EXHIBIT 126



HIGHLY CONFIDENTIAL
ATTORNEYS' EYES ONLY

Transcript of Dario Amodei, Ph.D.

Date: January 8, 2026
Case: Concord Music Group -v- Anthropic

Planet Depos
Phone: 888.433.3767
Email: transcripts@planetdepos.com
www.planetdepos.com

1

IN THE UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL., Plaintiffs, v. ANTHROPIC PBC, Defendant. — Case Number 5:24-cv-03811-EKL SVK

HIGHLY CONFIDENTIAL--ATTORNEYS' EYES ONLY

VIDEOCONFERENCED and VIDEO-RECORDED
DEPOSITION of
DARIO AMODEI, Ph.D.
San Francisco, California 94111
Thursday, January 8, 2026

Reported Stenographically by:
MARY J. GOFF
CSR No. 13427
WA CSR No. 21030779
Job No. 614290
PAGES 1-158

2

HIGHLY CONFIDENTIAL--ATTORNEYS' EYES ONLY

VIDEOCONFERENCED and VIDEO-RECORDED DEPOSITION of DARIO AMODEI, Ph.D., Volume I, taken on behalf of Plaintiff Publishers, Oppenheim + Zebrak, LLP, at WilmerHale, 50 California Street, Suite 3600, San Francisco, California 94111, beginning at 1:44 p.m. and ending at 4:30 p.m., on Thursday, January 8, 2026, before MARY J. GOFF, California Certified Shorthand Reporter No. 13427 and WA CSR No. 21030779.

3

APPEARANCES:

For Plaintiff Publishers
Oppenheim + Zebrak, LLP
BY: NICHOLAS C. HAILEY
Attorneys at Law
4530 Wisconsin Avenue, NW
5th Floor
Washington, DC 20016
202-480-2999
nick@oandzlaw.com

For Defendant Anthropic PBC
WilmerHale
BY: SONAL MEHTA
Attorneys at Law
2600 El Camino Real
Suite 400
Palo Alto, California 94306
sonal.mehta@wilmerhale.com
628-235-1026

ALSO PRESENT: Natalie Naugle, Brian Israel, In-house counsel Anthropic
Ellen Stephano, Esquire, WilmerHale (remote Boston)
Videographer: Addison McNaughton

4

INDEX

WITNESS		EXAMINATION
DARIO AMODEI, Ph.D.		
Volume I		
	ATTORNEY HAILEY	6
	ATTORNEY MEHTA	--

NUMBER	DESCRIPTION	PAGE
Exhibit 246	1-29-24 Slack message Anthropic_0000605700-785	9
Exhibit 247	Transcript: Ezra Klein Pages 1-18	45
Exhibit 248	Exhibit 40 Anthropic_0000462964-968	89
Exhibit 249	Anthropic Plan for 2023 Anthropic_0000463241-255	108
Exhibit 250	Why Anthropic CEO Dario Amodei Spends So Much Time Warning of AI's Potential Dangers	136
Exhibit 251	AI Misalignment Risk Anthropic_0000478585	142
PREMARKED EXHIBITS:		
Exhibit 235	1-8-2023 email Anthropic Plan	123

Pages 45-48

45

This comment didn't -- I -- again, I don't remember this comment. But from looking at it and from the context, my belief is that this content did not intend to address copyright one way or another.

Q Well, sitting here today, sir, do you agree that copyright infringement falls within any

[redacted]

A I would describe it as a -- you know, I -- again, I would describe copy -- generating copyrighted output that's not within the bounds of fair use as, you know, something that the model should not do.

(Exhibit 247 was marked for identification and is attached to the transcript.)

Q (BY ATTORNEY HAILEY) Mr. Amodei, I'm showing you what we have marked as Exhibit 247 to your deposition.

ATTORNEY MEHTA: Do you have more copies of this one?

ATTORNEY HAILEY: Yes.

ATTORNEY MEHTA: Thank you.

Q (BY ATTORNEY HAILEY) It's a New York Times article titled "The Ezra Klein Show: Transcript:

46

Ezra Klein Interviews Dario Amodei," dated April 12, 2024.

Do you see that?

A I see that, yes.

Q And you recall being interviewed by Ezra Klein for the New York Times?

A I -- I recall being interviewed by Ezra Klein. I don't recall the content of the interview in detail, but I recall being interviewed by Ezra Klein.

Q And you agree this appears to be a transcript published in the New York Times of that interview, correct?

A This -- this appears to be a transcript of -- of that interview. Of course, as with all transcripts, there could be errors in this transcript. It may not be exactly what I said, but yes, this is a transcript.

Q And you're truthful and accurate when you give interviews?

A I attempt to be truthful and accurate, yes.

Q Please turn your attention to page 15, please.

Do you see two-thirds of the way down the

47

page, the question from Ezra Klein that begins:

But there's also the question of training data and the intellectual property that is going into things like CLAUDE, like GPT, like Gemini?

A Yes.

Q And I want to focus on the first sentence of your response, Mr. Amodei, which appears at the bottom of page 15. Are you with me at the bottom of page 15?

A I am with you on the bottom of page 15.

Q You respond:

I think everyone agrees the models shouldn't be verbatim outputting copyrighted content. First of all did I read that correctly?

A You read that correctly.

Q And that's a true statement that "everyone agrees AI models shouldn't be verbatim outputting copyrighted content"?

A Again, you know, if -- if I were given full time to consider this question, I would -- I would make that statement.

48

But of course, many statements that are made without caveats on, you know, interviews where you have to respond quickly, you know, if -- if you had the fullness of time, you would make that -- you would make that same statement with a number of caveats or elaborate and describe the nuance, so I would say, again, you know, models should not be verbatim outputting copyright cont -- outputting copyrighted content, again, subject -- subject to the bounds of fair use.

For example, fair use allows, you know, short snippets of content to be -- to be -- to be outputted. Everyone would -- you know, everyone would agree that that is the case.

Q And what's that statement based on, sir?

A That statement is based on my -- my personal understanding of how fair use works.

Q You didn't say that in this interview with Ezra Klein, correct?

A That's -- that's right. There isn't time in an interview to offer every caveat or every piece of nuance.

Q Instead you simply said:

I think everyone agrees that models shouldn't be verbatim

Pages 65-72

65

basically match strings -- match strings of word and block copyrighted content.

There was -- there's something called constitutional AI where CLAUDE trains models to be in line with the constitution. And that constitution at one -- that constitution at -- at one point included the idea that it should have -- the model's output should have respect for copyright.

And there was something called Prompt Shield -- I think Prompt Shield may have been developed later -- which provided -- which basically looked at user inputs and tried to detect whether those user inputs were asking for copy -- copyrighted information in an improper way and -- and blocked -- blocked outputs that -- that appeared to be improper.

That's my -- that's my best recollection of it, but I don't remember very clearly. I don't remember when each of those safeguards was put in place or which ones were put in place before launch and then improved after or -- or which launch -- which safeguards were put in place after.

Q So CopyFairy is an output filter; is that right?

66

A CopyFairy is -- yeah, it's -- it's an output matcher and blocker, correct.

Q And it was later referred to, I believe, as CCS?

A I have never heard that term.

Q Content Compare System, have you heard that term?

A I have never heard that term.

[redacted]

ATTORNEY MEHTA: Again, I'm going to caution you: With respect to any communications that you had with counsel, you can answer, if you know independent of that.

A If I know -- if I know independent of that? I don't believe I -- I don't believe I know anything independent of that, no.

Q (BY ATTORNEY HAILEY) So you don't know [redacted]

A Something -- something that I do know is that -- the way those filters work isn't that they add or don't add specific content. The way those

67

filters work is they include broad parameters such as -- include broad parameters such as length.

So I can be quite confident there was no specific decision with we said: Oh, yeah, we didn't include content X, and now we're -- now we're including content X.

CopyFairy doesn't work that way. CopyFairy isn't content specific. CopyFairy has broad parameters like length.

It is -- it is possible that as those parameters are adjusted, the balance of false positives and false -- false negatives could change, but -- but CopyFairy is not discriminative of content or specific content.

Q But you don't know what the specific parameters in terms of length of output for CopyFairy were at any given time?

A I -- I -- I knew at one point. I don't remember.

Q And you don't know when or even if Publishers' song lyrics were ever added to the output filter specifically, correct?

A I know that CopyFairy was updated -- CopyFairy was updated over time as we learned more, and that was generally -- generally, you know,

68

responsive to problems in the filter, but I couldn't give you a list of all the complete updates that were given.

Q The second safeguard you mentioned was constitutional AI?

A Correct.

Q Anthropic didn't update its constitution to include a principle against substantially reproducing copyrighted content until [redacted]; is that correct?

A I don't know. There was a point at which our constitution was published, and there was a point at which it was included in the training of the model.

I -- I'm not sure a date makes sense here. It would only -- it would only make sense -- because when something is included in the constitution, it then gets included in the next model.

So the right way to say it would be something was included in, say, CLAUDE 2.1, not CLAUDE 2.2.

You may be referring -- I'm not sure. You may be referring to the date in which the constitution is published, but that would -- that would not be the same as the date that -- the date

69

that the -- the content of that constitution was included in the model, nor would that be the -- the same date as the date that model was released.

Q Mr. Amodei, what I'm referring to here is an interrogatory response that Anthropic provided as part of this litigation where it stated that it added to its constitution, a principle against substantially reproducing copyrighted content on or around [redacted].

Do you have any reason to disagree with that date?

A I don't have -- I don't have any -- I don't have any specific information to that. If an Anthropic -- if an Anthropic witness said that, I would assume they're telling the truth.

Q And that is months after Anthropic launched CLAUDE to the public in May -- in March 2023, correct?

A That's -- that's right. I said that some of these -- I -- I said specifically in -- in the answer to the question you said before, that some of these guardrails were added before we launched our model and some of these guardrails were added after we launched our model.

In the -- in the case of

70

constitutional AI, the -- the method, constitutional AI, was not -- not available until a few months before the date you cited. And the methods for training with constitutional AI were -- because it was such a new method, were not reliable for a while after that.

So this was a new technological method. It's not like it was laying around and could have been used much earlier than it was.

Q And the third safeguard you mentioned was Prompt Shield?

A That's correct.

Q And Anthropic didn't implement Prompt Shield until [redacted]; is that correct?

A Again, I -- I cited these three things and said that I didn't know when each of these -- I said that I didn't know when each of these things were implemented. I don't know or recall the specific date when it was implemented.

Q No reason to disagree that it was [redacted] when Prompt Shield was first implemented?

A I -- I have no information about this one way or another.

Q And you could have implemented those safeguards earlier, correct?

71

A No. I disagree with that.

As I said, on constitutional AI, the method of constitution AI was a new research method. You know, it was invented -- it was invented by Anthropic. It was a pioneering safety method that allowed models to be steered in a way that was not possible literally anywhere else in the world before.

That method was developed and published, I believe it was, around the end of 2022. And then it took a substantial amount of time to actually get the method to work well, to actually develop the constitution, to decide what should be in the constitution, to iterate in the constitution.

As I -- as I -- as I said before, it is not necessarily the case -- in fact, it is probably not the case that, you know, at -- at arbitrary times, this -- this method could have been -- could have been implemented because it was a newly developed research method.

Q But the first version of Anthropic's constitution did not include a principle against substantially reproducing copyrighted content, correct?

A I don't -- I don't actually -- I don't

72

actually recall or -- I don't actually recall or know what you're referring to.

Again, the first version of the constitution could be a document we made that we never actually trained models with. So again, there's many different dates you could mean here.

Q You don't disagree that Anthropic launched CLAUDE to the public before adding a copyright principle to its constitution?

A Again, you have -- you have stated that this principle was added on [redacted] Yeah, if that's true, then -- then that principle was added afterwards.

But that doesn't mean that we used constitutional AI or that the method of constitutional AI was reliable or that it was anything other than experimental method at the time it was first developed.

ATTORNEY HAILEY: Counsel, we have been going for over an hour. I could use a bio break, so why don't we take a break and we can come back and --

ATTORNEY MEHTA: That's fine. But I think we're going to do one break and then finish. So if you want to take that now, that's fine. And if not,

Pages 125-136

125

Q So if we turn back to Exhibit 235, you see that same language, "buy books and scientific papers"?
A That's -- that's -- that's correct.
Q But you understand Mr. Kingsland to be commenting on that language in this --
A Again, I -- I don't see these things together, but this could very well have been on comment on that, yes. Sure.
Q Mr. Kingsland writes here:
I think we need to have more specific approach here? Others might have a clearer vision, but I don't know what our strategy is.
Did I read that correctly?
A Correct.
Q And then you respond to Mr. Kingsland:
"I think we have many places from which we could, in principle, buy these things, but it's a huge legal/practice/business slog."
A Correct.
Q Did I read that correctly?
A You read that correctly.
Q And when you say "buy these things,"

126

you're referring to buying or licensing copyrighted content like books and scientific papers for AI training --
A Specifically --
Q -- right?
A -- to books and scientific papers, yes.
Q And you write on January 8, 2023, that:
"There are many places from which Anthropic could, in principle, buy these things,"
correct?
A That's correct.
Q In other words, it's possible to pay for the copyrighted content that you use for AI training?
A Again, it --
ATTORNEY MEHTA: Objection to the form.
Go ahead.
A -- it's possible to pay for books and scientific papers that are not freely and legally available.
Q (BY ATTORNEY HAILEY) And you don't limit it that way in the comment you wrote in this document, correct?
A Well, I think the reference "these things

127

here" probably refers to the books and scientific papers. That's -- that would be the commonsense meaning of these -- these comments.
Q But you don't say in this document that you're only referring to books and scientific papers that are not freely and legally available, correct?
ATTORNEY MEHTA: Objection to the form.
A Yeah, so -- so generally books and scientific papers are the kinds of things that are typically not available on the Internet, so scientific papers would be behind a paywall. And most -- most books are not available on websites, so these are -- these are categories that are frequently or usually not freely and legally available.
So this discussion, which I vaguely remember, was specifically about how to buy content that was not freely and legally available.
Q (BY ATTORNEY HAILEY) And books and scientific papers are copyrighted content, correct?
A Books -- books and scientific papers are often copyrighted content. Sometimes they're in the public domain, but yeah, sometimes books and scientific papers are copyrighted content.
Q But the books and scientific papers that

128

you're referring to here are examples of copyrighted content, correct?
[redacted]
Q And with respect to that copyrighted content, you say, "There are many places from which Anthropic could, in principle, buy these things," right?
A That's right.
Q But you go on to say that, "It would be a huge legal/practice/business slog to do so," correct?
A Yes. This is the same sentiment that I think was expressed in one of the other documents you showed me, which is that the -- you know, Anthropic at that time was a small company. And you know, I think it was less than, I don't know, 70 people or something at the time. I don't know if that number is exactly correct, but that's my approximate recollection.
You know, we had -- we had a very small

129

set of people in charge of -- you know, in charge of business functions. So doing -- doing -- this is just a factual statement that, you know, doing a lot of effort to -- to buy these -- you know, to do a business deal where we -- where we buy this kind of content would be a lot of effort.

That's just a true factual statement.

Q And you didn't want to undertake that effort, correct?

A I -- no. I believe that we explored this effort. I -- I don't have an exact recollection of it. But similar to this comment in the same doc about "buying access to data from large players like Google and Amazon," my recollection is that we -- we did are conversations with Google and with Amazon about buying this data, but those conversations never worked out.

Q And you ultimately chose not to undertake the huge legal/business/ -- excuse me. I'll start over.

You ultimately chose not to undertake the huge legal/practice/business slog of buying access to this copyrighted content for AI training?

A Well -- well --

ATTORNEY MEHTA: Objection, misstates the prior testimony.

A -- yeah -- yeah, again, we -- we -- there was a significant amount of slog. Like, talking to these large companies and trying to work out deals or agreements with them was a lot of effort, even though it didn't work out in the end. So -- so we did attempt to -- we did attempt to undertake these -- these activities.

But just to be clear, like, you know, the alternative to -- you know, there wasn't some alternative that was, like: If we don't buy these things, we'll obtain them in some sinister other way.

The decision was to either buy these things and have them or not buy them and not have them, which is a perfectly leg -- that is a perfectly legitimate decision.

You can choose to buy something and have it or you can choose to not buy something and not have it.

Q (BY ATTORNEY HAILEY) Licensing training data requires identifying the rightsholder, reaching out to them, negotiating a license, and paying for that license, correct?

A It requires a number of steps. That could -- that could include those steps, yes.

Q And that's the type of legal/practice/business slog you're referring to, correct?

A I'm just -- I'm just referring to the complexity of doing a business deal at a small start-up. That's right.

Q And complex deals are expensive, right?

A I wouldn't say that they're expensive. They're time-consuming. And -- and they're time-consuming and they use human labor. That's right.

Q And if you could get information on the Internet without paying for it, you didn't want to undertake the effort of this huge legal/practice/business slog --

A Yeah --

Q -- right?

A -- unfortunately --

ATTORNEY MEHTA: Objection to the form.

A -- yeah, unfortunately, I think what you're saying is a misappropriation and ignores what I just said before, which is that this wasn't about whether to get information freely available on the Internet, which our contention is that we don't have to buy versus buying that information.

Those two aren't the same thing. There are two pools here. There's pools of information that are freely and legally available on the Internet. This -- this, we used.

Then there is different information that could be bought or licensed. The choice isn't to buy -- the choice isn't to buy or license information or obtain it on the Internet in some illegitimate way.

The -- the choice is to -- to buy or license data and have that data or not buy and license that data and not have that data and only use data that's freely available on the Internet.

So there was not a choice between buying data and saying: No, this is too much of a slog; we're going to obtain it in some other illegitimate way.

That was not the choice at issue here, and it is misrepresenting to portray that as being the choice at issue.

Q (BY ATTORNEY HAILEY) If you could find information on the Internet and download it, you didn't pay for it; is that right?

ATTORNEY MEHTA: Objection to the form;

133

asked and answered multiple times.
Counsel, you're also getting into territory you know you're not allowed to get into based on the court's order. And we have a limited amount of time, so I suggest you move on.
ATTORNEY HAILEY: Counsel, the speaking objections are improper. We'll raise this with the court.
A I -- I believe you already -- I believe you already asked this question, and I don't have anything to add to my answer.
Q (BY ATTORNEY HAILEY) Mr. Amodei, I do need an answer to that question. I'm going to ask it one more time.
If you could find information on the Internet and download it, you didn't pay for it; is that right?
A My --
ATTORNEY MEHTA: Asked and answered.
A -- my view is that information that is freely and legally available on the Internet in an analogy to the way that humans consume information and use it to produce their own ideas, if information is freely and legally available on the Internet, then that information can be ingested by

134

AI models similar to how a human can absorb AI models into their brain and use it -- and use it to generate their own ideas.
That is our contention.
Q (BY ATTORNEY HAILEY) And when you say that "the information is freely and legally available on the Internet," that's just your own view as to whether it's legal or not to use it for AI training?
ATTORNEY MEHTA: I'm going to instruct you not to answer with respect to any advice you had from counsel. You can answer with respect to your own view as to whether --
A That's right.
ATTORNEY MEHTA: -- it's appropriate.
A That's right. Again -- again, I'll say -- I'll say again my own view is that -- you know, there's this analogy to humans. Right.
As I have said many times, I have a background as a neuroscientist. I have thought about artificial neural nets. I have thought about natural neural nets. Humans learn by absorbing information. They take inspiration from many different sources of information, and then they generate their own ideas.
My view is that AI models should work in

135

the same way. If there's information that is not freely -- freely and legally available on the Internet, then AI -- then humans have to pay for it. Therefore, AI models should have to pay for a copy of it -- a copy of it as well.
If the information can just be read by a human without -- without breaking any laws -- I'm not talking about laws relating to copyright. I'm just talking about can I actually view the information without violating any laws.
If a human can view it without violating any laws, then an AI model can view it and absorb it without violating any laws. That is my -- again, without reference to any -- any advice of -- of lawyers, that -- that was my commonsense view of the situation based on these unexplored legal issues many years ago.
Q (BY ATTORNEY HAILEY) Mr. Amodei, you agree that AI models like CLAUDE have both potential benefits and risks, correct?
A That's -- that's correct. I think that's absolutely central to the -- to the -- to the views at Anthropic, that this technology can have great benefits and be transformative for society and that it also has risks. And that -- that Anthropic

136

exists to try and manage the benefits and the risks, correct.
Q And I want to focus on the risks --
A Correct.
Q -- of AI technology for a moment.
You have called AI "the greatest risk humanity has ever faced"; is that correct?
A I don't -- I don't remember that quote. It's possible I could say that. I think it's the greatest benefit humanity has faced and also the greatest risk humanity has faced, yes.
(Exhibit 250 was marked for identification and is attached to the transcript.)
Q (BY ATTORNEY HAILEY) I'm showing you what I have marked as Exhibit 250.
A Okay.
Q It's an article that I printed off of CBSNews.com titled "Why Anthropic CEO, Dario Amodei, Spends So Much Time Warning of AI's Potential Dangers," dated November 16, 2025 --
A Yes. This --
Q -- by Anderson Cooper.
A -- this looks like it was associated with a 60 Minutes interview. That's right.
Q And you gave an interview to Anderson

Pages 141-148

141

"that doesn't sound good" in the middle of the page?
A Yes.
Q Then skip down three lines.
A Oh. Skip down three lines. Okay.
Q You say:
"Because AI is a new technology, just like it's going to go wrong on its own, it's also going to be misused by -- you know, by criminals and malicious state actors."
A Correct.
Q What do you mean when you say "AI will go wrong on its own"?
A "Because AI is a new technology, just like it's going to go wrong on its own, it's going to be misused by criminals and malicious state actors."
So here I believe I'm referring to a distinction between two types of risks. One risk is that the autonomous behavior of an AI system may be threatening to humans. So this is the risk that -- that an AI system may autonomously behave in ways that are dangerous to humans.
For example, a robot might physically injure a human being or -- or an AI system, in the

142

example I give to Anderson Cooper, might -- because it -- you know because our process of controlling the training of the AI system is not precise, the -- the AI system may blackmail someone or exfiltrate -- or exfiltrate information or, you know, may interact with a user in a -- in a manipulative fashion or, you know, in the long run could -- could be kind of a -- you know, a grand threat to the security of human -- of human civilization.
That's what I mean by AI going wrong on its own. And that's contrasted to the idea that AI would be misused by humans. So there are two dangers here. There is the danger of the AI system itself and there's the danger of humans misusing the Ai system.
I believe that's what I meant there.
(Exhibit 251 was marked for identification and is attached to the transcript.)
Q (BY ATTORNEY HAILEY) Mr. Amodei, I'm showing you what I have marked as Exhibit 251 to your deposition.
A Okay.
Q It's a document bearing the Bates Number Anthropic 478585.
Do you recognize this document?

143

A I do not -- hold on. I do not specifically recognize this document, no.
Q You can see that it's a memo you wrote, titled "AI Misalignment Risk: How to Improve Our Research Agenda," by Dario Amodei, dated October 20, 2022?
Do you see that?
A Yes.
Q And I want to start at the top of the first page, page 585.
A Okay.
Q You write in the first sentence:
"I think that existential risk from AI misalignment is real and that we should worry about it a lot."
Did I --
A That's right.
Q -- did I read that correctly?
A That's right.
Q Under "Summary," you write:
"I'm very worried about AI risk. I think it is the greatest risk humanity has ever faced. My very imprecise estimate is that

144

there's a 25 percent chance that AI leads to a very bad outcome."
Did I read that correctly?
A Yes.
Q At the time you wrote this document, you believed that AI was the greatest risk humanity has ever faced?
A That is correct.
Q And is that still your belief today, sir?
A Yeah, I believe AI is the greatest risk humanity has ever faced and the greatest benefit humanity has ever faced.
Q And if you could turn to the next page, page 586, please. Do you see the first bullet in the middle of the page?
A Yes.
Q You write:
"The current trajectory of AI development makes me think that there's a 20 to 30 percent chance that AGI will lead to the destruction or permanent stunting of humanity, as AGI seems more powerful/dangerous than nuclear weapons.

145

And looking back on The Cold War period, 20 to 30 percent seems like a reasonable guess for the probability of nuclear weapons causing a huge setback to civilization during that period. AGI may be more risky because it is a less stable technology. Did I read that correctly?

A That's correct.

Q And when you say "AGI" you're referring to Artificial General Intelligence, correct?

A Correct.

Q So you say in this Anthropic memo that you believe there's a 20 to 30 percent chance that AGI will lead to the destruction or permanent stunting of humanity, right?

A I -- I want to be specific here. I said that on the current trajectory. In other words, in the beginning -- in the way that companies other than us are currently behaving.

Anthropic, that probability is not a fixed amount. It's not like we're pressing a button and saying: Oh, we pressed this button, and there's a 20 -- 20 to 30 percent chance that -- that humanity

146

is destroyed.

Anthropic exists to try to change that trajectory, to try and change the development of AI in -- in a direction that reduces or eliminates the risks and increases the benefit.

The whole purpose of Anthropic's mission is to identify the risks and act in such a way that reduces those risks, to develop the technology in a different and alternative way.

Q Anthropic is developing AI models just like your competitors, OpenAI and Google and others, correct?

ATTORNEY MEHTA: Form.

A No. No. False. False. It's actually false.

Anthropic is not developing AI models that are just like its competitors. Anthropic measures its models for misalignment risk. And as of the most recent models, Anthropic models perform substantially better than other models on misalignment risk.

Anthropic is substantially more transparent, as in the Anderson Cooper interview that you described, in being honest about these -- these -- being honest about the risks of these

147

models in disclosing the risks of these models.

Anthropic has also advocated for regulation, for legislation, forcing companies to disclose the risk of these models and take safety measures against the risk of these models.

Now, that's actually a very costly action. Talking about the risk of AI models is an incredibly costly action. Because when you talk about the risks of AI models, people then blame you for this risk.

In fact, lawyers will even try to use the fact that you have talked about these risks to -- to try and catch you in "gotchas" and say: Oh, yeah, I -- I can -- I can exploit this. Because -- because you said that there was a risk, because you were honest about this risk, a lawyer can use this to try and say: Oh, you're -- you're a bad person because this risk exists.

But even though lawyers do that -- even though lawyers do that, I am still willing to talk about these risks, because it's so important to society to do it.

Q Is it your testimony under oath, Mr. Amodei, that Anthropic's CLAUDE.AI models are risk free?

148

A No. No, model can be risk free. The situation is similar to self-driving cars. There's no such thing as a risk-free self-driving car.

What there is is self-driving cars that are much safer than human drivers. Self-driving cars can be much more dangerous than human drivers. They can be equally dangerous or they can be safer.

Because the world is unpredictable -- for example, a duck could run in front of the road or there could be snow or there could be a hailstone or -- or another driver could crash into the car, creating risk that's completely exogenous and not related to the self-driving car.

So in the real messy world, there's no such thing as being risk free. It's impossible to be risk free. What we aim for is -- is to reduce risk over and over again to add more nines to the level of reliability.

There's no such thing as a risk-free world. Anthropic aims to reduce risk as much as possible. That is all that is physically possible. It is not physically possible to reduce risk to zero.

Q Turn to page 587, please, Mr. Amodei. Do you see the paragraph beginning "in closing"?