AI race. If people are not willing to slow down, at least let's get in front by trying to think really hard about what the possible harms are and how we can use regulation to really prevent harm as much as we possibly can.

SCORING BUMP

NOAM: If there's one thing I've learned while working on this series it's that no one knows what's going on here.

But that's not the way I often hear people talking about it.

There are extreme claims all over the place.

There are the people that are sure AI poses an extreme, immediate, existential risk.

And then there are people that are absolutely confident that this catastrophic risk from AI is overblown. That this is just another hype bubble about to pop.

But without projecting into the far future, I keep coming back to what's true right now. Scientists have created AIs that can pass tests of higher-order thinking, that have shown potential to help companies and militaries with strategy, that are already helping advance scientific and medical research.

But they still have these blind spots. GPT-4 even struggles to play tic-tac-toe.

And the people creating these systems, they can't really explain why.

Which can feel pretty risky if powerful institutions start to rely on these systems for guidance.

So slowing down AI might not be enough. We should be honest about what we don't know here. So hopefully the powerful actors who are actually shaping this future—companies, research institutions, governments—will take this seriously. And remember to be skeptical of all of this potential.

Because if we're open about how little we really know, we can start to wrestle with the biggest question here: are all of these risks worth it?

SCORING BUMP

NOAM: **That's it for our Black Box Series.** This episode was reported and produced by me, Noam Hassenfeld. We had editing from Brian Resnick and Katherine Wells, with help from Meradith Hoddinott, who also manages our team. Mixing and sound design from Vince Fairchild, with help from Cristian Ayala. Music from me, and fact checking from Tien Nguyen. Mandy Nguyen is potentially a werewolf. We're not sure.

And Byrd Pinkerton sat in the dark room at the octopus hospital, listening to the prophecy:

SCORING OUT

"Three thousand years ago, we were told that one day there would be an Octopocalypse. And that only a bird would be able to ensure the survival of our species. You are that bird, Pinkerton."

SCORING

Special thanks this week to Pawan Jain, José Hernández-Orallo, Samir Rawashdeh, and Eric Aldrich.

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 would love too.

Unexplainable is part of the Vox Media Podcast Network, and we'll be back in your feed next week.