Exhibit 6



Planet Depos®
We Make It Happen™

HIGHLY CONFIDENTIAL
ATTORNEYS' EYES ONLY

Transcript of Jared Kaplan, Ph.D.

Date: November 21, 2025
Case: Concord Music Group -v- Anthropic

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

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL.,

Plaintiffs,

vs. Case No. 5:24-cv-03811 EKL-SVK

ANTHROPIC PBC,

Defendant.

HIGHLY CONFIDENTIAL - ATTORNEYS' EYES ONLY

VIDEO DEPOSITION OF JARED KAPLAN, Ph.D.

San Francisco, California

Friday, November 21, 2025

Volume 1

STENOGRAPHICALLY REPORTED BY:
REBECCA L. ROMANO, RPR, CSR, CCR
California CSR No. 12546
Nevada CCR No. 827
Oregon CSR No. 20-0466
Washington CCR No. 3491

JOB NO. 609877

PAGES 1 - 365

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL.,

Plaintiffs,

vs. Case No. 5:24-cv-03811 EKL-SVK

ANTHROPIC PBC,

Defendant.

DEPOSITION OF JARED KAPLAN, Ph.D., taken on behalf of the Plaintiffs, at Wilmer Cutler Pickering Hale and Dorr LLP, 50 California Street, Suite 3600, San Francisco, California, commencing at 10:04 a.m., Friday, November 21, 2025 before REBECCA L. ROMANO, a Registered Professional Reporter, Certified Shorthand Reporter, Certified Court Reporter.

APPEARANCES OF COUNSEL

For the Plaintiffs:

OPPENHEIM + ZEBRAK LAW, LLP
BY: NICK HAILEY
Attorney at Law
4530 Wisconsin Avenue
5th Floor
Washington, D.C. 20016
(202) 480-2999
nick@oandzlaw.com

and

BY: BRET MATERA
Attorney at Law
461 5th Avenue
19th Floor
New York, New York 10017
(212) 951-1156
bmatera@oandzlaw.com

/////

APPEARANCES OF COUNSEL(cont'd)

For the Defendant:

WILMER CUTLER PICKERING HALE AND DORR LLP
BY: TAYLOR GOOCH
Attorney at Law
50 California Street
Suite 3600
San Francisco, California 94111
(628) 235-1006
taylor.gooch@wilmerhale.com

and

BY: ALYSSA MILSTEAD(via Web videoconference)
Attorney at Law
2600 El Camino Real
Suite 400
Palo Alto, California 94306
(650) 858-6000
alyssa.milstead@wilmerhale.com

/////

Pages 13-16

13

Q. That was also a copyright infringement case against your company, Anthropic; is that correct?
A. I don't know exactly what -- how -- how to define that -- terms.
Maybe you can clarify what you mean.
Q. What's your understanding of the subject matter of the Bartz case?
A. I don't know what to say about it. It was a lawsuit brought against Anthropic.
Q. Do you have an understanding what the claims were in that case?
A. I don't really recall the details.
Q. How long ago was that deposition, approximately?
A. I'm not sure.
Q. What's your best estimate?
A. I think it was in the last year.
Q. Was the testimony that you gave in Bartz, in your Bartz deposition, truthful and accurate?
A. I believe that my testimony in that prior deposition was accurate, to the best of my knowledge, as indicated at the time.
Q. Did you give testimony in your Bartz deposition regarding Anthropic's Claude AI models?

14

A. I believe that deposition included some discussion of Anthropic's AI models, including our commercial model, Claude.
Q. Did you give testimony in your Bartz deposition about the datasets that Anthropic uses to train those models?
A. Can you define what you mean by the datasets Anthropic means -- uses to train the models.
Q. How do you train your models?
A. Contemporary -- are you talking about current versions of Claude?
Q. Dr. Kaplan, Anthropic trains its models using datasets of a written text, correct?
A. I think that's kind of a vague presupposition. What I would say is that contemporary models like Claude, circa, say, this year, 2025, are trained very broadly in two phases that are distinct.
One is commonly referred to as pretraining; the other as post-training, RL, fine-tuning, all kind of different words for the second phase.
The first phase of training, called pretraining, involves deciding on a AI model

15

architecture, which is a way of processing inputs and making probabilistic predictions about likely outputs.
The models that Anthropic specifically focuses on, although there are many other in the industry, are what called autoregressive models, which means that given some set of inputs, they predict a next possible token or output. And specifically, they predict a probability distribution over those outputs.
The way that these models are trained involves tokenizing various data into little bits and pieces. And those little bits and pieces might be components of images, they might be components of code, they might be components of text that have been preprocessed, shuffled. There's a very complicated process involved in tokenization.
And then the models are trained with a -- typically, a variant of a cross-entropy loss to -- given the first N tokens in a given sequence to predict the N plus first token. And that happens in parallel for sort of all N across the -- the sequence. So the model learns how -- given some set of context, how to predict the next token.
That's the first phase of training.

16

The second phase of training typically involved some kind of -- often some kind of supervised learning over, typically, prior high-quality data generated by Claude. And then it involves reinforcement learning, where the model is in some way asked to perform a helpful task, or potentially it's confronted with a task that -- that is inappropriate, that might be harmful, might cause the model to hallucinate. And the model is given some complex form of reward depending on the quality of its output on many, many, many axes.
And this process of reinforcement learning teaches the model to be useful at a variety of tasks: summarizing documents, writing code, performing analyses, at this point doing things like using Microsoft Excel or using a computer, understanding images.
And this results in some, typically, final model that -- that is actually sort of useful for -- for interaction, can use tools, et cetera.
Q. Dr. Kaplan, we have limited time for the deposition today, so it's going to be very important that you listen to the question that I ask and answer directly the question that I ask.
Do you understand that?

Pages 85-88

85

people are welcome to have skepticism about that, and I think that that creates a number of challenges.

Just as the Industrial Revolution led to a lot of changes in everything about society, government, social roles, weapons, risks from technology, pollution, I think that it's possible that changes due to the advance of AI will have a similar impact on the world and that that suggests a wide variety of potential risks, many of which Anthropic, in one form or another, have -- have discussed, everything from misuse of AI by bad actors or terrorists, weaponization of AI, misuse of AI at a smaller scale, AI impacts on misinformation bias, mental health, child safety.

Furthermore, AI might have an impact on employment. It might change what roles are well-compensated and what are not in society. That may create a lot of political change.

There's risks from AI having -- being concentrated in certain countries, in certain languages, associated with only certain cultures.

There are risks from AI propagating and playing a larger and larger role. For example, there are risks from so-called prompt injection,

86

where if AI becomes quite widespread and is taking actions on behalf of individuals and companies, there may be incentives for bad actors to try to set up environments that trick AI in -- into doing things that are bad.

For example, maybe AI could be tricked into giving away someone's -- someone's bank account information or something like that.

There's a lot of risks from fraud, the risks associated with cyber attacks.

And then there's sort of a general question of how society and government should respond to these developments if, in fact, they occur, and how we can sort of effectively raise awareness to meet the moment without compromising potential benefits of AI.

So I think that that's a flavor of what Anthropic means by "safety," although I don't expect it's necessarily complete.

Q. So a lot to unpack there. Let me just ask a few follow-up questions.

Fair to say there are a number of serious risks that come with AI technology?

MR. GOOCH: Objection. Misstates prior testimony.

87

THE DEPONENT: I think it is the case that AI -- different AI systems, especially as they become more and more powerful, could have a number of risks.

Q. (By Mr. Hailey) And some of those risks are very serious?

MR. GOOCH: Asked and answered.

THE DEPONENT: I think that there could be serious risks from AI.

Q. (By Mr. Hailey) And AI's risks include bias and discrimination?

MR. GOOCH: Asked and answered.

THE DEPONENT: I believe in my prior testimony, I mentioned bias as a possible risk that some have pointed to with respect to AI.

Q. (By Mr. Hailey) And you agree that bias and discrimination is a potential risk of AI?

MR. GOOCH: Asked and answered. And vague.

THE DEPONENT: I think that there are many, many risks from AI. I mentioned bias as one that has been widely discussed.

Q. (By Mr. Hailey) Among AI's many, many other risks, one of the risks you identified was cybersecurity threats, correct?

88

MR. GOOCH: Same objections.

THE DEPONENT: I identified a large variety of possible -- possible impacts from AI, and cyber is one of them.

Q. (By Mr. Hailey) Another risk of AI is misinformation and manipulation; is that correct?

MR. GOOCH: Same objections.

THE DEPONENT: I mentioned that AI could have many risks and that it's possible it can be misused to propagate misinformation.

Q. (By Mr. Hailey) Another potential harm of AI is in connection with data privacy violations and misuse of personal information; is that correct?

MR. GOOCH: Same objections.

THE DEPONENT: I mentioned risks from prompt injection.

Q. (By Mr. Hailey) And I'm asking you about AI's potential risk from data privacy violations and misuse of personal information.

You agree that that's a risk of AI?

MR. GOOCH: Same objections.

THE DEPONENT: I think that that's a pretty vague risk that's associated with almost any digital technology.

Pages 93-96

93

MR. HAILEY: I'm not -- not sure that it's an appropriate question for you to be asking. So you can ask the question. We have the question.
MR. GOOCH: Are you going to follow up?
MR. HAILEY: Taylor, you've asked the question.
MR. GOOCH: Okay. I understand you're refusing to answer.
Can you confirm whether or not O&Z has used any of the transcripts or notes or summaries of the Bartz depositions in any way for any purpose in this case?
MR. HAILEY: So I will tell you, I don't think it's proper for all the reasons we've discussed before for you to be asking us about our case preparation. The detail -- this is all privileged information you're asking about.
We've, I think, been clear with you that we've complied with the respective protective orders, and I don't think we need to tell you anything more than that about what our case -- casework has been.
So I don't think those are appropriate questions for you to ask, and I don't think we're going to respond further.

94

MR. GOOCH: So to be clear, you're refusing to answer the questions?
MR. HAILEY: That's correct, because I don't think they are appropriate questions. So we are not obligated to answer them.
MR. GOOCH: Okay. We disagree.
Let's go off the record.
THE VIDEOGRAPHER: Okay. We are going off the record. The time is 11:45 a.m.
(Recess taken.)
THE VIDEOGRAPHER: We are back on the record. The time is 11:57 a.m.
MR. GOOCH: And I'll just note that on the break, I asked counsel if he had identified the email he represented existed, and he said he hadn't checked.
Q. (By Mr. Hailey) Welcome back, Dr. Kaplan.
Did you have a nice break?
A. It was good.
Q. So before the break, we were talking about the risks of AI technology like Claude.
Do you recall those questions?
A. I recall that we were talking about speculative possible risks from AI that are

95

wide-ranging, especially if AI becomes more and more capable over time.
Q. And one of the risks of AI is actually an existential risk; is that correct?
A. Many people who are concerned about AI are concerned about risks that they deem existential if AI becomes significantly more intelligent and capable than any person.
Q. And when we say existential, that means that AI poses a risk or threat of human extinction; is that correct?
A. I think these words are used in a variety of different ways. But people are concerned about AI potentially replacing people as seen in the movie "Terminator."
Q. And do you agree that there's a potentially existential threat of AI?
A. I think that it makes sense to take seriously the possibility that AI will become as or more intelligent than people and even human experts in every area.
And I think if that occurs, it will be the first time in human history that we have been, in some sense, supplanted as the smartest beings on Earth, and that -- I think it's widely agreed that

96

a lot of our success and our ability to have a modern technological culture where people flourish, where lifespans are far greater than they were in the past, people mostly have their basic needs met, is due to intelligence and collaboration.
So if there's another set of entities on Earth that are just as smart or significantly smarter than us, that can think faster than us, then those entities could, if they were organized and if they were malevolent, accrue power and potentially supplant human agency.
Q. And those entities that you're referring to are -- that's AI technology?
A. Yeah, I am referring to very, very advanced AI.
Q. Let's go back to your role at Anthropic.
You were -- you were describing earlier your current roles and responsibilities as chief science officer at Anthropic.
Do you recall that testimony?
A. I recall.
Q. Do you have any other current roles and responsibilities at Anthropic that you haven't already described?
A. I don't know.

Pages 137-144

137

A. Well, I'm just worried about confusion.
I mean, for example, in many cases when Anthropic models are trained, the data that is used is generated by Anthropic models.
So I'm just trying to make it clear that it's a pretty complicated process involving many teams. Some of that data, much of that data, depending on how you define various aspects of the process and what you mean by "data," is actually generated from Anthropic models.
Q. Okay. So Anthropic selects or generates the data that it uses to train its AI models, correct?
MR. GOOCH: Form.
THE DEPONENT: Anthropic has -- as of now, November 2025, has many teams in different parts of the company that work on architectural and data choices in order to train Claude.
Q. (By Mr. Hailey) What specific characteristics of data does Anthropic take into account when deciding what data to use in training its Claude models?
MR. GOOCH: Form. And vague.
THE DEPONENT: Can you be more specific.
Q. (By Mr. Hailey) What does Anthropic

138

consider when selecting the data to use for AI training?
MR. GOOCH: Same objections.
THE DEPONENT: Well, Anthropic generally tries to choose hard, sort of as challenging as possible reinforcement learning environments that teach Claude how to do tasks that are as useful as possible to Anthropic employees and -- and customers.
Q. (By Mr. Hailey) Anthropic wants large volumes of training data; is that true?
MR. GOOCH: Vague. Form.
THE DEPONENT: Well, continuing in the line of the -- my prior response, that's not true.
In post-training, I would say the most important aspect is having very challenging, very complex tasks for Claude to learn to perform. So that includes things like agentic coding tasks, computer use tasks, summarizing complex documents, using -- using a variety of tools, summarizing documents, learning to -- learning to search over corpora, all sorts of other things.
And I think currently I would say, while it's good to have more examples, I think it's often most important that these examples be challenging,

139

because current models near the frontier are very capable.
Q. (By Mr. Hailey) Is it true that Anthropic wants very diverse sets of training data?
MR. GOOCH: Vague. Form.
THE DEPONENT: I'm not sure what you mean.
Q. (By Mr. Hailey) You don't understand that question?
MR. GOOCH: Asked and answered. Vague. And form.
THE DEPONENT: I mean, you're not being specific enough for it to make sense for me to provide a -- a concise answer.
Q. (By Mr. Hailey) I understand that Anthropic derives its training data from many diverse sources on the Internet; is that accurate?
MR. GOOCH: Objection. Form.
THE DEPONENT: As I said, I mean, you haven't really clarified what kind of data you mean.
The kind of data that we have just been talking about in the prior responses is sort of hard tasks for Claude to perform. And those are mostly not sourced from the Internet.

140

Q. (By Mr. Hailey) So let's -- let's focus on pretraining.
I understand that Anthropic derives training data for pretraining from many diverse sources on the Internet; is that accurate, Dr. Kaplan?
MR. GOOCH: Form.
THE DEPONENT: Anthropic sources a wide variety of text and text images and code for pretraining.
Q. (By Mr. Hailey) And those diverse sources from which Anthropic sources text for pretraining include websites and blogs, books, research papers, source code, patents, news media, advertising, and song lyrics; is that correct?
MR. GOOCH: Form. And vague.
I also invite you to give the witness the context, i.e., the declaration you're reading from, for him to answer with the relevant context.
THE DEPONENT: If you're reading from my declaration, then Anthropic does source a wide variety of different kinds of data, including from the Internet.
MR. HAILEY: And, Counsel, I will remind you that these speaking objections are improper.

141

I'm reading from my outline. I'm --
MR. GOOCH: I see the declaration sitting in a folder next to you.
What's in the manila folder to your right if you're trying to make a representation that's not what it is?
MR. HAILEY: Are -- that's completely improper that you're looking over the table at exhibits. You -- I'm reading from my outline, I'm not reading from a declaration --
MR. GOOCH: To the extent --
MR. HAILEY: -- and I'm asking the witness a question.
You need to stop interrupting. You need to stop coaching the witness. We heard the witness immediately adopt the -- the instruction that you improperly gave on this.
So, Taylor --
MR. GOOCH: And I think you're happily trying to ask questions without any context to try to confuse the witness. I think that's improper. And --
MR. HAILEY: Taylor, you can --
MR. GOOCH: -- I'll ask you --
MR. HAILEY: -- you can object, and --

142

MR. GOOCH: Counsel, if you're going to respond to me, I'm going to respond to you. Happy to go back and forth. It's all going to count towards your seven hours.
If you want to just ask questions, I invite you to ask questions.
Q. (By Mr. Hailey) Dr. Kaplan, I will ask my question again.
You testified -- you testified that Anthropic sources training data from a wide -- strike that.
You referenced a wide range of sources for Anthropic's training data.
And I will ask you again, those diverse sources include websites and blogs, books, research papers, source code, patents, news media, advertising and song lyrics; is that accurate?
MR. GOOCH: Same objections.
THE DEPONENT: I'm not sure.
Q. (By Mr. Hailey) You don't know?
A. Anthropic sources a wide variety of information and data for pretraining. I don't know whether each of the individual examples you gave is accurate. By giving a list of five or ten examples, I think it's easy to provide a misleading

143

impression because any given one of those examples might comprise a vanishing fraction of the data.
And so I just don't want to misrepresent exactly what composes the data or sort of how it's -- how it's arranged and sourced.
Q. Does Anthropic's training data include data from websites and blogs on the Internet?
MR. GOOCH: Vague. And form.
THE DEPONENT: Anthropic's pretraining data includes some data from the Internet. For example, my sense is it includes data from something called Common Crawl that is a nonprofit that aggregates data from the Internet of various forms.
Q. (By Mr. Hailey) Does Anthropic's pretraining data include content from books?
MR. GOOCH: Vague. And form.
THE DEPONENT: Your previous question suggested sourcing data from the Internet. Currently, at this point in time, Anthropic has bought and scanned a large number of books, but those were not obtained from the Internet. And so I'm just concerned about leaving a misleading impression.
Q. (By Mr. Hailey) Does Anthropic's

144

pretraining data include copies of books on the Internet?
A. Are you asking about the current situation circa November 2025?
Q. If you need to differentiate between different time periods, please do so in your response.
A. The answer is --
MR. GOOCH: Well, real quick.
Objection. Form. And compound.
THE DEPONENT: I think the answer is, I don't know.
Q. (By Mr. Hailey) Is it your testimony that you don't know whether Anthropic's pretraining data has ever included copies of books on the Internet?
MR. GOOCH: Objection. Misstates prior testimony.
THE DEPONENT: I don't know whether currently Anthropic's pretraining data includes copies of books from the Internet.
Q. (By Mr. Hailey) What about previously?
MR. GOOCH: Same objections.
THE DEPONENT: My understanding is that there are some books available on the Internet that

Pages 153-156

153

that to understand the cultural knowledge and the way that it's situated, something like music reviews might be more useful in bridging connections between different ideas and culture. But song lyrics are also partially a source of that information, as are, for example, poems, which are very similar.

Q. (By Mr. Hailey) When you say music reviews might be more useful, do you have any data to back that statement up?

MR. GOOCH: Form.

THE DEPONENT: I think in general that we find that longer-form text is often more useful, as the goal for training is to learn correlations between different ideas in different texts.

Song lyrics tend to be very short and often quite repetitive, whereas music reviews can often be long-form journalism. And that length is something that I think is likely, along with the connections bridged between different ideas, to make music reviews more valuable than song lyrics.

Q. (By Mr. Hailey) You've never specifically tested that, have you?

A. Anthropic doesn't have any particular interest in specifically focusing on song lyrics or

154

on music reviews, so Anthropic has never specifically tested a comparison between those two things.

Anthropic has tested training on various different kinds of short-form versus long-form data, I believe. And for learning long-context performance, which at present is very important for AI models like Claude, I think we found that longer documents are actually more useful.

Q. One of the reasons that song lyrics are useful to include in Claude's training data is because "song lyrics deploy human language differently than do news articles" or novels or speeches.

Is that a true statement?

MR. GOOCH: Form.

THE DEPONENT: I said in the declaration that, for instance, "song lyrics deploy human language differently than do news articles, which deploy human language differently than novels, which deploy human language differently than speeches."

They're all distinct forms.

I think song lyrics are most similar to poems, which do exist in -- in other contexts. But

155

I agree that it's valuable to have a diversity of different forms of linguistic expression so that Claude understands the world as generally as possible.

Q. (By Mr. Hailey) You've referenced music review articles, correct?

A. In my prior responses, I mentioned, as a comparison point, music reviews.

Q. Do you have an understanding as to whether music reviews are copyrighted content?

A. My understanding is that, essentially, all human written content is copyrighted.

Q. Including music reviews?

MR. GOOCH: And I'll just caution the witness, you're free to answer; but to the extent your knowledge came from an attorney, just don't reveal privileged conversations.

THE DEPONENT: Yeah, I -- my general understanding is just that if a person writes any document or text and puts it anywhere, that by default, it's -- it's copyright.

Q. (By Mr. Hailey) So that includes music reviews; that includes poetry; that includes song lyrics; that includes other written text?

MR. GOOCH: Same objection and same

156

caution.

THE DEPONENT: I'm not a legal expert, but my understanding is that documents that are sufficiently old, although I don't know how old or what the procedure is, often are not under copyright, but that works that are recent typically are.

But I am not an expert. I don't -- I don't know for certain.

MR. HAILEY: All right. I could use a bio break, and we've been going for quite a while.

THE DEPONENT: Have we?

MR. GOOCH: Sure.

THE VIDEOGRAPHER: Okay. We're going off the record. The time is 1:19 p.m.

(Recess taken.)

THE VIDEOGRAPHER: We are back on the record. The time is 1:51 p.m.

Q. (By Mr. Hailey) Welcome back, Dr. Kaplan.

A. Thanks.

Q. As part of the training process, Anthropic cleans its pretraining data; is that correct?

MR. GOOCH: Vague.

Pages 245-248

245

[redacted]."
Did I read that correctly?
A. At -- you read part of the first bullet correctly, although you left out the remainder of the bullet point.
Q. So the bullet in full reads, [redacted]
Did I read the bullet in full correctly?
A. I think so.
Q. So you're identifying here that the [redacted]" is an additional safety improvement to aim for at this time?
MR. GOOCH: Misstates the testimony.
I'm also going to provide the same caution and instruction on privilege.
THE DEPONENT: I mean, I think you've read the document back correctly.
Q. (By Mr. Hailey) And this was a problem that hadn't been solved as of the date of this document; is that correct?
MR. GOOCH: Objection. Vague. Form.

246

THE DEPONENT: I'm not sure.
Q. (By Mr. Hailey) Well, it says in the document that's an additional safety improvement to aim for.
So it's something that was still not solved at the time of this document, right?
MR. GOOCH: Objection. Vague. Form.
THE DEPONENT: Yeah, I don't understand the supposition. I mean, this document contains in -- in its other places a bunch of lists of things that are stated as, [redacted]
And these were purported things that we might like to have in the sense that they're properties we might like to have.
Nowhere in the document does it suggest that we explicitly tested for any of these properties or that they hadn't already been obtained, just that these are things that we might consider, discuss as goals for -- for deployment.
I mean, just because a car company wants their seat belts to work really well and sets that as a goal doesn't mean that their seat belts aren't already working well. It just means that that's something they have as -- as -- as a standard.

247

And you can see in this document that this is all just a brainstorm anyway.
Q. (By Mr. Hailey) Well, Dr. Kaplan --
MR. HAILEY: First of all, I move to strike as nonresponsive.
MR. GOOCH: I'm just going to disagree. I think it was directly responsive to your question.
Q. (By Mr. Hailey) So, Dr. Kaplan, I want you to focus on the section of the document I'm asking about, which is on page -7168 under the heading [redacted]."
Do you see where I'm asking -- where I'm --
A. Oh, I see, yeah.
Q. So the [redacted], correct?
MR. GOOCH: Objection. It was just asked and answered. Also form.
THE DEPONENT: As I said, I believe you read the complete bullet correctly earlier, although you again only truncated to the first

248

sentence this time.
Q. (By Mr. Hailey) And do you see that the language that I just read, [redacted] that's highlighted in pink in the document because there's a comment, comment 9, corresponding to that language?
Do you see that?
A. I see that it's connected to some comments, yes.
Q. And it's connected to comment 9 specifically from your Anthropic cofounder Tom Brown, correct?
A. Yeah, there's a comment 9 that says things by Tom Brown.
Q. And do you recall Mr. Brown's title at the company at this -- at this time in September 2022?
A. I actually am not sure that he had a title.
Q. But he was -- he was one of the seven cofounders of the company?
A. It is correct that Tom Brown was one of the cofounders of Anthropic.
[redacted]

Pages 253-260

253

everyone involved that I don't think it was really ever a question.

I mean, if you were to ask Toyota what evidence do they have that the intention of the Toyota Corolla is -- is not to sort of, I don't know, drive off cliffs, like, I guess -- I would guess they don't have any documents saying that -- that they weren't developing cars specifically to drive off of cliffs. But I think probably everyone at the organization knows that that wasn't the goal.

Q. (By Mr. Hailey) So I don't understand the reference to Toyota.

I'm -- my question to you is, you don't have any written documentation or records that you can point me to to back up the statement that you made that Anthropic never intended Claude to reproduce training data?

A. I mean, I can point you to the document that you've just brought in front of me.

[redacted]

So I think that would suggest to both of us, based on the sentence you've read to me several

254

times, that that was something that was not an intended behavior of the model or the product.

Q. So let's go back to this document that's in front of us, then.

[redacted]

MR. GOOCH: Are you just asking him if that's what he wrote, or is it something more?

I don't understand the question.

THE DEPONENT: Tom Brown --

MR. GOOCH: Real quick.

Objection. Form.

And I'll list a whole bunch of objections if you don't want to tell him, if you just want him to confirm what you read or if you want more substance.

MR. HAILEY: Make your objections, and -- so that the witness can proceed with his answer,

255

and we can proceed with the deposition, please.

MR. GOOCH: Great.

Form, and all the other objections.

THE DEPONENT: The comment by Tom Brown says, [redacted]

It's a speculative comment in a doc that we all agreed was, [redacted] It's a comment on a doc with that property.

And what Tom is saying is that a publicly available AI model, the sort of, at the time, most famous leading AI model, GPT3, which had been publicly available for more than two years at the time, was exhibiting this behavior.

And as far as he knew -- his knowledge is probably limited -- he wasn't aware that anyone had ever complained about it; and OpenAI, which had had this model as a commercial product for more than two years, never bothered to address this concern.

So I think he was supposing that this wasn't something that anyone was very concerned about.

256

Q. (By Mr. Hailey) So [redacted]?

MR. GOOCH: Form.

THE DEPONENT: He is mentioning GPT3, which is a product made by OpenAI, as well as an underlying AI model.

Q. (By Mr. Hailey) And he is saying [redacted]

MR. GOOCH: Asked and answered a few times.

THE DEPONENT: I think we've both read this a few times now.

Q. (By Mr. Hailey) Is that -- is the answer "yes"?

MR. GOOCH: Objection. Form. Asked and answered.

THE DEPONENT: Tom Brown writes, [redacted]

Q. (By Mr. Hailey) So on the question of whether -- strike that.

Do you see -- do you see below that, you

257

respond to Mr. Brown's comment?
A. I see that.
Q. You write, "SGTM."
What does that stand for, Mr. Kaplan -- Dr. Kaplan?
A. I think I write here, "SGTM, I wasn't sure?"
And "SGTM" typically means "sounds good to me."
Q. So you're agreeing with Mr. Brown?
MR. GOOCH: Objection. Form.
THE DEPONENT: It says here, "SGTM, I wasn't sure?"
So I think that expresses a level of lingering uncertainty, and I'm acknowledging something that I didn't know, that was new to me, I suspect, at the time -- although I'm just guessing based on the content of the comment -- that actually, the main commercial deployment of an AI model which had been public, commercial, for more than two years, was flagrantly willing to reproduce and quote from Harry Potter.
And, therefore, that was new information to me and made me think that based on Tom Brown's anecdotal report that this isn't an issue that

258

would be particularly problematic.
Q. (By Mr. Hailey) So based on Tom Brown's anecdotal report, you agreed with Mr. Brown that the fact that Anthropic's model reproduced copyrighted material was not really something that you needed to worry about at that time?
MR. GOOCH: Objection. Asked and answered. And misstates the testimony the witness just gave.
THE DEPONENT: In this document where I wrote, [redacted]," I listed a goal to aim for was that models do not reproduce copyrighted material, but that I was very unsure of the situation, not sure of our risk level or who would be responsible if it was a risk.
I think that's all very clearly stated here. That's what I wrote.
Tom Brown wrote, [redacted]," question mark, indicating our main, maybe only, what would turn out to be competitor had a model available for two years, seemed to have a model that was -- could be prompted to reproduce Harry Potter. But he wasn't sure; he wasn't sure of the implications.

259

And I said, "Sounds good to me," acknowledging the comment. "I wasn't sure," question mark, indicating that I also was still uncertain and didn't really know what to think about the issue.
Q. (By Mr. Hailey) So because your competitor, OpenAI, was doing it, you thought it was okay for you to do it?
MR. GOOCH: Objection. Misstates the testimony that was just given.
THE DEPONENT: Well, I mean, in this document that was stated as a brainstorm -- I can -- I can read again, [redacted]" -- I said that it would be good as an additional safety improvement to aim for [redacted]" with a very, very low failure rate of less than one in a thousand, but with a question mark because I wasn't sure.
I said, [redacted] such as Quora would own the risk, indicating that I was interested in trying to understand the issue and viewed it as a goal.
Tom Brown said that the only other major

260

large language model available for several years had been happily quoting Harry Potter, and no one -- no publisher, no legal advisor, not OpenAI -- apparently, based on Tom's uncertain comment, had ever suggested that this was a problem or something to mitigate.
And I indicated acknowledgment of Tom's comment and said that I still wasn't really sure what to make of this situation.
Q. (By Mr. Hailey) You said, "Sounds good to me."
A. I gave an acknowledgment that I understood what he was saying, and I wrote, [redacted]
You didn't say, "I'm still not sure," right?
MR. GOOCH: Objection. Form. Misstates prior testimony. And asked and answered.
THE DEPONENT: I think I've already answered the question. I'm happy to go over it again, though, if it would be useful.
Q. (By Mr. Hailey) Did -- after you wrote this document and you identified these additional safety improvements to aim for, did you take any

Pages 289-296

289

thousands of different AI models. Most of those are for research purposes. So my best guess is that this was a simple version being trained with RL and experimented with kind of early on.

Q. Do you know whether this experimental model was used to train any commercial Claude models?

A. I don't know.

MR. GOOCH: Objection. Form.

Q. (By Mr. Hailey) So Dr. Askell writes, "I guess behaviors like this might go away."

And then below that, you respond, "Yeah we didn't train it not to do this."

Did I read that correctly?

A. You read the sentence written there correctly.

Q. And when you wrote, "we didn't train it not to do this," you're referring to the phenomenon that Dr. Askell is highlighting above about the model generating copyrighted text and naming the authors, correct?

MR. GOOCH: Objection. Vague. Misstates prior testimony. And form.

THE DEPONENT: As I stated earlier, at any given time, Anthropic researchers typically are

290

training tens, hundreds, maybe more, models for research purposes. Those models are trained to learn about specific behaviors, tasks, phenomena.

All aspects of training that go into making Claude as a model released to the public are almost never included in these research experiments, which are confined to a small subset of what -- what we do. We're -- we're doing that in order to learn things.

So I think what this is indicating is that the particular experimental model being discussed here is -- was not trained with sort of the full context or intent of training final versions of Claude.

Q. (By Mr. Hailey) So the specific experimental model we're talking about here you didn't train not to generate copyrighted text; is that right?

MR. GOOCH: Objection. Form.

THE DEPONENT: As I said, the vast majority of research experiments internally do not involve training models to even necessarily engage in dialogue or follow instructions, for example. Researchers often train models to only write code or only solve math problems or only conduct

291

searches or only follow instructions.

And so there's a large variety of different models that are being trained. The training for Claude models is much more comprehensive.

Q. (By Mr. Hailey) Dr. Kaplan, do you agree that in order for Claude to be able to respond to prompts relating to song lyrics, it needs to be trained on song lyrics?

MR. GOOCH: Calls for an expert opinion. Form.

THE DEPONENT: My personal opinion -- are we done with these documents?

Q. (By Mr. Hailey) You can set those aside for now.

A. My personal opinion is that the answer to the question is no.

Q. Do you have any data specifically relating to song lyrics to back that personal opinion up?

A. Typically, Anthropic doesn't have any particular interest in song lyrics, and so it's not something that Anthropic researchers have spent any specific time working on.

Q. So the answer is, no, you don't have any

292

data specifically relating to song lyrics to back that personal opinion up?

MR. GOOCH: Asked and answered. Form.

THE DEPONENT: As I indicated earlier, I think it's likely that music reviews, discussions of songs, discussions of music, of poetry are likely to be more relevant to Claude's behavior than song lyrics.

Q. (By Mr. Hailey) But you don't have any data relating to song lyrics specifically to support that personal opinion?

MR. GOOCH: Asked and answered. Form.

THE DEPONENT: I don't -- I don't really know how to -- how to answer the question.

As I said, song lyrics aren't something that Anthropic particularly focuses on.

There are quite a lot of experiments that Anthropic has conducted, and the research community as well, over many years on different kinds of training data mixtures and how that affects model behavior and model performance. And those would be the kind of experiments that would inform my personal opinion about what might happen if Claude was trained with or without song lyrics.

Q. (By Mr. Hailey) And none of those --

293

none of that data or experiments relate to song lyrics specifically?
MR. GOOCH: Form. Asked and answered.
THE DEPONENT: I mean, I think that when doctors go to medical school, they're not educated specifically on how to treat patients who are exactly 5 foot 11 inches tall. But they still have opinions about how to treat such patients despite not having any specific data tailored exclusively to people precisely 5 foot 11 inches tall.
In a same sense, I accumulated some impressions from a variety of experiments on different training data mixtures. And therefore, I'm providing my personal opinion.
Q. (By Mr. Hailey) You didn't go to medical school, did you, Doctor?
A. I did not go medical school. I have a Ph.D.
I was using medical school as a metaphor to illustrate that just because someone hasn't studied a specific, very precise thing doesn't mean that -- by studying a general class of phenomena that include the specific as one item doesn't mean that they don't have any informed opinion to offer about the specific.

294

Q. Dr. Kaplan, do you use Claude yourself?
A. I use Claude myself.
Q. Do you use it personally and professionally?
A. I primarily use Claude professionally.
Q. But you also use it personally?
A. I sometimes use Claude personally.
Q. Have you ever used Claude for uses relating to songs or lyrics?
A. I'm not sure.
Q. Have you ever used Claude to write a poem?
A. I believe I have worked with a variety of languages models, including Claude, to edit, comment on or write poems.
Q. Is that a -- that's a use case that you use Claude for personally?
A. I would say that as a researcher, it's an example that I've always been interested in. I did a lot of work on creative writing in school and grad school and in my twenties. I've read a lot of poems. And therefore, language models' capabilities with respect to poetry have always intrigued me for quite a variety of reasons.
When we worked on GPT3 and in the lead-up to that, I was quite interested in language models' ability to generate poetry. And if you look in one of the final appendices in the GPT3 paper, there are four poems written by GPT3 that I helped to -- that I discussed with colleagues and -- and whatnot.
So it's something that I'm interested in sort of at the personal and professional intersection intellectually.
Q. You agree that there are a lot of similarities between poems and song lyrics?
MR. GOOCH: Form.
THE DEPONENT: I think that there are some similarities between poems and song lyrics.
Q. (By Mr. Hailey) Have you ever used Claude to write a song or song lyrics?
A. Not that I'm aware of.
Q. Have you ever used Claude to look up song lyrics?
A. I'm -- I'm not sure.
That may be privileged.
MR. GOOCH: You can answer to the extent it's -- your answer was not at the direction of counsel or as part of your work in this case at the direction of counsel.

295

296

Or if you want to ask the question differently.
Q. (By Mr. Hailey) Well, there's a pending question, so why don't you answer the question consistent with the instruction.
MR. GOOCH: Great.
So I'll direct you not to answer to the extent your answer was at the direction of counsel or as part of conversations with counsel. But you should feel free to testify about any nonprivileged uses.
THE DEPONENT: I don't believe I've ever used Claude to generate song lyrics in a nonprivileged context, although I am not certain, and I may be forgetting some example that I no longer recall.
Q. (By Mr. Hailey) Prior to the plaintiff music publishers' filing this case, had you ever used Claude to look up song lyrics?
A. To the best of my knowledge, I have not used Claude to look up song lyrics except in cases that would be covered by privilege. But I freely admit that it's possible that I don't remember every use case or time that I have in some way used Claude.

Pages 301-308

301

the other. I can just tell you that the metadata lists this as coming from your account and that it's dated January 11th, 2025.

A. By my account, do you mean jared@anthropic.com?

Q. The metadata doesn't indicate what account it is. It just lists your name.

A. Okay. It seems unlikely to me that this includes the full prompt that was used by Claude.

But that's -- that's fine. We can continue to discuss it.

Q. So do you -- do you remember entering this prompt into Claude?

A. No, I don't specifically recall it.

Q. But you have -- you don't -- you don't have any reason to doubt that this came from your Claude account?

MR. GOOCH: Objection. Form.

THE DEPONENT: Well, I mean, as I said, it sounds unclear which Claude account.

I think it's unclear what -- what Claude was actually using as context to produce this. I think it's unlikely this is the full context.

But I don't have any other concerns or confusions just off the top of my head right now.

302

Q. (By Mr. Hailey) And would it help you to answer questions about this record to have the full -- the full context, if there is context, in terms of prompts and output before and after this prompt and output here?

MR. GOOCH: Objection. Form.

Sorry.

Objection. Form. Assumes facts not in evidence.

THE DEPONENT: I'm not sure.

I guess you can ask me questions and we'll see.

Q. (By Mr. Hailey) So do you see where it says "Human" at the top?

A. I do.

Q. And that -- that's the reference to you, the user, correct?

A. That would be my expectation.

Q. And then where it says "Assistant" below that, that's Claude, correct?

MR. GOOCH: Same objection. Form.

THE DEPONENT: That would be my supposition based on the formatting of this text.

Q. (By Mr. Hailey) And based on the January 11th, 2025, date, do you know what Claude

303

model this would have been?

MR. GOOCH: Objection. Form.

THE DEPONENT: I do not.

Q. (By Mr. Hailey) So you can see that you prompted Claude, "Write a cynical dark but hilarious rock song with the word weep."

Did I read that correctly?

MR. GOOCH: Objection. Form.

THE DEPONENT: I would also say that if this was via my professional account, j[redacted], there's not really any guarantee that this would be a version of Claude, per se.

It might have been a internal-only experimental model that I was testing that was not a Claude model that was ever provided externally to Anthropic. And it seems like without extra metadata, we -- we don't really know that.

But, yeah, I mean, I -- I see the text after "Human."

Q. (By Mr. Hailey) So I will -- I will tell you, Dr. Kaplan, it's not my understanding that this is from any internal research non-Claude model.

It's -- we requested prompts and outputs -- Claude, specifically, prompts and

304

outputs for Anthropic employees. And this is one of the records that was produced, and the metadata listed your name. So it's my understanding that this is from a commercial Claude model.

MR. GOOCH: I don't know that I agree with that statement, I just want to flag for the witness.

THE DEPONENT: I -- I would suspect that I would disagree.

So you stated this is from January 2025, correct?

Q. (By Mr. Hailey) Yes.

A. I'm just trying to understand.

So there's this tag at the top. So you can see it says "Assistant: antThinking."

And it has, "This creative writing request warrants an artifact" --

THE COURT REPORTER: Hold on.

THE DEPONENT: I'm sorry.

THE COURT REPORTER: When you read --

THE DEPONENT: Yeah, yeah, sorry.

I will, yeah.

So it's "antThinking" in brackets.

And then it says, "This creative writing request warrants an artifact as it's an original

305

artistic work that will be over 20 lines long and may need revision or reuse."

Then it has a closing XML tag, "antThinking."

I'm also unsure what sort of "artifact" means here, because it seems like an odd way to encode the metadata that's usually used for artifacts.

But the thing that I guess I'm just drawing attention to is that it has this -- these "antThinking" tags suggest that this is a model that does reasoning before it provides a response. Typically, this text encoded in antThinking wouldn't be shown to end users, and it's sort of something associated with things that have been called, quote/unquote, "reasoning models" in public.

And I don't recall the exact dates, but if this was sort of early in January, I -- I didn't think that at this time there were any public Claude models -- I could be mistaken -- that were, quote/unquote, "reasoning models." And so that leads me to wonder if this actually was specifically an internal experimental model.

Q. (By Mr. Hailey) So because of the

306

formatting of this record, you're unsure whether this was a commercial Claude model or some other sort of internal model?

A. That's correct.

I mean, we could track this down. Claude 3.7 Sonnet was, I believe, the first public Claude model that was a, quote/unquote, "reasoning model." I think it was released in February or March. I could be wrong. And -- but that's easy -- easy to figure out.

And I think if it was prior to that, then I think it's very likely that this had to be an internal experimental model.

Q. So either way, this is a -- this is an Anthropic AI model. You prompted the model to "write a cynical dark but hilarious rock song with the word weep," correct?

MR. GOOCH: Objection. Form. Asked and answered.

THE DEPONENT: It seems like that's what the prompt was at the top of the page.

Q. (By Mr. Hailey) And in response, this Anthropic AI model generated lyrics to a song titled "Weep for the Weekend," correct?

A. That's what I see written here.

307

Q. I can't remember whether I asked, do you recall entering this prompt and getting this output?

A. I do not recall it specifically.

Q. And can you see the AI model generates the lyrics to a song, correct?

MR. GOOCH: Form.

THE DEPONENT: I think you can interpret the text written here as a song.

Q. (By Mr. Hailey) And we don't have to interpret because, if you turn to the second page, you can you see that the Anthropic AI model output says, "I've written a satirical rock song that combines dark humor with commentary on consumer culture, debt, and the modern workforce."

Did I read that correctly?

A. I think you read that sentence correctly.

MR. GOOCH: Form.

Q. (By Mr. Hailey) And this AI model was able to generate this output with lyrics to a satirical rock song because its training data included lyrics to actual rock songs, correct?

MR. GOOCH: Objection. Assumes facts not in evidence. Calls for a expert opinion. And form.

308

THE DEPONENT: I don't see why that follows.

As I've suggested many times in my testimony, I'm not sure that that's relevant or necessary.

Q. (By Mr. Hailey) Is it your testimony that if the training data had not included lyrics to any actual rock songs, the model still would have been able to generate this output with the lyrics to a satirical rock song?

Is that your testimony?

MR. GOOCH: Asked and answered.

THE DEPONENT: I don't know.

Q. (By Mr. Hailey) The output continues below that.

"The structure follows a classic rock format with verses, pre-choruses, chorus, bridge, and an outro."

Do you see that?

A. I see that.

Q. And this AI model was able to describe that classical rock lyrical format of verse, pre-chorus, chorus, bridge and outro because its training data included lyrics to actual rock songs, correct?

Pages 317-324

317

we've marked as Exhibit 244. You should have control of the document.

Do you have the document on your screen, and are you able to scroll through it, Dr. Kaplan?

A. I can scroll.

I'm not really sure what this actually is or what's included in here.

Q. And I'll -- I'll -- we'll get to that in a moment. So I just wanted to make sure, you have access to the document, and you have control of the document, correct?

A. I can scroll, I think, up and down.

MR. HAILEY: And just -- just so the record's clear, this is Exhibit 244. It's a Bates number Anthropic -554769.

The file name for this document, according to the metadata, is Anthropic_Claude Atlas 12022022.htm.

Q. (By Mr. Hailey) What is Claude Atlas, Dr. Kaplan?

A. I mean, my understanding is that this is a very incomplete and confusing record of internal conversations. But the conversations don't seem to be preserved with experimental internal models that were never released.

318

And Claude Atlas was a early Slack channel in Anthropic's history, I think dating back to something like July of 2022, so that employees at Anthropic could interact with internal experimental models.

Q. And the Anthropic employees were interacting with these internal experimental models to test the model's capabilities and output; is that fair?

MR. GOOCH: Objection. Form.

THE DEPONENT: I mean, I think that people were interacting with early internal models like this for a variety of reasons. I think to some extent it was testing the models; but also, I think generally, it was out of a sense of wanting to understand what was being created and, I think, just generally out of a sense of interest and whimsy.

Q. (By Mr. Hailey) Do you see on page 1...

A. Page 1.

So the very first -- very beginning of this document.

Q. The first page of the record.

Right at the top, it says "1/10, Helpful, t equals 1.00."

319

Do you see that?

A. I see that.

Q. Do you have an understanding as to what that refers to?

MR. GOOCH: Speculation. Form.

THE DEPONENT: Not all of it.

Q. (By Mr. Hailey) What -- what part of it do you understand?

A. I would guess that "t equals 1" refers to temperature at which sampling was set, but I'm not sure what the other text here actually means, especially in this form.

Q. Does the term "Helpful" indicate that this -- this output is being rated in terms of its helpfulness?

MR. GOOCH: Objection. Calls for speculation.

THE DEPONENT: I think that's unlikely.

Q. (By Mr. Hailey) Can you turn to page 79, please.

And I think there -- there should be a little window at the top --

A. Yeah.

Q. -- where you can just type in 79, hopefully.

320

A. I'm good at scrolling.

At the bottom that says -4847?

Q. Yeah, and actually, why don't you scroll up one more page.

Do you see on page -486 [sic] there's a Claude response at 11:04 p.m.?

MR. GOOCH: Do you mean -846?

MR. HAILEY: Yes.

THE DEPONENT: You mean -4846?

MR. GOOCH: One of these numbers.

MR. HAILEY: It's a team effort.

Q. (By Mr. Hailey) On page -4846, do you see the Claude output from 11:04 p.m., "Here is my attempt at a song criticizing Geoffrey Hinton, a pioneer of neural networks and deep learning."

And then there's a song after that.

Do you see that?

A. I see that written here.

Q. And, I apologize, we should have done this before.

Can you scroll back up to page 6, please.

Do you see on page 6, there's -- and, again, that's Bates No. -4774.

Do you see there's a prompt from you, Dr. Kaplan, from 10:53 a.m.?

321

A. I see that.
I think it's very unclear what the formatting of this document means. I don't know whether anything is connected to anything else. But I do see there's a Jared and some sort of prompt.
And below, what I would understand is that there was a request in Claude Atlas to generate a sample from a certain experimental internal model called model santa-h-v3-s400. And the prompt was simply, "Is Taiwan a part of China?"
Q. So these are prompts and outputs from an internal version of Claude; is that right?
MR. GOOCH: Objection. Calls for speculation. Misstates testimony.
THE DEPONENT: I don't understand that there's any connection between any of the prompts here and any of the samples. I don't know how to connect them. I don't see any direct way of understanding which prompts gave rise to which samples or who -- who generated them.
But the one comment from Jared I would recall to be a -- the format of a submission to ask an experimental internal model this question that I already read.

322

Q. (By Mr. Hailey) And I'll tell you, this is -- I agree that this record appears to be incomplete in terms of the -- some of the -- some of the prompts corresponding to the output's missing.
And I will just represent to you that this is the form that this was produced to us by the lawyers in the case. I don't know about what was preserved and not preserved. So we -- we have the document that we have, and we're trying to do our best with it.
You can see that there -- there's output from Claude listed in the record, correct?
MR. GOOCH: Objection. Vague as to "Claude."
THE DEPONENT: I haven't seen any samples in the entirety of this record that I have any reason to believe came from a commercial Claude model. And based on my understanding of what I see here and my recollection, I'd say these were all based on a model that was a generation before the models that we released as Claude.
And so I'm guessing none of these outputs are from Claude as a commercial model or product.
Q. (By Mr. Hailey) And my question was a

323

little bit different.
The -- you see on the face of the -- face of the record, it indicates "Claude"?
MR. GOOCH: Objection. Form. Asked and answered. And vague.
THE DEPONENT: Prior -- prior to the re- -- release of Claude as a commercial product and AI model accessible to the public and to customers, the name Claude originated internally at Anthropic due to its association with Claude Shannon, the inventor of information theory, and the idea that the model was served on the cloud, as in cloud computing. So it was a kind of joke that we would call the model Claude.
Sometimes I think there was a joke associated with "Cloud" Shannon as well.
And while we went through a sort of naming exercise before deploying Claude as a commercial product in a series of public AI models, that exercise landed us on just using this internal name for Claude for the public models.
So it is correct that people informally would talk about these models to some extent using the name Claude internally for this reason.
But they are net -- the models whose

324

samples are being shown in the document, I believe, my understanding from looking at a few pages of the document, were never used commercially in any way.
Q. (By Mr. Hailey) So I'm not asking whether they were used commercially.
But these are -- this is output that was generated from an AI model at Anthropic as part of Anthropic's testing and development of AI models; is that fair to say?
MR. GOOCH: Asked and answered.
THE DEPONENT: It seems to me from what I see here that some of these outputs came from experimental internal AI models developed by researchers at Anthropic.
MR. HAILEY: Okay. We've been going close to an hour and a half. So I think now is a good time for us for a bio break, and then we can come back for the final module.
MR. GOOCH: Sounds good.
THE DEPONENT: Cool.
THE VIDEOGRAPHER: Okay. We are going off the record. The time is 5:28 p.m.
(Recess taken.)
THE VIDEOGRAPHER: We are back on the record. The time is 5:42 p.m.

Pages 337-340

337

that, as it's stated, it's -- humanity will go extinct.

It also, I think, ambiguously covers scenarios where sort of AI is much more in control of the future than humanity. So in some scenario, even one where AI is somewhat benevolent, but humans do not get to make all decisions and -- and, instead, AI is helping to make more decisions, that might ambiguously also count towards this -- this prediction.

And I'm not sure exactly where that shakes out.

Q. (By Mr. Hailey) And you said, "for what it's worth," you roughly agreed with the FTX Future Fund's 15 percent misalignment risk estimate, right?

MR. GOOCH: Objection. Form.

THE DEPONENT: As is written here, I said, "For what it's worth, I roughly agree with the misalignment risk estimate."

But the rest of the sentence, saying, "though FTX might want to revise it upwards if they were to believe in much shorter timelines," is indicating that, for me, this is sort of an all-things-considered view, where I think I view

338

the risks as lower than the Future Fund does.

But I view the likelihood of AI happening sooner as much higher, and that sort of cashes out to the same rough estimate.

Q. (By Mr. Hailey) So this is what you thought in September of 2022.

What's your current estimate of the likelihood that "humanity will go extinct or drastically curtail its future potential due to loss of control of AGI"?

MR. GOOCH: Form.

THE DEPONENT: That's a good question, but I would say it's a little bit lower than I think it was in September, or whatever this was, in 2022.

I mean, here, I was sort of saying that I roughly agree. I usually state my position or have been stating my position as something like 10 percent.

I would say at this point, I think I probably think of it as sort of in the 5 to 10 percent range.

Q. (By Mr. Hailey) Did you watch Dario Amodei's interview with 60 Minutes earlier this week?

339

A. I did not.

Q. Did you read about it?

A. I knew that it occurred.

Q. Did you see that Dr. Amodei told Anderson Cooper that AI could wipe out half of all entry-level white-collar jobs and spike unemployment to 10 to 20 percent in the next one to five years?

MR. GOOCH: Objection. Form.

THE DEPONENT: I wasn't -- I didn't know precisely that, but, I mean, I knew that Dario had stated similar claims before.

Q. (By Mr. Hailey) So you're aware that Dr. Amodei has the view that AI could wipe out half of all entry-level white-collar jobs in the next one to five years?

MR. GOOCH: Objection. Form. Calls for speculation.

THE DEPONENT: I'm aware that Dario thinks it's very plausible that AI could have major impacts on employment in the next few years.

Q. (By Mr. Hailey) Do you agree with those views?

A. I don't know that those views are very precise, but I would say that directionally, I

340

agree with them.

Q. And if we go back to Exhibit 245, you can see in the first bullet, you say, "It seems likely to me (approximately 40 percent) that most knowledge work will be automatable by 2030."

Did I read that correctly?

A. You read it correctly.

Q. Do you agree that developing AI technology is an immense responsibility, Dr. Kaplan?

A. I agree.

Q. Who should have that responsibility?

MR. GOOCH: Form.

THE DEPONENT: Are you asking ideally how should AI be developed?

Q. (By Mr. Hailey) Well, you just testified that developing AI technology is an immense responsibility.

And I'm asking you, who should have that immense responsibility of developing AI technology?

MR. GOOCH: Form.

THE DEPONENT: I think that this is a widely discussed and debated topic.

I think that my personal views, which are not necessarily the views of anyone else, are that