Hello, everyone. Welcome to another Q&A for… about science and technology, both kids and others.
I see a number of questions here.
Gregory is asking, what would count as proof in a world where AI is doing most of the science?
That's sort of an interesting question. Well, what… the…
What is the purpose of proof?
That's a… it's an idea, particularly coming out of mathematics, originally from geometry, where when saying.
That one has kind of a… an argument that is sort of purely formal.
Purely, kind of, abstract and logical.
That can explain why something is true.
Now, the fact is that the idea, when, for a long time in the history of mathematics.
to know that something was true required that you had a proof that people could look at and say, yes, I understand all the steps in this proof, therefore the result must be true.
In a sense, in the last, I don't know, half century or so that we've had computation, there's kind of an alternative way to find out that something is true.
You say, well, I know the steps that are being done in the computation, and I believe that my computer actually followed those steps. Therefore, whatever the computer came out with is the correct answer.
as I say, in the past, the only sort of way to say, oh, it's the correct answer, is to give something which people would describe as a proof.
Where then you could say, well, a human could go check the steps.
If they wanted to.
Whereas with a computer.
It's like you've defined the steps, and you can trust that the computer is just going to follow through those steps.
Of course, in the case of a proof that is being checked by humans, you have to trust that the humans will correctly check it. So… and in fact, the computer, one can expect to be a lot more likely to do it right than the humans.
So people had long believed that to know that things were true, let's say in mathematics and doing mathematical kinds of things, that you had to have sort of a proof. But once we have computation, you just have to have the computation.
And when we introduced Mathematica in 1988, there was sort of a… people were, at first, at the very beginning, a bit confused, like, it's giving a mathematical result. How can we know if it's correct?
They looked like, where's the proof, type thing. And it took only about a year for people to realize that, no, actually, the fact that this came out of a precise computation was the thing that told you it was true, so to speak.
Now, in doing pure mathematics, mathematics with sort of abstract mathematics, there's kind of a different role that emerged for proof.
Which is kind of a way of explaining, kind of at a more intuitional level, why something is true.
So, there's one thing is to just go through the steps, calculate it, compute it, and no human may ever understand those steps. Another is to be able to kind of, write
an exposition that explains why the thing is true. And that second use of proof is something that has been kind of important in the development of pure mathematics.
Particularly in the last century or two.

Now, when it comes to, sort of, the relationship between proof as exposition and proof as making sure the answer is right.
Those are… those are things that, kind of…
Interact in a complicated way when you're dealing with sort of computer-based proofs.
So… There are cases where you are…
In, for example, mathematics, where you're just trying to work out the answer.
There are other cases where you're asked the question,
where it's… and as I say, that's probably the vastly dominant use of mathematics, and the world is…
What's the answer?
Not explain why this is true.
There are cases in which it's sort of interesting.
To ask, why is it true?
But… When that sort of explanation, that working out of why it's true, is done by a computer.
There's no guarantee that that working out should be something that a human will understand.
And, in fact, there's a thing I did, oh, 26 years ago now, of a kind of a proof done automatically by computer that,
was… Something where one could have just thought of it as a computation. This does this.
But it was, in sort of the tradition of mathematics, it was more thought of as a proof that there is this mathematical statement
That implies this and this and this.
That proof
generated by computer is completely incomprehensible. I mean, I've tried, other people have tried over the course of years to understand what's going on in this 100-step proof, and it's been essentially impossible to do that.
that proof doesn't really contribute much to the question of, sort of, why is the result true? It just says, it is true, and now you know it, essentially, by computation.
So… In… now, when it comes to… to modern neural net AIs.
There's another complication to the whole thing.
Which is… It's…
When one's dealing with, sort of, precise computation, it's like you set up the computation, it goes crunch, crunch, crunch, and you get a definite answer.
neural net AIs
are not set up that way. They often even explicitly involve a certain amount of randomness. But in general, it's kind of like we poured in all of this training data, we've kind of trained the AI based on what's on the web and in books and so on.
And now it's like, okay, let's hope it does the right thing.
And we've tried to set it up so that it's trained so that, with high probability.
it will have done the right thing of the things we trained it to do. Now, when we actually give it a new thing to do that wasn't what we trained it to do, we just have to sort of hope that it will extrapolate from its training data to do the thing that we consider to be the right thing.
So it's no longer one of these, setups where…
like a sort of precise computation where you're giving precise rules and a precise result is coming out. It's something where you just have to kind of hope that the result that comes out will be kind of the thing you want.
So when it comes to asking, you know, does the AI prove this?

Well, in the usual setup, with kind of just a large language model, neural net AI, whatever.
What comes out, even if the thing gives a long chain of reasoning.
There really isn't much idea that that is really, sort of, precisely correct.
Now, there are all sorts of attempts to connect LLMs, To so-called proof assistant systems.
What is a proof assistant system?
It's… something where… the…
You've kind of defined, kind of, steps that you can make, and you've said, these are steps that are valid.
And now you've asked… now you ask, can you construct a proof that only involves valid steps?
And…
In the past, proof-assistant systems were operated by humans. Humans would try and put these sort of LEGO blocks down in just the right configuration, or puzzle pieces down in just the right configuration, so that the system could verify that, yes, those pieces fit together, and and when they… and if they get to where you want it to go, and they all fit together, then you could say, yes, we have a proof.
And…
Traditionally, that was done by humans. I mean, proof assistants were never that popular, but they were a lot of work to operate, but that work was done by humans, sort of fitting together the puzzle pieces. Kind of a new game in town is try and have an LLM fit together those puzzle pieces.
And if the puzzle pieces fit, you might say, then bingo, we've got an actual proof.
Unfortunately, it is not quite as simple as that. The thing is that there's the question of whether the thing you thought you were proving, that was some statement that you made, maybe in mathematical notation, or in English, or something like this.
did that correctly get converted to something that sort of fits in with these precise puzzle pieces?
That's one issue.
Another… another issue is, even if you get those puzzle pieces fitted together, what kind of a thing is that proof?
Is that proof?
a, something like sort of a computation where you don't really… can't really look inside, and it's, like, too complicated to look inside. It may be precise… you may be able to be sure that it's right, but it's not something that provides useful kind of exposition.
You know, I will say that one of the things
that, again, makes things complicated in modern times, is that LLMs
In the past were just purely, going, kind of, from their own
kind of training data, working out, sort of, what output to give. Increasingly, LLMs use tools like our technology stack, where the LLM gets to a point and says, okay, let me just compute that with all the language.
and then get back a precise result, and kind of jump forwards to do that. And even more than that, there's the following possibility, that if the LLM is doing something where you can… if it got
a right answer, then it can tell that it got a right answer, and if it didn't get a right answer, it can tell it didn't get a right answer. Then it can kind of loop around, keeping on trying, you know, is this the right answer? Oh, no, let me go back and retry. Is this the right answer? Let me retry, and so on. This is a thing that with

so-called reasoning models is a… is an increasingly, sort of common thing to try to do, is to say, if we're trying to produce a thing where we can sort of tell if it's right, we might be producing a piece of code, for example, where we can tell, does the code run? Does the code run and produce the result we wanted? Maybe it's even kind of a mathematical derivation, where we can say, did it get to the thing we wanted it to get to?
Did it get to something where we can check it, and so on?
And then we end up with this loop where you're going around saying, did it get there? No. Okay, let's try again. Did it get there? Et cetera, et cetera, et cetera.
And the thing that is important to think about in that is when you say, oh, it didn't get there, let me try again.
how do you try again? What… if you're picking, sort of, at random, where to try next.
that… how efficient can you make that? Because sometimes you don't have any choice, you don't really know where to go, you just have to do it really at random.
But often, you can say, well, I kind of can get a sense of which way one should go, so let me kind of weight things so that it's going more in the direction that we think we want to go. And LLMs seem quite good at doing that kind of weighting, of saying, wait a minute, you know, I'm trying to write this piece of code, and it's not, and I vaguely have seen that people who write this kind of code
use this kind of construct to do that. So let me, instead of just picking constructs at random, let me try mostly to use the construct that I've kind of vaguely seen people using.
And that approach of kind of picking, sort of, which direction to go, when you're… you're even… you're being a bit random, but you're also informing, sort of, that… what… what you choose out of that randomness from things that
the system has been… that the LLM has somehow been trained on. And that… that seems to be a fairly powerful technique.
It's… it's very reminiscent of, kind of, biological evolution, this whole process of you make some little tweaks in the program that you're using, then that develops into, in the case of biology, this whole organism, then that whole organism does better or worse.
And then you look at that whole result, and then you tweak the program again.
And what… It's kind of remarkable how much… how far you can get by having
sort of just tweaking the program, then seeing how it runs, tweaking the program again, and so on. I've been very surprised in a bunch of experiments I've done the last couple of years, particularly aimed at biology, but also with applications to machine learning, the extent to which It's a very powerful thing to just be able to make small changes to a program, and it sort of wouldn't work if every change to the program just made the program not operate.
But once the program is, it doesn't have to be very big, it could be, like, 27 bits long or something, but just so long as there's enough different choices of what the program could be, it turns out that making a small change
at least sometimes, doesn't make the whole thing just go crazy. It just makes the thing change a bit, and that changing a bit allows you to kind of home in in the direction you want to go.
So…
this… there's sort of a certain magic to this kind of adaptive evolution. Well, it's a combination of the adaptive evolution and the kind of
sort of power of computation going from the program to what the program generates. But that's sort of a surprising and magic thing, and if you combine that with sort of the… not just, sort of changing the program at random, but

Learning, learning how to change the program, learning what kinds of changes are more likely to work.
That seems to be something that can be very powerful. In fact, biological evolution, to some extent, does the thing of making certain pieces of our program, our genome, change more rapidly in certain kinds of situations.
And it's probably… it's probably learned how to sort of optimize the process of adaptive evolution in certain cases.
Anyway, a few, a few thoughts about, about that.
Let's see…
Lux is asking if two AIs trained separately develop similar ideas, does that mean those ideas are somehow inevitable?
It's an interesting question.
I think… Well.
Okay, so I've seen many examples of related phenomena in studying biological evolution, in studying, kind of, adaptive evolution for biology.
And what you see there is that…
you can… well, in the things I've done, you can explicitly see these patterns of behavior.
And let's say you have a particular objective, like, have the pattern live as long as possible, let's say.
There are many different ways to achieve that objective.
And when you look at these patterns, you see that they're visually very different. There are many different branches and paths of how you can achieve that objective.
I think… This question of when are those pods
sort of achieving their objective using the same idea is a very interesting question, and one I don't really completely know how to… how to answer properly. It's like, when is the idea the same? Well, there's one that has big blocks of purple or something. There's another one that has, you know, a kind of yellow
vertical lines in it, and so on. And visually, you can say, there are… there are these,
there are these structures which, you know, you can identify, sort of, there is a structure that developed here, and it went and used that structure to go and achieve its fitness objective. As it sort of iterated on changing the program. That particular structure kind of survived and got built on.
What you see is a certain… with a certain amount of the time, you see convergent evolution. That is, two things that maybe even started working rather differently, as they keep adapting and evolving, they end up discovering the same thing. Like, they both discover that you can have a big block of purple to extend the lifetime or something.
And so… I think… That same kind of thing, of there's a certain set of possible
there's a certain set of possible ways to achieve something, and sometimes you'll fall into the same way to achieve things. I would say that most of the time, at least in things I've looked at with biological evolution, there's an awful lot of ways to solve the problem. It is…
Rare.
that there's just one way to solve the problem. You can ask the same question in mathematics. Is there one way to prove a particular result? Or are there many possible proofs you can give of a given result? I would say in most of the things I've seen.
In, that involve any kind of non-trivial computational steps, that there end up being a lot of choices.

Although some of those choices may appear quite similar. They may… it may be sort of examples of convergent evolution. It may almost seem like they had the same general idea, even though the details of how it worked were different. But characterizing when they have the same general idea, I don't really know completely how to do.
And that's… that's kind of what you would have to do to ask whether two LLMs that sort of do something in,
develop similar ideas, what does that mean? You know, how do you… how do you characterize the similarity of ideas?
I think… kind of… for certain…
there's a complicated landscape of different ways to solve particular problems. My guess is that if you see LLMs following, sort of, two very similar paths.
That's more because the LLMs are similar inside than it is because there are only those parts. But one would have to actually see particular examples to know that for sure.
Robot is asking, could an AI get bored or curious like people do?
Well…
You know, one of the things about AIs is something kind of has to keep them going on running. With… Probably the same is true, I mean, with, with,
It's kind of like the chemical processes that keep us humans running, you know, if we eat and sleep and do other things, we'll sort of just keep running.
And similarly, with an AI, if it's just, you know, it's running on a computer, the computer has still got power and so on, the thing is going to keep running. The question of
of, well, one could ask that for the operating system of a computer. You've just got your computer, and it's running, and it's doing things, and if something, you know, if the camera of your computer is on, and a cat walks past or something, the computer is going to do lots of kinds of things inside as it processes that image.
The, the question of, so…
You know, is that… does that mean that the computer is kind of seeking out
new frontiers to do things. Well, not really. It's just following… it's just following the operating system that it had set up.
And one might… You know, to ask whether
The computer is bored, curious, whatever else.
Those are… those are concepts that we feel familiar with because we experience them ourselves as individuals.
And sort of the inference that anything else experiences these things is just something we have to extrapolate from our own experiences. So, you know, whether we should call our computer bored when the computer isn't
doing anything other than running, sort of, the basic things it does in its operating system, I'm… I'm not sure.
in terms of… Will…
sort of an AI spontaneously do things, well, if it's set up to just keep going, and keep trying to work out, you know, what am I going to say next, what am I going to say next, etc, etc, etc, then yes, it will… it will keep doing things.
Gosh, you guys are asking difficult questions today.
Let's see… Jamie is asking, could future scientists think directly in computational structures instead of language?
Well…

There's been a long sort of debate in philosophy about whether you need language in order to think.
And to what extent language affects how you think?
My view tends to be that in our brains, we've got, you know, 100 billion neurons that are constantly firing and affecting other neurons and so on, but the question is, does all that activity Become something that we can summarize and remember.
It's kind of like, you're not going to remember. There's no way to sort of store in the brain of a certain size, there's no way you can store the full history of what every neuron did. You have to kind of summarize that history somehow. And in a sense, language is in…
when it comes to, sort of, thinking, language is what provides us ways to summarize all that activity. We can say, you know, I,
sort of, I ate a cookie or something, and that's… the process of eating a cookie involves lots of neuron firings and lots of kind of imagining, you know, how you move it to your mouth and start chomping on it and all those kinds of things, but
Those things are all kind of lost from what we remember. At best, we remember, I ate a cookie, and somehow language
captures the things that we are going to remember about, kind of, eating the cookie. Now, there may be details about what the cookie smelled like, or whatever else that are not so easy for us to put into language, but in a first approximation, the things we're remembering are things that we have managed to summarize by talking about them in terms of language.
So, I think language is kind of the… the way that we
Sort of set things up so that we can remember them, and so that we can operate on them in future with our brains.
Now, the question is, what?
What is that language like?
And human natural languages have evolved over the last, I don't know, some number of hundreds of thousands of years to have… to be the languages we have today.
And languages are a complicated social thing, because it doesn't… the… we… we kind of learn and transmit languages by communicating with other people, and that's kind of one of the roles of language is not just for us to remember internally to our brains.
What happened, but also to be able to communicate that to other people.
And I think one could imagine that if one was just sort of isolated from everybody else, that one could invent a language that one could use in one's own mind to remember things and to sort of put pieces together and so on. I think… but, you know, the languages that
Inventing a language from scratch as a newborn human, so to speak, not so easy.
A lot easier if there are 100 billion humans who've lived and have collectively figured out, you know, this is how language works, and this is how to add this kind of word on, and so on.
And if you get the sort of accumulated wisdom of those 100 billion humans, then you're off to a good start, and you're going to end up using the language that has been derived from all that sort of human activity and communication.
Well, the question is, what, you know, how does that kind of language, how does the idea of language, which our brains are pretty well set up to deal with, how do you extend that idea of language to things other than sort of inherited natural language that we have traditionally had?
Well, any kind of formal structure is sort of an extension of language, whether it's logic, whether it's pieces of mathematics.

It's… it's something where we're putting together these kind of symbolic ideas, these lumps of kind of symbolic thought, and we're combining them as, you know, X plus Y, or, you know, if P and Q, then et cetera, et cetera, et cetera.
These are things that are very language-like.
But… and we're using probably the same kind of brain structures to be able to process these things, but there are things where we have constructed something that is often a more precise, more buildable-on language-like structure than we've had before.
Well, so…
I've spent a good part of my life trying to build our Wolfram language, computational language, where we're kind of taking
We're trying to make precise the way that we can talk about things in the world in computational terms.
And we're often using, to describe the computational constructs we have, we're often using words that come out of natural language, because that's what people are already familiar with.
But we, in our computational language, have a much more precise meaning
for those words that allows us to understand and the computer to understand precisely the consequences of those things. And so, I think in terms of the question of
can people, do people think in terms of computational language, more so than in terms of natural language? I know that when I'm trying to figure out, figure things out, you know, I can write Wolfram language before I could say what I'm going to write.
And certainly well before I could explain to an LLM what I want to write, and have it write the Wolfram language for me, I can think pretty well directly in Wolfram language, in terms of, you know, I don't know, fold list of function of this, of that.
and so on. I'm not sure I could verbalize it that well, but I can certainly type it in without having sort of, without having thought to myself, what am I doing? Let me make a plan in English, then let me translate that to this computational language. I mean, for me, it's, and I think for lots of other people, it's kind of a language that you can think in, as well as a language that you can kind of communicate with your computer in.
And so, yes, I think that the idea of, sort of, computational language as a language to think in that's relevant to building
sort of new science, it's a… it's an important idea. I mean, kind of a precursor is the language of mathematics, things like mathematical notation. And again, there, mathematical notation has been an important sort of building block for thinking about things in mathematical terms.
when it comes to thinking about things in computational terms, a much broader set of things that can be thought about that way, computational language and the kinds of things I spend a long time building is, I think, what one needs to be able to… to have… to talk about and think about things in those kinds of terms.
Boy, Libra is asking, how are you holding up in the age of AI? Are you using it, enjoying it? Any favorite models or anecdotes?
You know, for me, I'm always interested in automating as much as I can, because I'm kind of, like, I want to go from ideas that I have
to kind of, things that I can build on as efficiently as possible.
And so I've spent a lot of effort building tools that allow me to build these sort of tall towers of capability, and that's really what computational language is about. You can precisely build sort of each brick in the tower so that you can build a long way and not have the tower fall over.

So, when it comes to, sort of, modern AIs, they're… they're very useful for, kind of, broad, but somewhat shallow kinds of operations.
So, what do I use them for? I mean, I use them for lots of, sort of… rather than reading lots of web pages, it's like, okay, tell me what the basic answer is. It's pretty good at extracting the information and summarizing it.
Do I completely… am I completely sure it's right? If I really, really care about the answer, I'm going to drill further. But in many cases, if the answer seems plausible to me, and then it's like, good, okay, I can move on to the next thing.
I would say that I've been interested in using AI to kind of extract thematic results from millions of papers in the scientific literature. I don't think that's been quite as successful yet for me. I think I need to push harder on trying to do that. There are things that I'm interested in, which are kind of areas of science where I think that there's ways to sort of pluck results from
many different places that have never been combined together before, combine them, and produce, kind of, interesting results. However, what I've certainly noticed in sort of many uses of AI, if you understand what you want with more clarity, there's a much better chance it's actually going to work in some sensible way, and not just go off and sort of wander around and do crazy things.
I've… I've used… AI a bit to write Wolfram language code.
usually things where I want some detail of something, some plot, where I want it to have a border in this way and that way, and I can't remember exactly how to do it, and I can't be bothered to look up the documentation. I just ask the AI to do it, and it can pluck things from different pieces of documentation and so on, and put them together and say, this is the result. Maybe it's right, maybe it isn't quite right. If it's almost right, I can kind of look at the Waltham Language code and change it. I mean, I think a key thing for Waltham Language is that it is
a language that humans can read as well as humans writing. So if you're trying to build some sort of tower of computational capability, you can say to the LLM, just go write this thing, and it will do something where you're not going to understand the code. If the thing you're producing is just something that has to look roughly right.
Well, great, that's probably enough.
If the thing you're producing has to be this sort of precise brick that fits in this big tower, that's not going to be okay. You're going to need to look at that thing and understand for yourself whether it is actually the thing you wanted it to be.
And that's kind of a unique feature of Wolfram language, is it's a language that computers understand and humans also understand. So a workflow that seems to be quite good is you sort of say something vague to the AI, it comes back with a piece of quite succinct, but short, precise Wolfman Language code. You look at it, you say, no, you didn't get that quite right.
You know, either you tell it to modify it, or you modify it yourself. Then you have this sort of very solid brick
of Wolfram language code that you can use. It's a precise thing, you can use it to keep building from there. So that's a useful kind of concept. It's for all the computations where you need to know what you're computing, it's kind of the main game in town.
And it's when… if you are… if it's good enough to have a computation where you don't really know what got computed, you just know that the result was roughly, you know, roughly what you wanted, you know, you're making some image.
and you just want a picture of a, you know, a duck flying over an elephant or something. Well, if it looks roughly right.

then that may be good enough for the purposes you had. If you're trying to figure out the precise trajectory of this or that thing, then you're going to need to know, did you… did you compute… are you computing the thing you thought you were computing, and then are you computing it correctly, so to speak? And that's kind of the… the role of… of,
Our language, it always has been the role of our language, is to formalize things, make things precise, so that you can build with them in a systematic way. And that has been the
that's been the sort of great advance of science since the 1600s, has been this notion of formalizing things. I mean, there are plenty of areas of science where, really, you don't get to formalize much. You know, if you're doing life sciences, biology, whatever else, it's just there's a lot of facts.
and a lot of things fit together in complicated ways. There's not a lot of places where you can say, this is kind of the formal structure, and now we can use that formal structure to build this big tower of consequences. That's something that's been very successful in physics and mathematics. It's not really been nearly as successful almost anywhere else. It's sort of beginning to have some success in chemistry, for example.
That's a thing where what I've been interested in is kind of contributing to the ability to sort of formalize things and build these big sort of towers of consequences, which is the big thing that started happening as mathematics came into the physical sciences in the 1600s.
And it's something where, that… that really is quite different from what kind of a neural net AI is about.
there's a… there's a quite separate issue of can you analyze neural net AI and what it's doing using ideas that come out of, sort of, the… the computational paradigm. That's another very interesting area that I've been interested in. But, that's,
that's a bit of the story there. I mean, I suppose in, in modern times.
with AIs, it's, I mean, it's a very… it's a strange and swirling world out there, where people are saying, oh, I've got this benchmark that proves that my AI has almost reached AGI, or whatever else.
it's… these things are fairly crazy. I mean, it's, and there's a lot of, kind of, gaming that's going on in the, sort of, let me… let me choose a benchmark that is exactly right, so that my AI happens one time in five to solve this problem, and so on.
That doesn't mean that it's completely useless in assessing what's happening with the AI, but when it's kind of like, we're going to… we're going to do this race
and it's going to be this elaborately made benchmark against that one. It's, it's kind of like… like you're setting up a trial of something. Could be… could be a medical, clinical trial, or something like this. And you are putting a lot of effort into making sure that the way you do the test
It's such that you're going to get a good result.
And that's… that makes it much harder to interpret the test if you know that lots of effort was gone to… to really set the test up so you have the absolute best chance to get a good result. That kind of changes how you should think about the fact that a good result or not a good result was got in that test.
Let's see… There's a question here.
From Creature. Will future humans look at our science the way we look at ancient myths?
It's an interesting question. I mean, I think that… As you look at… kind of…
descriptions of, I don't know, the… the,
why… why the world works the way it does. You know, the spirit of the wind pushed this or that thing.

Might be sort of the ancient version, or the, or… This was determined by some
Something which is, you know, some… some sort of spirit, god, whatever else, did this thing in the world.
And that's our explanation of how that happened in the world.
As we've gone through developing science.
One thing that's happened is the explanations get in, sort of.
The exact science is a lot more elaborate.
It's not…
And then the wind god decided that this or that should happen, and maybe you go through a few steps of the story, but then basically you got to the answer.
It's not like you have to go through millions of steps to get from the original statement of how things were set up to the answer, which is what you can have to do in more formalized areas of science.
So, in terms of the way that we describe things, it's sort of been interesting. I mean, when we look at mathematics, for example, if you look at Babylonian mathematics from 4,000 years ago. we can still plainly recognize a lot of the constructs that we're using today. Some of the presentation is different, I mean, the way that numbers were written out was very different. Those kinds of things, but the concepts are still very much the same.
The, when it comes to other areas of science, things have changed a lot in terms of what
you know, are things described in words? Are things described more formally? I think it is a general trend in science that one wants to formalize things. One wants to get things to the point where one can build a tall tower of consequences, rather than just having to say, this is how it appears to work, and that's the end of the story.
And I think… that,
Is… is no doubt a trend that will continue, and as it continues, things that
looked to us like things we just sort of have to wave our hands about today become things we can talk about formally. For example, things we were talking about earlier today, about when do two systems develop with different ideas? When do things build on different ideas to go forward?
Well, at this point, we just have to talk about that. We don't really know how to characterize what's one idea versus another idea, what does it mean to build on ideas, and so on. There will be a formalization of that, hopefully soon.
Once that's formalized, it will sound very different to talk about these things. And that's a place where, looking back at sort of earlier science, one will say, that's very vague, that's just a bunch of words, whereas now, we can do things in a formal way and be able to build much further from it.
I mean, in, in, I don't know, for example,
a lot of ancient Greek physics.
was very vague. It's like, oh, there are atoms, and they're in the void, and the atoms are bouncing around, and there's an atom that's a certain shape, and that represents heat, and there's another atom of another shape that represents water, and so on. It's all rather
sort of descriptive, but not completely off in the wrong direction. Yes, there's molecules that make water that are different
from molecules that make fire, let's say… well, molecules don't make fire, but fire is a different kind of thing that wasn't recognized when people were talking about atoms in ancient Greek times.

But I think this… this notion that what starts as words ends up formalized is a general trend in science, and that will make some of the things we talk about today that we talk about only in terms of words, look very myth-like.
To… to science of the future.
I mean, I think that, in… sort of… a science done by AIs
Tends to operate… well, okay, it's a tricky thing, because
You know, it has to understand, what is the point of science? What is science trying to do?
In a sense, the natural world does what it does.
And we have certain things we can think about. The role of science is to make this bridge between what the natural world does and what we can understand in our minds.
So it's a way of, kind of, having a narrative about how the world works that somehow fits in our minds. And in the past, that narrative will be in terms of, you know, spirits and gods and so on.
In modern times, that narrative is more in terms of mathematics and computation and so on.
It's, and as I say, the great advantage of that formalized way of doing things is that you can build this kind of tower of consequences.
So, I think,
the answer is that there are areas of science, like, for example, in biology, I'm quite certain that there are more formal theories of biology that can be given. The way that, kind of, molecules are kind of orchestrated in what they do at a
a molecular scale, and so on. And I think the things that we discussed today about, kind of, sort of how biological organisms work.
will look very primitive in the future. There'll be much more global things we can say, much more formal things we can say there, probably true in a bunch of other places as well.
So, yes, I mean, that is the… that's the way that things work. In the progress of science, things get formalized, and the unformalized version looks like just a bunch of, sort of, almost myth-like words.
Let's see…
Robot is asking, are neural networks the final form of AI, or just a weird early prototype?
I think the main thing to understand about that is…
At some level, it doesn't matter. What do I mean by that?
We have…
We have brains that sort of do thinking-like things using neurons and electrochemistry and all those kinds of things. We have computers that do sort of thinking-like things using silicon and electrons and transistors and so on.
and using particular setups for their machine instructions. But in the end, those thinking-like things, those computational things, can be done
on different substrates. They can be done on, sort of, the substrate of neurons, they can be done on the substrate of transistors and electronics, they can be done on the substrate of artificial neural networks. These are… but…
Well, actually, that last thing is not… that probably shouldn't be quite so closely connected, but the point is that there's a way of kind of saying, we can compute this stuff, we do it using these foundational elements.
the same things can be computed with different foundational elements. That's one of the big surprises that sort of started emerging in the 1930s, and then I think emerged with much more clarity, probably even through my efforts in the 1980s and so on, that… that there are different bases for computation that all sort of can work equally well.

And so when you say, so neural nets, it may be the case that People say, well.
You know, you could tweak this, you could make it a little bit more efficient there. You could even use a completely different foundation.
But still, the basic kind of workflow of what's happening will be very similar. I mean, the same thing has happened in the history of computers. You know, there were vacuum tubes, then there were transistors, there were… there was core memory, then there was semiconductor memory. These all work differently, but somehow the functions they achieve are the same.
And my guess is.
That with neural nets, certain things will be done quite efficiently with current neural net architectures, some other things not so efficiently, but in the end, there will be a certain sort of universal equivalence between these different architectures, which means there's no fundamental difference between what you can do in these different places.
So, it's really more of a practical, technological question, what the right kind of hardware architecture for these things is. And I've done things myself, looking at, sort of.
generalizations or different directions than neural nets, looking at sort of discrete things where every… everything is, every… every sort of memory… in a neural net, memories are kind of stored
with continuous numbers that can have any value. I've been interested in whether you can store and compute with things where you just have only discrete values and so on, and it seems that you can.
But…
whether that may be more efficient in certain ways, given certain pieces of electronics that you use to implement things, I… I don't think… I think in the end, it will be somewhat equivalent.
To the neural nets that we have today.
So, while, as I say, there may be specific advances, and there are particular cases where you need advances to be able to do the things we want to do with the, with kind of the, the kind of inputs that we have today, oh, for example, the,
I don't know, people in robotics, for example, are very concerned about the fact that it's not so easy to get
enough training data to train a robot with all those different motions and so on. It's like, well, you could simulate the robot, you could run the robot in practice. When you… when you simulate it, it's hard to get all the sort of results about how much force you should put to pick something up and so on, to get all those things right, but
you know, people tend to think that you have to sort of burn into the structure of the… of the neural nets, or the AI, more information about, kind of, the physical
Kind of things that can happen with actual robots in the physical world, and that that will be important, and that sort of burning in that architecture will be important to making neural nets really be able to do robotics fluently, and such like.
But I think in the big picture, it's, like, you can imagine different kinds of bases for, kind of, computation that can be adaptively evolved and can do learning, and they're a different sort of raw material, but in the end, they're all going to be somewhat equivalent.
Let's dot C.
There's a question here about the next steps after the Artemis moon program. I don't know, those are always,
It's always a, A moving target that depends on a lot of

both politics of a country, and competition with other countries, and… and so on. And politics within these, organizations
that are, you know, sending spacecraft to here or there. I have to say, it was sort of interesting for me to watch the Artemis II launch, because when I was a kid.
I watched the… all the… the Apollo moon mission launches.
And, it was kind of notable to me what,
You know, 50 years later, 55 years later now, you know, how different
Was the kind of, was what was happening in this launch.
And I have to say, an awful lot was somewhat similar. I mean, there were sort of societal changes, like, you know, back 50 years ago, there were basically only male voices to be heard in, in all these, you know, the ground control people and the astronauts and so on.
And that's… that's now sort of equalized. That's… that's one sort of very superficial, change.
And then…
You know, a lot of the, you know, we launched a rocket, it's going up, it's reached this altitude, it's this far downrange, and so on.
All of that is just the same as it was before. I have to say that the one thing, you know, when the Apollo missions were launched for this Saturn V booster thing, there were these big black stripes, vertical stripes, that were painted on the white background of the rocket.
And, those were to… so that you could just look with a telescope or something, and see, was the thing rotating around, or was it staying kind of, going, you know, not… not twisting around.
They don't seem to have those stripes anymore, possibly because they have solid fuel boosters on the side, that you could see what's happening to those. But I think the other reason is that I'm sure there's a great deal more instrumentation
of, with accelerometers. Accelerometers have become easy to make now. Back in those days, the only way you could really make accelerometers was with gyroscopes embedded in capful cages and so on, whereas now you can make an accelerometer out of just solid-state
Microelectronics components.
And so on. So it's much easier to tell, you know, how things are moving. I also… it was notable that at some point it was, the thing had sort of locked on to GPS, and back in the day, there wasn't any GPS. The GPS satellites were far in the future, back in 1969, and so on.
the, It's,
So I think it's sort of remarkable that in the end, rockets are rockets, and they're always very complicated. I mean, when you're trying to, you know, take sort of…
The fuels have changed a bit, but let's say taking hydrogen and oxygen, and you're mixing them together, burning them, and getting them to produce energy that way, and expel gas out of the back of the rocket type thing.
Well, you know, when you look at rocket engines, you might have thought, oh, somebody will eventually make a simple rocket engine. That hasn't happened yet. I mean, it's a little bit like in cars, you might say, well, you know, an internal combustion engine is a lot more complicated than an electric motor, but with rockets, it's still the same basic physics
That it's always been. And it turns out that when you're doing the combustion for rocket.
you are sort of operating in various kinds of extremes of the materials that you're dealing with.
So, you know, you have liquid fuel being pumped in, or whatever it is, and then it's mixing in this combustion chamber, and you've got pumps, you know, trying to pump the usually turbopumps, which are trying to pump fuel in as quickly as possible, and mix it in certain ways, and then things are making transitions from one

phase of the material to another phase of the material, and it's all very complicated, and that translates into the typical, your typical rocket engine is full of all kinds of tubes and sort of chambers and so on, and valves and all that. And it just, it always seems to look very complicated, and I think that that complexity is an inevitable consequence of the properties of the materials
And the kinds of things you're getting them to do.
I mean, in, in, when it comes to,
just sort of… that's for kind of heavy lifting rockets. When it comes to rockets that you're using to kind of reposition a satellite, then there are simpler things. Ion thrusters and so on do look a lot simpler, but they don't provide the same kind of, sort of, heavy lift that you need to launch a big rocket. There's something where you've got a satellite and you just want to slowly turn it around. Then you can use these other things that are… that do look a bit simpler.
I think, the,
Let's see, yeah, another thing that's interesting about the Artemis II mission is that its, trajectory is rather different from the trajectories of,
Of the, of the Apollo missions.
it's, it's doing this thing of kind of looping around the Earth for multiple days, kind of getting to a higher and higher… it's not really an orbit. An orbit is a closed thing where you're coming back and coming to the same place again. It's kind of spiraling out, and then it kind of makes it
to the moon, and there's a nice animation that somebody posted, made with Golfram language, where you kind of see the thing, the spacecraft is going out towards the moon, and then the moon rushes by, and the spacecraft happens to arrive there right at the time when the moon is there, and get to go around the moon.
it's, it's, you know, when the Apollo missions were happening and the earlier moon… moon probes that started in 959, I guess, as the first, object that, Soviet Union landed on the moon.
or at least sent to the moon, it just crashed into the moon, the first ones did. There's kind of the question of, you know, can you hit the moon? You know, it's going really fast, and your spacecraft is going really fast, and, you know, if you don't get all the mechanics correct, you could miss the moon.
And, it took a lot of effort and computation to be able to work out the trajectories so that you didn't miss the moon, so to speak.
The, but in any case, the thing that… that ends up happening… so… so normally.
Okay, few basic facts of mechanics. If you have something that's not being acted on by gravity, and you point it in some direction, and it's moving in that direction, kind of Newton's first law of mechanics says it will just keep going in that direction, a straight line.
If you're in orbit, Then, in a first approximation.
you're gonna… going in a circle around some object, like the Earth or the Moon. Actually, in more of an approximation, you're going in an ellipse, and more than that, you're going in kind of a wiggly path
That depends on the mass distribution
where the mass is in the object you're dealing with. So the Earth, for example, has its… its mass is roughly evenly distributed in different directions around the Earth, and that means that the gravity doesn't vary much in different directions. On the Moon, that's not the case. The Moon has a sort of big lump on one side of it. I mean, it's spherical on the outside, but the density of rock is different on one side of the Moon than the other.

And so that means that orbits around the Moon don't follow that same kind of simple trajectory. In fact, it's hard to have an orbit that keeps going and just sort of closes on itself and keeps going around the Moon, because of that, sort of, that distortion in the gravity field of the Moon.
But in any case, the… the sort of… the basic way that… that spacecraft tend to, kind of.
they have to kind of transfer, for example, from one orbit to another, or they're kind of transferring, they're moving from one body, like the Earth, to the Moon, or whatever else.
And usually what happens… the way one thinks about that is they're going to fire a rocket, and that rocket is going to eventually change the velocity, the speed and direction that the thing is going at. So it's usually referred to as giving delta V, giving a difference in velocity, V,
To the spacecraft by firing some thruster, and, you know, you might fire it for a longer time, and it's pushing less… harder, or a shorter time, and it's pushing…
longer time is pushing less harder, a shorter time it's pushing harder. It leads to the same delta V, the change in velocity, at the end of the burn, at the end of the time when you're actually, you know, doing the, running the rocket, so to speak.
And so it's always a question of how many, sort of, delta V changes do you need to make to, to get the rocket to go to where you want it to go to. And,
The, but in the case of, and it's always a complicated thing, how you calculate where you want where you want the thing to go to get to the destination it's going to. So, for example, if you're sending a spacecraft to Jupiter.
It's, you might think you just go from the Earth.
you fire a rocket engine, and you go off towards Jupiter, and hopefully Jupiter will be in its orbit where you, you know, you're pointed in a different direction, you're going in a straight line. Jupiter is moving around its orbit. You kind of have to have calculated it right so that you get to Jupiter when Jupiter is where your straight line would take you.
But actually, in modern times, spacecraft have tended not to just go straight to where they're going to. Instead, they've done this idea of gravity assists, where the spacecraft goes… kind of loops around another body, like the Earth, for example. So even if you're going… you're on your way to Jupiter, but you get more or less to Mars, but then you kind of come back.
And the Earth is in just the right place that you can kind of do a loop around the Earth. And as you loop around the Earth, the gravity of the Earth is causing your trajectory to be such that you speed up as you kind of… you get this kind of slingshot around the Earth.
And that gives you enough speed that you can go off to Jupiter. Why do you need speed to get to Jupiter? You need speed to get to Jupiter because the Sun, the gravity of the Sun is pulling you back.
towards the sun, and you need to… you need to have kind of escaped that enough to be able to go out
out away from the Sun to get to Jupiter. And then you have to get there at a time when you can… when the gravity of Jupiter is going to pull you in, but
if you… you typically have to, actually change your velocity with some delta V thing at the point where you are… you would otherwise just fly by Jupiter, but you need to change your velocity so that you actually go into orbit around Jupiter, and you get fully captured by the gravity of Jupiter.
So that's, and one of the things that's really changed, actually, over the last 50 years
is the trajectories of spacecraft have got much more complicated, and both because I think the control… the spacecraft has gotten better, and because it's been easier to calculate these more elaborate kinds of trajectories, and so on.

I mean, it's always a challenging thing, you know, if you've got a spacecraft and it's out there, you know, you don't get to go and say, hey, let me push this button to reset it. Whatever happens has got to happen autonomously, and it's, you know, there's…
there's a, you know, the spacecraft is set up so it kind of, if its computer crashes, it will reboot, and usually it has multiple computers and so on. And, you know, the spacecraft is most of the time just going and doing its own thing, because
you can't… if you have a spacecraft going into deep space, you know, going off to Jupiter or something like that, you don't get to, sort of, do radio communications with it every day. Instead, there are a small number of antennas on the Earth that are powerful enough to do those communications, and typically it's timeshared, so a typical spacecraft
will be communicated with once a week, and once a week, it will kind of wake up and say, I'm waiting for my transmission, and actually, sometimes it's continuously waiting, but the antennas on the Earth only turn towards that spacecraft once a week, and give it a bunch of instructions, and try and reload programs on the computer inside the spacecraft.
And so on.
Let's see, Tony is asking…
Can you actually stop at a spot in space where you will never move from that spot?
Well, of course, it depends move relative to what?
I mean, if you're saying, can you be… so, one straightforward thing for the Earth, the Earth is rotating roughly once every 24 hours, and so you could ask, can you be at a place over a single spot on the Earth?
Well, the answer is, you're in orbit. If you're in orbit around the Earth, you better be orbiting sort of the same way as the Earth, and you better be orbiting above the equator of the Earth, because if you want to be sort of staying in the same spot relative to the surface of the Earth.
And it turns out that about 24,000 miles… so, okay, the period of an orbit depends on how far the orbit… how high the orbit is, how far away it is from the center of the Earth.
So, a low Earth orbit.
the satellite will go around the Earth once every 90 minutes, roughly. As you make the orbit further away from the Earth, the time it takes the satellite to go around the Earth will increase.
And so at some point.
the time it takes for the satellite to go around the Earth becomes basically 24 hours, the length of a day, and that happens about 24,000 miles away from, away from the Earth.
the orbit takes 24 hours to go round. So if you're orbiting over the equator, and it takes you 24 hours to go around, you're basically sitting stationary with respect to one point on the equator of the Earth.
And there's a whole bunch of satellites in so-called geostationary orbit, all lined up around the equator, 24,000 miles away.
That, are sort of in a fixed place. So, you know, for example, if you're setting up your antenna to, to communicate with one of those satellites, you can just fix your antenna, given that your building is in a particular place. You kind of fix your antenna to point at the angle you need to point
To reach that satellite that's staying in the same place relative to you.
Actually, they move around a little bit, but they're close enough that a typical antenna, you just point them at the thing, and they will be… it will be used for that. Now, the good news about something in geostationary orbit
is…

that you don't have to move the antenna to be looking at it, so to speak. The bad news is that it takes a while for a radio signal to get to 24,000 miles from the satellite, or it's actually more than that, because there's a… if you're… unless you… that would be the distance if you were looking straight up at the satellite, but typically you're looking at a slant angle.
So,
So for things where you are saying, let me talk to somebody, and you say something, they say something, that wouldn't work so well if you were going all the way to geostationary orbit and back. That works much better for things in low Earth orbit that are only, like, let's say, 100 miles away, in orbit.
Rather than 24,000 miles away, then so long as you can hand off
kind of, oh, I'm talking to this satellite now, oh, that one moved out of range, I'm talking to this one. That's sort of a solved problem at this point.
And, so that is what allows you to… to have, kind of, to do those things from,
much better with low Earth orbit. Now, in terms of can you stay stationary with respect to other things, the,
There's a question of, of can you stay, in…
In the solar system, if you just sort of put yourself somewhere, will you stay there? Or will you just, like, be falling into the sun? Well, if you're in orbit.
there are many orbits in the solar system. For example, there are many orbits around the Earth where the orbit stays stable, but sort of, as I say, you're moving relative to a particular point on the Earth.
in… around the Sun, there are, you know, the planets are all orbiting at different distances from the Sun. The asteroids, zillions of asteroids, are orbiting around the Sun. There are many… there are many ways that you can stay, sort of, at least a fixed distance from the Sun.
Now, if you say… if you… if you're just placed somewhere with zero velocity, you will just fall into the sun.
You have to be going at a given distance from the Sun, there's a definite speed you have to be going at to maintain your orbit around the Sun.
Now, if you ask the question, well, let's say you're orbiting around the Sun, but you want to stay in the same position relative to the Earth, can you do that?
not orbiting around the Earth, but orbiting around the Sun, but staying in the same position relative to the Earth, because something like another planet, like Mars, for example, it's orbiting… because it's a different distance away from the Sun, it's orbiting at a different… the time it takes for it to go around its orbit is different from the year that the Earth takes to go around its orbit, and so we've never lined up.
Mars and the Earth aren't lined up, sort of going straight out from the Sun. They go at different speeds around the Sun. And that's why, for example, if you want to send a spacecraft to Mars, there are particular times, I think every couple of years, when the Earth and Mars are kind of enough lined up in their orbits that it's a fairly short distance to get from Earth to Mars, comparatively short distance. Whereas, if they're not lined up in their orbits.
then Mars might be on the opposite side of the Sun from the Earth, and then it's a much longer trip.
to get to… to… to Mars.
So, question is, can you put things in a place where they will be at the same kind of location relative to the Earth, still going around the Sun? Well, there are places, the, well, you can have things that are in the same

let's see. The… the issue is, if you just put things kind of in the same orbit as the Earth.
then… They… let's see, how does this work?
That isn't a stable situation.
Because… why is that?
Well, there clearly are some things that can be stable, sort of, all the way around the orbit. That's why you get things like the rings of Saturn.
that have little… little pieces of rock and ice and so on, that are… that are in different places around the orbit, but they're all kind of, they're all… they're all just in different places in the orbit. And…
let's see, there's a problem with the stability of that, which I'm not immediately remembering, but the end result for the,
right, for… for… for the Earth and Sun.
There are these so-called Lagrange points.
which are places where the gravity of the Sun and the Earth matches to the point where something will be sort of stationary with respect to those two things. And so there are five Lagrange points for the Earth-Sun system, or for the Earth-Moon system, for that matter.
And, those are places where, relative to the Earth.
you stay in the same position as you go around the Sun, and so there are some spacecraft that have been put at the Lagrange points, in the Earth-Sun system. So that's kind of a way that you can stay stationary with respect to the Earth.
sort of still orbiting around the Sun.
Now, if you go… I mean, nothing stays fixed, like, in the galaxy.
we're orbiting the Sun, and it's orbiting around the center of the galaxy every couple hundred million years, 100 million years or so.
And, so that's, again, it's not staying in the same place, so to speak. The whole universe is expanding, so galaxies are getting further apart. So, again, they're not sort of staying in the same place. It's always relative to something you could sort of stay in the same place.
It's never, it's never that the,
sort of things are moving around all the time. Now, if you really go out to sort of the extremes, the universe… all the matter in the universe is… it's arranged in different places, it's moving in different ways, and so on and so on and so on, but in the end, there's a center of mass, there's an average
There's a… there's… there's kind of a… a… a thing that is stationary relative to all those things moving around, because it is… it is the place where, sort of, the,
The… the average of all those masses is…
And all the motions that are happening is some particular… is some… is… is something where, sort of, there's an average motion, and the… so, for example, the Earth is moving relative to the center of mass of the universe at about a thousandth of the speed of light.
And and that… that motion of the Earth
Part of that is the Earth going around the Sun, part of that is a small part, a bigger part is the Earth going around the center of the galaxy, another part is the galaxy moving around our cluster of galaxies, and another part
is our galaxy receding from other galaxies. Those all contribute, but in the end, there is a kind of a center of mass of the universe that the Earth is moving with respect to. And you can ask the question, could you put something so that it is not moving relative to the center of mass of the universe?

And the answer to that is, we don't completely know, and there is a… there's sort of an overall expansion of the universe,
that…
It's not completely clear whether the overall expansion of the universe is just because there was a big bang at the beginning, and then everything's going outwards, and if there's a change in the speed of things going outwards, it's because things are getting pulled together by their gravity, whether it is purely, kind of, you started off with the big bang at the beginning of the universe, and then everything's kind of flying apart from there, or whether there's something that is affecting the speed that
Things are kind of going apart, that is continually happening.
And the thing that people imagine might be doing that is what's called dark energy.
The… it's kind of like this ordinary matter.
then there's something people call dark matter, which I don't actually think is matter at all. I think it's a feature of the nature of gravity. But dark matter seems to be kind of concentrated… it's kind of a thing that sort of may be associated with galaxies and things like that.
And then there's sort of various effects that have been observed that make people think maybe the universe is accelerating in its expansion. Well, something has to be doing that.
And so, what's been introduced there is this so-called dark energy, which is essentially negative mass matter. Something with negative mass. It's not, in that case, not thought to be ordinary matter, like atoms and photons and things like this, but something else.
And if you have, sort of, that something else that is continually making the expansion of the universe accelerate, then no, you can't put things in a fixed place in the universe.
But we don't really know whether that's what's going on right now. So, you know, going out at the level of the whole universe, it might be possible, but it's, you know, we don't get to do that. I mean, you know, to escape the sun.
the gravity of the sun, you have to be going at 100,000 miles an hour. To escape the gravity of our galaxy, you have to be going at a million miles an hour. You know, rockets of the kind we have right now can't even do that. You'd need some other kind of propulsion mechanism that has more kind of bang for the buck
or more particularly, more thrust for mass than we can get with rockets done with chemistry.
Maybe nuclear fusion or something like that. That seems to be an idea that's coming back in vogue, is the idea of using nuclear fusion to power rockets.
Even if you can't control nuclear fusion on the ground, if you're kind of… if you're just having it sort of… you don't have to control it as much if it's kind of spewing things out into,
Into, into space.
Let's see… Well…
No good is asking, if we gave AI the goal of exploring the universe, what strange strategy might it come up with? It's a reasonable question. It's like,
Well, I don't know. It's like saying, if you give, sort of, early forms of life on this planet.
The kind of, Go be successful. You know, replicate a lot.
How does it do that? Well, you know, there are lots of forms of life that exist on this planet, all with different, slightly different solutions. I mean, realistically, at a base level, all forms of life on this planet are extremely similar. They all use proteins and DNA and
and ribosomes that transcribe things to make proteins and so on. All that machinery is probably more than 2 billion years old in the history of life on Earth.

And it's the same for every piece of life on Earth that we currently recognize. I mean, I have to say, I'm sure that the current mechanisms for life on Earth aren't the first form that life had on this planet.
In other words, the ribosome, for example, or the spliceosome, all these complicated molecular machines that do things that make life work the way that life works. These didn't just arrive one day. These are the result of many steps of adaptive evolution of some kind.
And my guess would be that there were previous forms of life
that sort of provided the cocoon, the environment, in which life of our kind could come into existence. And I think one of the interesting kind of questions is, you know, if we are Life 7.0 or something, and there were six forms before us, are there any signs, are there anything leftovers from those earlier forms of life?
And more extremely, could some of those earlier forms of life still exist today and not have been wiped out by our form of life, sort of taking all the nutrients and so on? And, you know, one wonders if there was some earlier form of life that had some different way of making molecular machinery that does something, would we even know it was here?
would it be sort of alien life on our planet that we never recognized? Because we're used to, when we say, is it alive?
Well, we don't have great tests for that, but, you know, one thing we can easily do is say, let's get its DNA, let's sequence its DNA.
well, we've already kind of, you know, we've completely predetermined things by the time we're even asking about DNA. If we say, well, does it have certain metabolic features? Does it, you know, ingest sugar? Does it like sugar and do things chemically with it? Well, that again is probably not the thing that is sort of the most, in order to sort of
create kind of a thing like life, you probably don't need sugar. There's probably other ways to do it. The particular chemistry that we have needs that.
But other chemistries that might be the things that supported making the chemistry we have might not need that. I think, you know, it raises the question, what is the fundamental defining feature of life?
And I've wondered about this for a long time. I've basically come to the conclusion that the defining feature of life is what one can call bulk orchestration. That is, that everything about life when it comes to anything other than the smallest molecules is somehow orchestrated. There's somehow some sort of way in which those molecules are moved around, and these molecules determine how those molecules move around, and so on. There's something where, in the end, the whole structure of the organism is kind of having
It's like all those molecules build up to make the whole organism, but yet the whole organism somehow… the fact that the whole organism behaves in the way that it does eventually has the consequence that it has all of this sort of orchestration of molecules moving around in just this way, and binding to each other, and building up more larger aggregates of molecules, and all those kinds of things.
So, if you ask the question, well, you know, what else is there that works that way? The answer is nothing works that way, with such a big tower. I mean, in a sense, when we look at towers of technology, we're starting to get, sort of, layers of… there's… you start from the transistors in the computer, and then you build up this, and build up that, and build up that. We're starting to see towers that are
Of somewhat comparable height.

When it comes to other things in the world, like, for example, I don't know, the structure of a star, or the structure of geology.
There's a lot…
it's a lot less… there's a lot less of a sort of tower of different kinds of mechanisms that are going on in geology. It's like, well, you know, there are… one kind of rock can dissolve and make another kind of rock, and this one can, under pressure, transform into that. But it's not nearly as big a tower of things that have happened as happens in life.
And the thing that is still unique about life for us is that it's kind of a molecular-scale machine, molecular-scale computation. The things we can do with computers, you know, we have a…
a computer, and it's got electrons, and they're going around in transistors and so on, but there are, you know, there are tens of thousands or hundreds of thousands of electrons for every bit in the transistor, millions of atoms in every transistor in the computer. Whereas in life.
There are molecules that are… well, they're molecules with maybe tens of thousands of atoms that are sort of carefully orchestrated and moved around to do things with other… with other molecules and so on.
And it's like, right down to the scale of those individual molecules, things are kind of orchestrated. And right up to the scale of the whole ecosystem of life on Earth.
That's… it all kind of makes this kind of somewhat unified, if messy, kind of orchestrated system. And that's… that's really different from what we see in geology.
It's what we… we haven't yet seen that in technology. We're getting closer to that in technology. I mean, it's sort of an interesting question whether, as we start to see, kind of,
Oh, things like… we start to do sort of adaptive evolution in things like neural nets. To what extent are we seeing those same levels, hierarchy of kind of bulk orchestration? And the answer, I think, is not yet at the same kind of level. In other words, but…
But it's sort of interesting to compare those, and I don't think I've… I haven't really finished my own efforts to compare those things. I mean, what tends to happen, this is sort of a thing I discovered fairly recently, is that it is inevitable
When you have a system that is sort of adapting its underlying rules, its underlying program, to achieve some simple-to-described purpose, like grow as tall as possible or something.
Then it becomes inevitable that inside the achievement of that, when it is done by sort of adaptively evolving, going from one program to the next, and so on, by little mutations and such like.
It turns out to be the case that there are little pieces of mechanism, I call them mechanoidal behavior, they're little, little mechanism-like things that the whole object might
combine together lots of little mechanisms, but you can identify, beyond the level of the individual elements, you can identify sort of a big structure with a thousand elements in it that is, oh yes, I recognize that mechanism, it's just, you know, periodically repeating, going on, off, on, off, or something like this.
That it's sort of inevitable that there are these little pieces of mechanoidal behavior that exist within the whole system.
And, I think… That's something that,
I don't know whether we've observed that in neural nets. It's a good question whether we've observed that. I've been trying to figure out even whether one observes that. In other words, a neural net right now, it's just sort of like a black box.

you feed an input, it gives output. If the output isn't what you like, you tweak the neural net, then it operates differently, and so on. And the question is, does the fact that you are adaptively evolving the neural net
to achieve the purpose you want, like recognizing cats from dogs, or producing English text, or whatever else, does the fact that you've done that adaptive evolution inevitably imply that there must be these little patches of mechanism, mechanoidal things, within the neural net?
I think the, it's a… it's an interesting question. You can ask the same thing about brains.
you know, this is an issue for neuroscience. Is there something that you can say about how brains work that is between, oh, this is how an individual neuron works, and this is how people sort of speak as, you know, in language or something like that at the level of the whole brain?
And what has been the activity of lots of neuroscience is finding little circuits, little mechanisms that exist in the brain.
And how those fit together to make, sort of, the whole thing the brain does, that's a different story. But the existence of those little circuits, is probably inevitable from, sort of, the adaptive way that brains learn and so on. And so, probably the same kind of thing is inevitable in neural nets, and people have certainly identified they have this
This concept of mechanistic interpretability, identifying little patches of mechanism.
Do those little patches of mechanism add up to the whole story of what the thing will do? I suspect not.
must there be some tantalizing little pieces of mechanism? Probably yes. It's really confusing, because it's like, I'm gonna find out how the whole thing works. It's gonna… I'm gonna unpack the whole end-to-end secret of the neural net, because I… because, look, I just found this one piece of mechanism.
it doesn't follow, and it probably isn't going to work that way. In fact, I think I have pretty good arguments that there never will be kind of a fully, sort of, easy-to-understand interpretation of what the whole neural net is doing. If there was, you wouldn't need that whole neural net. You would just be able to use that description to do whatever the neural net can do.
So… oh, I see that it is time for… I need to go back to my day job.
Lots of interesting questions here. These were difficult questions today. Well, thank you for those, and
I look forward to chatting with you another time.
So, bye for now.